Second phase of online safety regulation: Protection of children

Call for evidence

Call for evidence – Second phase of online safety regulation: Protection of children – Welsh overview

CALL FOR EVIDENCE:

Publication date: 10 January 2023
Closing date for responses: 21 March 2023

Contents

Section

Annexes

1. Overview

1.1 We are preparing to take on a new role as online safety regulator. As part of this role, we will be publishing codes of practice setting out the steps platforms can take to protect children online as well as guidance on how platforms should assess risks of harms to children. This document solicits evidence on risks of harms to children online and how they can be mitigated. As well as informing our media literacy work, we will use this evidence to help us prepare codes of practice on protection of children and relevant risk assessment guidance.

Why is Ofcom calling for evidence?

- In July 2022, we published a roadmap to regulation,[1] which set out our plan for implementing online safety regulation. Alongside the roadmap, we published a call for evidence for the first phase of online safety regulation, which focused in particular on the duties relating to illegal content.
- This call for evidence is focused on the matters that we anticipate will be included in our second consultation, which we expect to publish in autumn 2023, for **protecting children from legal content that is harmful** to them, as set out in more detail in our roadmap. This will include draft guidance to assist services in complying with their children's access assessment duties, draft guidance about how services likely to be accessed by children can comply with their duties to undertake their children's risk assessment and draft codes of practice explaining how services can comply with their safety duties relating to protecting children from harmful content.
- We welcome responses from interested stakeholders and we anticipate this call for evidence will be relevant to a wide range of stakeholders, including those with an interest or expertise in protecting children online, including civil society organisations and providers of online services. Responses will supplement the evidence gathered from our first call for evidence.
- Respondents should not feel it necessary to repeat answers to questions in this call for evidence, if they provided answers to similar questions in response to the first phase of online safety regulation call for evidence published in July 2022. Not all questions will be relevant to all respondents, for more information on which questions to respond to, see Annex 1.4.
- •Our call for evidence will remain open for **10 weeks** from publication and we request responses back by **5pm on 21 March 2023**. We are also planning to engage with stakeholders throughout this period with a view to holding sessions in January and February 2023.

Background

1.2 Ofcom is the UK's communications regulator, overseeing sectors including telecommunications, post, broadcast TV and radio. We regulate online video services established in the UK, including on-demand programme services (ODPS) and video-sharing platforms (VSPs). In line with our statutory duties, we have a programme of work dedicated to promoting and carrying out research on media literacy.

1.3 In February 2020, Government announced it was minded to appoint Ofcom as the regulator for online safety in the UK. We have been working since then to develop our understanding of the opportunities and challenges of online safety regulation, build our

[1] As the Bill is still passing through Parliament, the timings and requirements as set out in our roadmap remain provisional. We expect to publish further information about our implementation plans in due course, reflecting any updated expectations of when the Bill will obtain Royal Assent.

internal capability and begin planning for our regulatory approach. Since December 2020, we have been funded by Government to develop and strengthen our capabilities to prepare for this new role, including creating a new Online Safety Policy team, a new Trust & Safety Technology function and growing our Enforcement, Legal, Research and Insight, and Data teams.

1.4 We have been informed by our experience in regulating video-sharing platforms (VSPs) under Part 4B of the Communications Act 2003.[2] We are also continuing to cultivate effective relationships with agencies and regulators in the UK (for example, through the Digital Regulation Cooperation Forum) and in other jurisdictions, and to participate in global conversations and forums that seek solutions to problems of online safety.

1.5 The Online Safety Bill ('the Bill'), as currently drafted,[3] will require services which host user-generated content and search engines to have systems and processes for protecting individuals from certain types of harm online

1.6 The new regulatory regime is divided into four key areas. Each of these areas will have specific requirements on services, and a timeline for implementation. These areas are:

a) measures to protect people from illegal content;

b) measures to protect children from content which is harmful to them;

c) measures to empower adult users of 'Category 1' services[4] and ensuring clarity about what is allowed on the service and that they consistently apply their terms and conditions; and

d) increasing public transparency about what categorised services[5] are doing to tackle online harms, what measures they are taking to comply with their duties under the Bill, and how effective their safety measures are.

1.7 For the purposes of this document, a child is anyone under the age of 18,[6] unless otherwise specified. The child protection duties referred to above include:

a) to establish whether children are likely to access their service or part of their service: to do so, services will need to conduct a 'children's access assessment';

b) for all services likely to be accessed by children, to assess risks of harm to children and use proportionate measures relating to the design and operation of the service to effectively manage and mitigate the risks identified; and

c) to take specific action under the child safety duties in respect of 'primary priority' and 'priority' content that is harmful to children. For example, user-to-user services that are

[2] As it stands, the Online Safety Bill will ultimately repeal the VSP regime, although the exact date of repeal has not yet been set. Ofcom will provide support to services which will eventually transition between the two regimes.
[3] The Bill is currently at report stage in the House of Commons. The latest consolidated version of the Bill, published on 19 December 2022 can be found here.
[4] These are expected to be the highest reach user-to-user services with the highest risk functionalities.
[5] As well as Category 1 services, these include Category 2A services, which are likely to be the highest reach search services, and Category 2B services, which are other user-to-user services with potentially risky functionalities or characteristics.
[6] As currently defined under Clause 207 of the Bill.

likely to be accessed by children must use proportionate systems and processes designed to prevent children of any age from encountering primary priority content.

