# Your response

# Volume 2: The causes and impacts of online harm

## Ofcom's Register of Risks

| **Question 1:** | |
|---|---|
| i) | Do you have any comments on Ofcom's assessment of the causes and impacts of online harms? |
| Response:<br><br>The safety of our users is paramount, and we appreciate Ofcom's commitment to base its proposals on analysis. We also appreciate Ofcom's candour in acknowledging that some gaps remain in its knowledge of the causes and impacts of online harms, and that it may not have a full picture of both the range of harms and types of services in scope. We would recommend ongoing and constructive engagement with services to better inform and address any information gaps and we look forward to working closely and constructively with Ofcom where we can. | |
| ii) | Do you think we have missed anything important in our analysis? Please provide evidence to support your answer. |
| Response: | |
| iii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: no | |

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

## Governance and accountability

| **Question 4:** | |
|---|---|
| i) | Do you agree with the types of services that we propose the governance and accountability measures should apply to? |
| Response: | |
| ii) | Please explain your answer. |
| Response:<br><br>We agree with the general principle that services with high risk from multiple kinds of illegal harms should be subject to more onerous measures. However, as outlined in more detail throughout our responses (especially questions 15 and 16), we do not agree with the proposed | |

definition of multi-risk services. We have concerns that the definition of multi-risk services as outlined in the consultation will lead to an oversimplified approach, where services with medium risk from two kinds of harm are automatically subject to the same measures as services with high risk from many kinds of harm.

iii) Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: no

**Question 5:**

i) 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?

Response:

We believe that ultimately, services are best positioned to understand how to structure their own content safety programmes, processes and tools in a way that most effectively addresses harms. We would therefore be concerned that, if this measure were to be introduced, third-party auditors simply would not have the necessary context, product, and business knowledge to produce adequate assessments. We believe that with the support, guidance, and oversight of Ofcom, services themselves are best placed to determine whether their processes are effective.

We are also concerned about the cost implications of this potential measure, which would likely present significant challenges, especially to smaller services, because third parties will be able to charge high fees for performing the audit. These cost challenges would also negatively impact on competition, as large companies with more resources and bigger compliance/legal teams would more easily be able to manage audit processes and fees.

Finally, this potential measure could have other unintended consequences for competition if care is not taken to ensure that the auditor is properly independent. It is possible that the largest, most well-resourced technology companies will wish to take on the role of auditor, which would allow them to become even more entrenched whilst smaller companies struggle to cope with the costs and resourcing demands of complying with the audit measure.

ii) Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: no

**Question 6:**

| | |
|---|---|
| i) | 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? |
| Response:<br><br>We would recommend very cautious consideration of any future measures that seek to link remuneration with online safety outcomes. Such prescriptive measures risk services taking overly intrusive measures which could undermine users privacy and freedom of expression. We welcome Ofcom's own assessment of such measures and its intention to recommend only well-evidenced risk-based measures that effectively support online safety outcomes. | |
| ii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: no | |

## Service's risk assessment

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

| **Question 8:** | |
|---|---|
| i) | 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? |
| Response: | |
| ii) | Please provide the underlying arguments and evidence that support your views. |
| Response:<br><br>We support the recommendation that services have a written policy in place to review their risk assessment at least every 12 months. The proposal of 12 months is sensible, and in alignment with similar compliance measures required by the Digital Services Act. | |
| iii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: no | |

| **Question 9:** | |
|---|---|
| i) | Are the Risk Profiles sufficiently clear? |
| Response: | |
| ii) | Please provide the underlying arguments and evidence that support your views. |
| Response:<br><br>We support Ofcom's aim to implement a risk-based and proportionate online safety regime. As Ofcom rightly notes, there are no "one size fits all" solutions, and this should be the case for all services. We also believe that before concluding that such services automatically pose an elevated risk of these types of harms, the Risk Profiles should take into account the efficacy of a service's content moderation practices, business model and product functionality, and likelihood/volume of harms on the platform should be considered. | |
| iii) | Do you think the information provided on risk factors will help you understand the risks on your service? |
| Response: | |
| iv) | Please provide the underlying arguments and evidence that support your views. |
| Response: | |
| v) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: no | |