1.8 The new duties will apply to a wide range of user-to-user and search services, with different reach, sizes and risk levels. For example, regulated user-to-user services will include social media platforms, video-sharing platforms, forums, messaging apps, some online games, cloud storage and sites hosting user-generated pornographic content. More detail on the scope of the regime can be found in Ofcom's roadmap to regulation.

1.9 In addition, the Bill also imposes separate duties on providers of online pornographic content. In particular, they are required to ensure that children are not normally able to encounter pornographic content (as defined in the Bill) on their service, for example, by using age verification.

1.10 Ofcom will not receive any new powers until the Bill has received Royal Assent. Further details about our plan for consultation and implementation are provided in our roadmap, which was published in July 2022 and reflected our planning assumptions on implementation timing at that time. As the Bill is still passing through Parliament, the timings and requirements as set out in our roadmap remain provisional, and we expect to publish further information about our implementation plans in due course, reflecting any updated expectations of when the Bill will obtain Royal Assent.

1.11 As explained in our roadmap to regulation, we are planning on publishing two consultations in 2023. The first will be focused on illegal harms and we anticipate this to be published in late spring, subject to the timing of Royal Assent. The second consultation will focus on the protection of children and we anticipate this to be published in autumn, also subject to the timing of Royal Assent and any relevant secondary legislation.

1.12 Alongside our roadmap to regulation, in July 2022 we also published a call for evidence for the first phase of online safety regulation. In the first call for evidence, we sought views on matters that we anticipate will be included in our first consultation in 2023, including illegal harms to children such as grooming and child sexual abuse.

This call for evidence

1.13 Today, we are publishing a second call for evidence focused on:

- draft guidance on how services should conduct a children's access assessment;
- draft guidance on how services likely to be accessed by children are to undertake their children's risk assessment;
- draft guidance on how services hosting online pornographic content can comply with their duties to ensure that children cannot normally access that content;
- draft codes of practice explaining how services can comply with their duties to protect children from harmful content.

1.14 The questions in today's call for evidence therefore focus on assessment of children's access to services, the risk of harm to children from legal content that is harmful to them,[7] and measures that can be taken to protect children from harmful content, for example content moderation and age assurance and verification. The questions we pose below differ from those in the call for evidence we published in the summer, which focused on illegal harms. However, if you have already provided evidence relevant to these questions in response to the first call for evidence there is no need to provide the same evidence a second time here.

1.15 Through the questions set out in Annex 1, we are seeking evidence from a wide range of stakeholders, to strengthen our understanding of the range of approaches and techniques that platforms can employ to help them meet their online safety duties in respect of children. We would welcome evidence on the efficacy of the measures that can be taken to mitigate harm, their costs and the practicability of implementing them – where relevant reflecting on how this may differ depending on service characteristics.

1.16 We would like to hear from providers whose services are likely to fall within scope of the online safety framework as well as regulators, academics, civil society organisations, consumer representatives and other stakeholders with interest and expertise in the protection of children.

1.17 Any responses will supplement the evidence gathered from our first call for evidence. Not all questions will be relevant to all respondents. For more information on which questions to respond to, see A.1.4.

1.18 We envisage that the evidence provided in response will be valuable in preparing future reports and initiatives under our media literacy powers and, assuming the Bill passes, be of use in informing how we carry out our functions. These include drafting codes of practice and regulatory guidance which will set out steps that companies can take to comply with their child protection duties under the Bill. This call for evidence is one part of our preparations for taking on our new duties, alongside a wider programme of research and extensive stakeholder engagement.

1.19 As we are exploring the ways that services may comply with their duties under the Bill, which is yet to be approved by Parliament, this document and our questions should not be seen as an indication or statement of policy intent, but rather as an opportunity for stakeholders to provide input.

Next steps

1.20 Our call for evidence will remain open for **10 weeks** from publication and we request responses back by **5pm on 21 March 2023**.

[7] See A.1.2 to A1.3 for more information on the indicative list of primary priority and priority harms to children.

1.21 We consider that the complexity and novelty of this new regulatory regime will benefit from close engagement with a wide range of stakeholders. This call for evidence is our second key opportunity to do this ahead of our 2023 consultations and we look forward to continued close engagement as we develop our policies and plans.

A1. Questions

A1.1 In line with the plan set out by our roadmap, the focus of the second phase of our work will be on the duties relating to protecting children from harm. Where we refer to 'legal content that is harmful to children' or 'content that is harmful to children'[8] in the questions below, we mean both 'primary priority' and 'priority' content that is harmful to children to be defined in secondary legislation, and non-designated content that is harmful to children.

A1.2 In relation to 'primary priority' content, regulated user-to-user and search services will have a duty to operate the service using proportionate systems and processes designed to prevent children of any age from encountering such content, by means of the service. In July 2022 the Government published a Written Ministerial Statement,[9] which set out an indicative list of the types of content it expected to be listed as 'primary priority' and 'priority content' that is harmful to children, subject to further engagement with stakeholders and Parliamentarians, as well as consultation with Ofcom. The indicative categories of 'primary priority' content are:

- **pornography;**
- **content promoting self-harm (with some content which may be designated as priority content, e.g. content focused on recovery from self-harm);**
- **content promoting eating disorders (with some content which may be designated as priority content, e.g. content focused on recovery from an eating disorder); and**
- **legal suicide content (with some content which may be designated as priority content, e.g. content focused on recovery).**

A1.3 In relation to 'priority' content, regulated user-to-user and search services will have a duty to operate the service using proportionate systems and processes designed to protect children in age groups judged to be at risk of harm from encountering such content by means of the service (as well as from non-designated content identified in their risk assessments). The indicative categories of 'priority' harmful content are:

- **online abuse, cyberbullying and harassment;**
- **harmful health content (including health and vaccine misinformation and disinformation); and**
- **content depicting or encouraging violence.**

A1.4 Our questions are divided into sections, each covering areas relevant to elements of the new regulatory framework proposed by the Bill. We anticipate that these questions will be relevant to a wide range of stakeholders and therefore we do not address questions to specific respondents. Under each question, we have provided prompts to expand on the areas in which we are particularly interested. You do not need to respond to every

[8] "Content that is harmful to children" means – (a) primary priority content that is harmful to children, (b) priority content that is harmful to children, or (c) content, not within paragraph (a) or (b), of a kind which presents a material risk of significant harm to an appreciable number of children in the United Kingdom.
[9] https://questions-statements.parliament.uk/written-statements/detail/2022-07-07/hcws194.

question or prompt. Please provide evidence to support your responses; clearly evidenced and reasoned submissions will be most valuable in improving our understanding of the questions below.

A1.5 Our normal practice is to publish non-confidential versions of responses on our website. As such, you should specify if your response or a part of it is confidential, where relevant. If someone asks us to keep part or all of a response confidential, we will treat this request seriously and try to respect it. Please refer to Annex 2 for further instructions on submitting a response.

A1.6 If you are a business, organisation or group with expertise and relevant evidence around the questions below, and would like to get in touch with us about this call for evidence, please contact OS-CFE@ofcom.org.uk.

Preliminary question

A1.7 Our first question asks respondents for information about themselves, so that we can categorise responses to later questions. See section A2 for information about confidentiality.

A1.8 The Bill applies to a wide range of 'user-to-user services', defined as 'an internet service by means of which content that is generated directly on the service by a user of the service, or uploaded to or shared on the service by a user of the service, may be encountered by another user, or other users, of the service'. It will also apply to search services, defined as 'an internet service that is, or includes, a search engine'. The Bill also imposes duties on certain other providers of online pornographic content which is not user-generated. Any service which has significant numbers of UK users or which is targeted at the UK market will have new duties, and must comply with the new law.

Q1. To assist us in categorising responses, please provide a description of your organisation, service or interest in protection of children online.

For providers of online services, please provide information about:

- the type of service and functionalities[10] you provide;
- number of users globally, and in the UK (including children and their ages);
- global and UK revenues; and
- your business models and revenue generation.

Please indicate where this information is confidential.

[10] Within the Bill, 'functionalities' of user-to-user services include: creating a user profile, including an anonymous or pseudonymous profile; searching within the service for user-generated content or other users; forwarding content to, or sharing content with, other users of the service; sharing content on other internet services; sending direct messages to or speaking to other users of the service, or interacting with them in another way (for example by playing a game); expressing a view on content, including, for example, by applying a "like" or "dislike" button or other button of that nature, applying an emoji or symbol of any kind, engaging in yes/no voting or rating or scoring content in any way (including giving star or

Children's access to services

A1.9 All in-scope user-to-user and search services will be required to conduct a children's access assessment to determine if they are to be treated as 'likely to be accessed by children'. Services must determine (a) whether it is possible for children to access the service or a part of the service, and (b) if it is possible for children to access the service or a part of a service, whether the child user condition is met. The child user condition will be met if there is a significant number of children who are users of the service or that part of it, or the service, or that part of it, is of a kind likely to attract a significant number of users who are children. Services treated as 'likely to be accessed by children' will need to comply with the child safety duties under the Bill.