# Volume 4: What should services do to mitigate the risk of online harms

## Our approach to the Illegal content Codes of Practice

| **Question 12:** | |
|---|---|
| i) | Do you have any comments on our overarching approach to developing our illegal content Codes of Practice? |
| Response:<br><br>The safety of our users is paramount, and we support Ofcom's commitment to implementing a risk-based and proportionate regime. There is no "one size fits all" solution to addressing illegal content, and a proportional approach will be key to ensuring the Act's effectiveness across the wide range of services that it covers - as Ofcom rightly recognises. | |
| ii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |

| |
|---|
| Response: no |

| **Question 13:** | |
|---|---|
| i) | 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? |
| Response: | |
| ii) | Please provide the underlying arguments and evidence that support your views. |
| Response:<br><br>We agree with a proportional approach that doesn't place an undue burden on small/medium sized companies with limited resources and services that are low risk.<br><br>For those risks and harms that we do experience, we have put in place a range of mitigating measures, in the form of best-practice policies, processes and tools designed to prevent and minimise these harms, including the use of industry standard hash-matching technologies, external reporting, and human review for the detection of known CSAM and terror content. | |
| iii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: no | |

| **Question 14:** | |
|---|---|
| i) | Do you agree with our definition of large services? |
| Response: | |
| ii) | Please provide the underlying arguments and evidence that support your views. |
| Response:<br><br>We welcome the effort to ensure consistency in definitional thresholds across the various international regulatory regimes, such as the Digital Services Act.<br><br>However, to help companies navigate and comply with obligations under the Act, we seek further clarification on the definition of a "user" for the purpose of establishing a service's size. The Act and consultation as currently written offer a broad definition that does not take into account entities that may have a mix of regulated and unregulated services, and services which may have a large number of dormant/inactive users. Where an entity is made up of various services, some of which may be regulated under the Act and some of which may not be, it is currently unclear whether UK users should be counted from only the regulated service(s), or from the entity as a whole. In addition, where services have a large number of registered yet inactive users, it is unclear whether these should be counted as users for the purposes of establishing the service's | |

size to check whether it meets the large service threshold. We would encourage Ofcom to provide further detail on these definitions.

For our part, we take the small number of instances where users have abused our service to share or store illegal content very seriously. To that end we have developed a robust set of policies and processes to address such content. We use a combination of industry-standard hash-matching technology and human review to detect content that violates our Terms of Service and Acceptable Use Policy and take appropriate action.

Please see question 16 for more details on how we address illegal content.

| iii) Is this response confidential? (if yes, please specify which part(s) are confidential) |
| --- |
| Response: no |

| **Question 15:** |
| --- |
| i) Do you agree with our definition of multi-risk services? |
| Response: |
| ii) Please provide the underlying arguments and evidence that support your views. |
| Response:<br><br>We agree with the general principle that services with risk from multiple kinds of illegal harms should be required to mitigate those risks. However, we believe that Ofcom should foreground overall severity of risks, rather than whether a service is single-risk or multi-risk, as a small number of high risks may often present more danger to users than a moderate number of medium risks. In addition, we have two concerns with the way Ofcom has defined multi-risk.<br><br>**Defining multi-risk as at medium/high risk of two or more harms.**<br>Firstly, the difference between services with medium/high risk from two kinds of illegal harms, and services with medium/high risk from ten kinds of illegal harms, is significant. We have concerns that the definition of multi-risk services as outlined in the consultation will lead to an oversimplified approach, where (holding risk level constant) services with risk across two harm categories are subject to the same measures as services with risk across many categories. This approach leaves little room for nuance or proportionality and is not in line with Ofcom's own proposals for a risk-based, proportionate approach.<br><br>**Few companies will be single risk.**<br>Secondly, we suspect that many companies will have more than one medium level risk, which means they will be covered by the term "multi-risk", and very few will be "single risk". This negates the usefulness of the classification.<br><br>Considering all of the above, we believe that a more appropriate approach for Ofcom would be to foreground severity of risk, and consider the number of harms each service is at risk for on a case-by-case basis, rather than applying a blanket classification. This would help to ensure that platforms that present the greatest danger to users will be subject to the most onerous measures, and avoid an undue compliance burden for less risky services. |
| iii) Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: no |

| **Question 16:** |
| --- |
| i) Do you have any comments on the draft Codes of Practice themselves? |
| Response: |

Generally we are supportive of the draft Codes of Practice. With that in mind, we have set out our approach to safety below and would appreciate further clarity from Ofcom about how such mitigation measures should be factored into services' risk assessments.