A1.10 The Bill currently states that *'providers may only conclude that it is not possible for children to access a service, or part of it, if there are systems or processes in place – for example, age verification, or another means of age assurance – that achieve the result that children are not normally able to access the service or part of it'.* Use of age assurance or age verification may be one way that providers could seek to fulfil their child safety duties, outlined in 1.7 particularly in relation to primary priority content. As referred to above, we will publish guidance for services on how they can comply with their duties to carry out their children's access assessment.

Q2. Can you identify factors which might indicate that a service is likely to attract child users?

In particular, please provide evidence explaining:

- the types of services which are likely to attract child users;
- any functionalities or other features of a service which are particularly likely to attract child users;
- the type of content that is likely to attract child users;
- if or how the factors you've identified may differ depending on the age of a child user; and
- whether there are any noticeable patterns in the activity of child users.

Where possible, please specify whether this evidence relates to child users in the UK or globally.

numerical ratings); sharing current or historic location information with other users of the service, recording a user's movements, or identifying which other users of the service are nearby; following or subscribing to particular kinds of content or particular users of the service; creating lists, collections, archives or directories of content or users of the service; tagging or labelling content present on the service; uploading content relating to goods or services; applying or changing settings on the service which affect the presentation of user-generated content on the service; accessing other internet services through content present on the service (for example through hyperlinks). 'Functionalities' of search services include: a feature that enables users to search websites or databases; and a feature that makes suggestions relating to users' search requests (predictive search functionality).

Q3. What information do services have about the age of users on different platforms (including children)?

In particular, please provide evidence explaining:

- the methods used to gather any information that can assist in estimating or assuring a user's age, either at the point a user first accesses the service or subsequently;
- what, if any, mechanisms are available to enable services to identify children in different age groups (for example children below age 13, aged 13-15, or aged 15-17); and
- how approaches to assessing the age of users are evolving.

Q4. How can services ensure that children cannot access a service, or a part of it?

In particular, please provide evidence explaining:

- how age assurance policies have been developed to date and what age group(s) they are intended to protect;
- if the service is tailored to meet age-appropriate needs (for example, by restricting specific content to specific users or part of a service), how this currently works or could work;
- how the efficacy of age assurance policies is or could be monitored; and
- how services can identify users that do not meet any relevant age limits and how is this appropriately addressed?

A1.11 Age assurance and age verification technologies could be an example of a type of system or process to protect children from harmful content. In the first call for evidence, which focused on illegal harms, we asked a question about the costs and efficacy of age assurance and age verification or related technologies. If you have evidence relevant to this question which you did not provide in response to the first call for evidence, for example evidence that relates specifically to the protection of children, please provide this.

Q5. What age assurance and age verification or related technologies are currently available to platforms to protect children from harmful content, and what is the impact and cost of using them?

In particular, please provide evidence explaining:

- how these technologies can be assessed for effectiveness or impact on users' safety;
- how accurate these technologies are in verifying the age of users, whether accuracy varies based on any user characteristics, and how effective they are at preventing children from accessing harmful content;
- any potential unintended consequences of implementing age assurance (such as risk of bias or exclusion), and how these can be mitigated;
- the safeguards necessary to ensure users' privacy and access to information is protected, and over restriction is avoided;
- which methods of age assurance users prefer, when offered a number of ways to verify their age;
- the cost of implementing and operating such technologies;
- how age assurance and age verification or related technologies may be circumvented; and
- what mitigations exist to reduce circumvention among users.

Content that is harmful to children

A1.12 Primary priority and priority content that is harmful to children will be defined in secondary legislation rather than in the Online Safety Act. Paragraphs A1.2 to A1.3 of this call for evidence sets out the Government's indicative list of 'primary priority' and 'priority' content that is harmful to children as published in the July 2022 Written Ministerial Statement.

Q6. Can you provide any evidence relating to the presence of content that is harmful to children on user-to-user and search services?

We are interested in evidence about the quantity or presence of such content on services of particular types or on user-to-user and search services in general. This could include, for example, the findings of relevant investigations, transparency reports and research papers that demonstrate how such content might vary across different services or types of service, or across services with particular groups of users, features or functionalities.

In particular, please provide evidence explaining:

- which groups or ages of children, if any, are more likely to encounter content that is harmful to children;
- the prevalence of primary priority content on user-to-user and search services;
- the prevalence of priority content on user-to-user and search services; and
- the types of service children are most likely to encounter different types of harmful content on.

Q7. Can you provide any evidence relating to the impact on children from accessing content that is harmful to them?

In particular, please provide evidence explaining:

- the impact of harm on children who have encountered primary priority content;
- the impact of harm on children who have encountered priority content;
- any specific risks children may encounter on search services associated with harmful content; and
- any differences in the impact of priority content on children in different age groups.

Risk assessment and management

A1.13 Risk assessment and management will be a cornerstone of the regime and is a core requirement of the Bill. The Bill sets out the duty on Ofcom to carry out a risk assessment to identify and assess risks of harm to children, in different ages groups, from content that is harmful to children presented by user-to-user and search services. The risk assessments must, amongst other things, identify characteristics of different kinds of user-to-user and search services that are relevant to content that is harmful to children and assess the impacts of those kinds of characteristics on this particular risk. Ofcom must develop 'risk profiles' for user-to-user and search services which relate to the risk of harm. Ofcom must also publish its risk assessment (in a register of risks) and risk profiles, keeping both up to date.

A1.14 Ofcom must prepare and publish guidance for providers of regulated services to assist them in complying with their duties to carry out their own risk assessments. In undertaking their risk assessments, services must take account of the risk profiles prepared by Ofcom that relate to the risk of harm to children.

Q8. How do services currently assess the risk of harm to children in the UK from content that is harmful to them?

In particular, please provide evidence explaining:

- how risks from harmful content are identified (including any relevant internal processes, policies and documents); and
- in considering the potential risk that children may encounter harmful content, the extent to which services factor in evidence on users' behaviour and age.

Q9. What are the exacerbating risk factors services do or should consider which may have an impact on the risk of harm to children in the UK?

In particular, please provide evidence of:

- how the user base of the service may have an impact on the risk of harm to children;
- how the business model of the service may have an impact on the risk of harm to children;
- the functionalities or features of services which may have an impact on the risk of harm to children; and
- what mitigations exist for these risk factors.

Q10. What are the governance, accountability and decision-making structures for child user and platform safety?

As part of your answer, please outline how different teams may consider child user safety risks across different business functions such as product development, management, engineering, public policy, safety, legal, business development and marketing.

Please consider:

- how staff are/should be trained in a service to understand how their own roles and responsibilities can create risks to child user safety;
- how services can ensure consistency in consideration of child user safety across teams;
- examples of best practice or innovative approaches to governance, accountability and decision making; and
- possible costs associated with assessing risks of harm to child users, including specific reference to costs associated with ensuring services have governance and decision-making processes for child user safety where applicable.

Terms of service and policy statements

A1.15 The Bill sets a number of expectations around regulated user-to-user services' terms of service. Terms of service are expected to be clear and accessible to users and consistently

applied. Search services will be required to outline similar provisions in public policy statements.

Q11. What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements for children (including children of different ages)?

Please submit evidence about what approaches make terms or policies clear and accessible to children.

Q12. How do terms of service or public policy statements treat 'primary priority' and 'priority' harmful content?[11]

Please outline as part of your answer:

- what services currently cover in their terms of service and public policy documents in relation to primary priority and priority harmful content and why, including any reference to what is considered harmful content, any measures to identify it, and sanctions applied to users who are in breach;
- whether when drafting these documents, the specific needs of children are considered;
- evidence of the process, time and any costs involved in developing these terms; and
- whether you have any evidence about how child users engage with terms of service or public policy statements, or whether children understand what they mean in practice.

Reporting and complaints

A1.16 Under the Bill, regulated user-to-user and search services likely to be accessed by children will be required to operate systems and processes for the reporting of content that is harmful to children.

A1.17 Under the Bill, regulated services will also be required to operate complaints procedures, allowing users and affected persons[12] to complain about content they consider to be harmful; or if they consider that the provider is not complying with its duties, or that their content has been removed or ability to make use of the service has been restricted unduly.

A1.18 Services likely to be accessed by children will also have to enable users to complain if they are unable to access content because measures used to comply with the child safety duties (for example, age assurance) have resulted in an incorrect assessment of the user's age.

[11] See A1.2 to A1.3 for more information on the indicative list of harms to children.

[12] The Bill currently defines "affected person" as a person, other than a user of the service in question, who is in the United Kingdom and who is— (a) the subject of the content, (b) a member of a class or group of people with a certain characteristic targeted by the content, (c) a parent of, or other adult with responsibility for, a child who is a user of the service or is the subject of the content, or (d) an adult providing assistance in using the service to another adult who requires such assistance, where that other adult is a user of the service or is the subject of the content.

A1.19 These questions focus on reporting and complaints mechanisms for children. If you're a service provider that has already provided all of the information you have on reporting and complaints in answers to these questions in our first call for evidence, including information that relates to the protection of children, please move to the next section. If you provided a response to the first call for evidence but there is additional information relevant to the protection of children that you could provide, please respond to this section.

Q13. What can providers of online services do to enhance children's accessibility and awareness of reporting and complaints mechanisms?

Please submit evidence about what features make user reporting and complaints systems accessible to children, considering:

- reporting or complaints routes for registered and non-registered users;
- how services could encourage children to report content;
- how to ensure that reporting and complaints mechanisms are not misused;
- the key choices and factors involved in designing these mechanisms;
- how to ensure that reporting and complaints mechanisms are user-friendly and accessible to children;
- whether particular consideration is given to the different needs of child users, for example children of different ages;
- whether user reports are anonymous to the service;
- whether users are notified that their reports are anonymous to other users; and
- what happens to users who have their content or account reported.

Q14. Can you provide any evidence or information about the best practices for accurate reporting and/or complaints mechanisms in place for legal content that is harmful to children, or users who post this content, and how these processes are designed and maintained?

We are interested in obtaining evidence on:

- how users, including children, report harmful content on services (including the mechanisms' location and prominence for users, and any screenshots you could provide) and whether this is or ought to be separate to complaints procedures;
- whether users need to create accounts to access reporting and complaints mechanisms;
- what type of content or conduct, users and non-users may make a complaint about or report, including any specific lists or categories;
- whether reporting and complaints mechanisms are effective, in terms of identifying content that is harmful to children, and how to determine effectiveness;
- whether there are any reporting or complaints mechanisms you consider to be less effective in terms of identifying content that is harmful to children, and how you determine this;
- the use of trusted flaggers[13] (and if reports from trusted flaggers should be prioritised over reports or complaints from users); and
- the cost involved in designing and maintaining reporting and/ or complaints mechanisms.

Q15. What actions do or should services take in response to reports or complaints about online content harmful to children (including complaints from children)?

Please provide relevant evidence explaining your response to this question.

Design and operation of the service, including functionalities and algorithms

A1.20 The duties set out in paragraph 1.7 also encompass measures relating to the design of functionalities, algorithms and other features of the service, and provision of user support measures.

A1.21 We are interested in understanding services' approach to design as a way to mitigate harm to children, including how safety is considered in the design of products and functionalities. The Bill provides a number of examples of 'functionalities'. For user-to-user services, this includes the ability of a user to have an anonymous profile, to like or dislike content, to share location information with other users, and to forward content to other

[13] 'Trusted flaggers' are organisations (such as civil society, government agencies, or other relevant groups) who have a specific expertise in online harms. 'Trusted flaggers' may partner with platforms to directly flag potentially violating content to the platforms.

users. An example of a functionality on a search service is auto completion, where a search engine predicts the rest of a query that a user has begun to type.

A1.22 We are interested in hearing about aspects of design that could be considered to help prevent harm to children. This could include both providing new features to child users (e.g. allowing control over what they encounter), or restricting functionality (e.g. limiting discoverability).

Q16. What functionalities or features currently exist that are designed to prevent or mitigate the risk or impact of content that is harmful to children? A1.21 provides some examples of functionalities.

In particular, please provide evidence explaining:

- any functionalities or features available to services, which you consider can effectively prevent harm to children; and
- any functionalities or features in development that services could consider implementing to mitigate the risk or impact on children of content harmful to them.

Q17. To what extent does or can a service adopt functionalities or features, designed to mitigate the risk or impact of content that is harmful to children on that service?

In particular, please provide evidence explaining:

- what control can or should children have over what they are shown or content that is delivered to them;
- to what extent the features or functionalities identified are reliant on other technology – for instance, age assurance, age verification or ID verification mechanisms;
- the costs or cost drivers involved in developing these features or functionalities;
- whether child safety is incorporated into the product design and development processes;
- to what extent evidence is considered about child user behaviour when developing features or functionalities intended to enhance user safety;
- what evidence can be provided relating to measures, including evidence around the impact and effectiveness of these techniques, in terms of reducing harm to children; and
- how services assess the impact of potential mitigations on users' privacy and freedom of expression and minimise the risk of over restriction.

Q18. How can services support the safety and wellbeing of UK child users as regards to content that is harmful to them?

In particular, please provide evidence explaining:

- whether this involves or should involve support provided through the platform (e.g. signposting to resources);
- whether this involves or should involve off-platform support (e.g. funding or facilitating programmes);
- how are these interventions or should these interventions be embedded into the user journey via service design;
- how effective these types of interventions, in terms of minimising harm from and impact of harmful content to children are; and
- the costs involved in implementing the support measures you have described.

Q19. With reference to content that is harmful to children, how can a service mitigate any risks to children posed by the design of algorithms that support the function of the service (e.g. search engines, or social and content recommender systems)?

In particular, please provide evidence explaining:

- if different from the risk assessment process outlined in response to Q8 how services assess the risk to children from algorithms central to the function of the service;
- what safeguards services have in place to mitigate the risks posed by algorithms (e.g. testing them before they are put into use, and monitoring their performance in real world settings);
- what safeguards services have in place to mitigate the risks posed using recommender systems in particular (e.g. providing users with controls over what they are shown, such as through keyword filters);
- additional requirements for safety in algorithms (e.g. accurate content categorisation);
- the costs involved in implementing these safeguards. In the absence of specific costs, please provide indication of the key cost drivers;
- how services can measure the effectiveness of these safeguards, in terms of reducing harm to users;
- what information services can provide to demonstrate the effectiveness of such safeguards; and
- how services can assess the impact of these safeguards on users' privacy and minimise the risk of over restriction.

Moderation

A1.23 Under their duties set out in Part 3 of the Bill, regulated services likely to be accessed by children must meet certain requirements in relation to 'primary priority' content, 'priority content' and non-designated content which is harmful to children. In relation to 'primary

priority' content services must operate a service using proportionate systems and processes designed to prevent children of any age from encountering it by means of the service. In relation to 'priority content' and other content that is harmful to children, services must operate a service using proportionate systems and processes designed to protect children in age groups judged to be at risk of harm from this content from encountering it by means of the service. In determining what is proportionate, services should consider the findings of the most recent children's risk assessment, as well as the size and capacity of the service as relevant factors. In meeting these requirements services may include use of content moderation (including systems used by search services to analyse searchable content).

Q20. Could improvements be made to content moderation to deliver greater protection for children, without unduly restricting user activity? If so, what?

In particular, please provide relevant evidence explaining:

- improvements in terms of user safety and user rights (e.g. freedom of expression), as well as any relevant considerations around potential costs or cost drivers;
- evidence of the effectiveness of existing moderation systems; and
- examples of where relevant content moderation processes are particularly good or poor.

Q21. What automated, or partially automated, moderation systems are currently available (or in development) for content that is harmful to children?

In particular, please consider:

- the suitability of automated (or part-automated) moderation systems to identify content that is harmful to children;
- whether and how automated moderation systems differ by the type of content (e.g. text, image or video);
- how in-house automated content moderation systems are developed and (in the case of technology which uses AI or machine-learning) trained or tested;
- how the data used to develop, train, test or operate content moderation systems is sourced and whether it is representative of the intended real-world scenario;
- the range or quality of third-party content moderation system providers available in the UK;
- how effective automated content moderation systems are, in terms of identifying target content that is harmful to child users, and how this may vary by harm;
- what evidence is available to assess the accuracy of automated moderation techniques (e.g. regarding the frequency of false positives/negatives);
- how action is taken in relation to content identified by automated means as potentially being harmful to children (e.g. automated action, human action, or further review);
- what safeguards are employed to mitigate adverse impacts of automated content moderation, e.g. on privacy and/or freedom of expression;
- whether certain types of automated moderation techniques might be better suited to certain harms or types of content and why; and
- what barriers and costs are involved in deploying these automated moderation systems.

Q22. How are human moderators used to identify and assess content that is harmful to children?

In particular, please consider:

- the typical role of an effective human moderator;
- how to determine the level of human moderation required by a platform, including by type of harmful content;[14]
- whether moderators are employed by the service, outsourced, or volunteers;
- whether moderators are vetted, and how; and
- the type of coverage (e.g. weekends or overnight, UK time) moderators provide.

[14] See A.1.2 to A1.3 for more information on the indicative list of harms to children.

Q23. What training and support is or should be provided to moderators?

In particular, please consider:

- whether certain moderators are specialised in certain harms or speech issues;
- whether training is provided or updated, and frequency of these; and
- whether moderators are trained to identify content that is harmful to children.

Q24. How do human moderators and automated systems work together, and what is their relative scale? How should services guard against automation bias?

In particular, please provide evidence explaining:

- the costs involved in these practices. In the absence of specific costs, please provide indication of cost drivers (e.g. moderator location) and other relevant figures (e.g. number of moderators employed, how many items the platform moderates per day); and
- how services can assess the accuracy and consistency of human moderation teams.

Other mitigations to protect children

A1.24 As well as content moderation, one of the child safety duties in the Bill requires services to use proportionate systems and processes designed to prevent children of any age from encountering primary priority content that is harmful to children. Aside from age assurance addressed in questions 4 and 5, we are seeking input on which other mitigations may be proportionate and effective at preventing children from encountering primary priority content.

Q25. In what instances is content that is harmful to children, that is in contravention of terms and conditions, removed from a service or the part of a service that children can access?

Please outline the circumstances in which content that is harmful to children is removed and how this may differ by type of harmful content. Primary priority content and priority content are outlined in A1.2 to A1.3.

Q26. What other mitigations do services currently have to protect children from harmful content?

In particular, please provide evidence explaining:

- available mitigations to prevent children from accessing primary priority content;
- available mitigations to protect children from accessing priority content;
- what types of service (e.g. social media, search, gaming etc.) available mitigations are suitable for;
- the costs of implementing mitigations;
- if applicable, the potential risks associated with mitigations;
- if applicable, any non-cost barriers of implementing mitigations;
- how effective the existing mitigations platforms have in place are;
- how 'other features' (alongside design of functionalities and algorithms) are designed to ensure safety to children; and
- what, if any, mitigations are feasible but may not be currently available on services likely to attract children.

Q27. Where children attempt to circumvent mitigations in place on a service, what further systems and processes can a service put in place to protect children?

Here we are particularly interested in the circumvention of measures that are not age assurance technologies, which we ask about in Q5. In particular, please provide evidence explaining:

- the ways in which some child users may attempt to circumvent mitigations put in place to protect them from harmful content;
- the ways in which services can combat these, particularly in relation to primary priority content, which children should be prevented from accessing; and
- examples of best practice for how to prevent children from encountering primary priority content.

Other

Q28. Other than those covered above in this document, are you aware of other measures available for mitigating the risk, and impact of, harm from content that is harmful to children?

We would be interested in any evidence you can provide on their efficacy, in terms of reducing harm to child users, cost and impact on user rights and user experience.

A2. Responding to this call for evidence

How to respond

A2.1 Ofcom would like to receive responses by **5pm on 21 March 2023**.

A2.2 You can download a response form from https://www.ofcom.org.uk/consultations-and-statements/category-1/call-for-evidence-second-phase-of-online-safety-regulation. You can return this by email or post to the address provided in the response form.

A2.3 If your response is a large file, or has supporting charts, tables or other data, please email it to os-cfe@ofcom.org.uk, as an attachment in Microsoft Word format, together with the cover sheet. This email address is for this call for evidence only, and will not be valid after 21 April 2023.

A2.4 Responses may alternatively be posted to the address below, marked with the title of the consultation:

Online Safety Call for Evidence
Ofcom
Riverside House
2A Southwark Bridge Road
London SE1 9HA

A2.5 We welcome responses in formats other than print, for example an audio recording or a British Sign Language video. To respond in BSL:

- send us a video of you signing your response (up to five minutes); or
- upload a video of you signing your response to YouTube (or another video hosting platform) and send us a link.

A2.6 We will publish a transcript of any audio or video responses we receive (unless your response is confidential).

A2.7 We do not need a paper copy of your response as well as an electronic version. We will acknowledge receipt of a response submitted to us by email.

A2.8 You do not have to answer all the questions in the call for evidence if you do not have a view; a short response on just one point is fine. We also welcome joint responses.

A2.9 It would be helpful if your response could include direct answers to the questions asked in Annex 1 of this document. It would also help if you could explain why you hold your views and provide supporting evidence.

A2.10 If you want to discuss the issues and questions raised in this document, please contact the Online Safety team by email on os-cfe@ofcom.org.uk.

Confidentiality

A2.11 In the interests of transparency and good regulatory practice, and because we believe it is important that everyone who is interested in an issue can see other respondents' views, we usually publish responses on the Ofcom website.

A2.12 If you think your response should be kept confidential, please specify which part(s) this applies to, and explain why. Please send any confidential sections as a separate annex. If you want your name, address, other contact details or job title to remain confidential, please provide them only in the cover sheet, so that we don't have to edit your response.

A2.13 If someone asks us to keep part or all of a response confidential, we will treat this request seriously and try to respect it, but sometimes we may need to disclose responses to fulfil certain legal obligations (e.g. where it is proportionate and fair to do so to enable appropriate consultation, or if we are ordered to disclose them). Please also note when responding to this call for evidence that we do not require information that might identify individuals that may be alleged perpetrators of harm(s).

A2.14 To fulfil our pre-disclosure duty, we may share a copy of your non-confidential response with the relevant government department before we publish it on our website. This is the Department for Business, Energy and Industrial Strategy (BEIS) for postal matters, and the Department for Culture, Media and Sport (DCMS) for all other matters.

A2.15 Please also note that copyright and all other intellectual property in responses will be assumed to be licensed to Ofcom to use. Ofcom's intellectual property rights are explained further in our Terms of Use. Please also see our Privacy Statement.

Updates on Ofcom publications

A2.16 If you wish, you can register to receive mail updates alerting you to new Ofcom publications.

A3. Response coversheet

BASIC DETAILS

Call for evidence title: Second phase of online safety regulation: Protection of children

To (Ofcom contact): **Online Safety team, OS-CFE@ofcom.org.uk**

Name of respondent:

Representing (self or organisation/s):

Address (if not received by email):

CONFIDENTIALITY

Please tick below what part of your response you consider is confidential, giving your reasons why

Nothing ☐

Name/contact details/job title ☐

Whole response ☐

Organisation ☐

Part of the response ☐

If there is no separate annex, which parts? _____

_____

If you want part of your response, your name or your organisation not to be published, can Ofcom still publish a reference to the contents of your response (including, for any confidential parts, a general summary that does not disclose the specific information or enable you to be identified)?

_____

DECLARATION

I confirm that the correspondence supplied with this cover sheet is a formal response to Ofcom's call for evidence that Ofcom can publish subject to the confidentiality section above. However, in supplying this response, I understand that Ofcom may need to disclose some information marked as confidential where it is proportionate and fair to do so to enable appropriate consultation, or if Ofcom is ordered to disclose them. If I have sent my response by email, Ofcom can disregard any standard e-mail text about not disclosing email contents and attachments.

Ofcom aims to publish responses at regular intervals; if your response is non-confidential (in whole or in part), and you would prefer us to publish your response only once the response period has ended, please tick here. ☐

Name Signed (if hard copy)