We believe that a more effective approach would be for Ofcom to consider each service on a case-by-case basis, taking into account the service's specific functionalities and user base to establish an overall level of risk. This approach would ensure that all services are placed on equal footing, and avoid small-yet-high-risk services falling through the cracks (equally, no larger-yet-low-risk services would be unduly burdened). It would also be more in line with Ofcom's commitment to proportionality.

However, despite our service being a low risk for most types of harm, we are aware that a very small minority of users do, unfortunately, abuse our service in ways that violate our Acceptable Use Policy. We take these instances very seriously and have invested heavily in a robust content safety programme to address such content. We set out below the steps we take to address this issue. This robust set of policies and processes helps to ensure that, whilst there is some risk of illegal content on our service, this risk is minimised at every possible junction, and addressed quickly and effectively when encountered.

- We maintain a clear Acceptable Use Policy prohibiting illegal and harmful content which users must agree to adhere to in order to access our service.
- We've made it easy to report harmful content through our reporting tools and complaints portal.
- We have a highly trained and experienced content safety team whose responsibility it is to review, action as appropriate, and – in specific circumstances – report harmful or illegal content.
- We use sophisticated technologies – including the use of industry standard hash matching technology to detect known illegal images or videos.
- We are an engaged member of a number of initiatives and member organisations working to combat harmful content, including the Global Internet Forum for Counter Terrorism (GIFCT), the EU Internet Forum, the Internet Watch Foundation (IWF), We Protect Global Alliance (WPGA) and the Tech Coalition.
- We have a trusted flagger programme through which we work with expert organisations such as the UK Counter Terrorism Internet Referral Unit (CTIRU) and Europol to expedite the removal of any terror-related content.
- We have URL sharing agreements with industry partners to more quickly investigate and remove violative material hosted on our service that's been shared on other platforms.

ii) Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: no

## Content moderation (User to User)

**Question 18:**

| | |
|---|---|
| i) | Do you agree with our proposals? |
| Response: | |
| ii) | Please provide the underlying arguments and evidence that support your views. |
| Response:<br><br>We broadly agree with the proposed measures outlined. Specifically, we welcome the flexibility provided to services to meet the measures in a way that is cost effective and proportionate to their needs and assessment of risk. The application and success of content moderation functions is different for every company and based on a range of factors such as size, business model and product.<br><br>We disagree with Ofcom's assessment that the wellbeing of content moderators would only be relevant if it impacted user safety. Ensuring the mental health and wellbeing of moderation teams should be considered as a core component for services that rely on content moderators to support their content moderation function. The mental health and emotional welfare of moderation teams is as important to the effective content moderation as providing access to training and materials. Neglecting this aspect not only undermines the overall health of moderation teams but also puts at risk the quality of content moderation and, consequently, user safety. We would advocate for a more holistic approach that recognises the link between moderator well-being and the success of content moderation functions. | |
| iii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: no | |

## Automated content moderation (User to User)

| **Question 20:** | |
|---|---|
| i) | Do you agree with our proposals? |
| Response: | |
| ii) | Please provide the underlying arguments and evidence that support your views. |
| Response:<br><br>However, we recognise that these services can pose some risk as a result of bad actors seeking to misuse the service functionality to store or share CSAM content and, without appropriate mitigation measures in place, they can be a high risk for such content. We therefore believe that Ofcom's proposal that such services should use hash matching to detect CSAM is appropriate.<br><br>In terms of our approach to CSAM, we use a variety of tools and processes including hash-matching, URL sharing agreements, trusted flagger programme, user reports, and human review, as well as working with industry initiatives, to swiftly find potentially violating content and action | |

it as appropriate - please see question 16 for more details on how we find and address illegal content.

We have not commented on URL detection and fraud keyword detection as we do not offer our users the option to search URLs on our service, and our functionality and the fact we don't have ads on the platform means that the types of fraudulent content defined in the OSA are not prevalent on our service.

| | |
|---|---|
| iii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |

Response: no

| **Question 21:** | |
|---|---|
| i) | Do you have any comments on the draft guidance set out in Annex 9 regarding whether content is communicated 'publicly' or 'privately'? |

Response:

Ofcom's decision to apply these proposals only in relation to content communicated publicly is a welcome one. It will help to ensure the technical feasibility of the Act, as well as proportionality.

| | |
|---|---|
| ii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |

Response: no

***Do you have any relevant evidence on:***

| **Question 22:** | |
|---|---|
| i) | Accuracy of perceptual hash matching and the costs of applying CSAM hash matching to smaller services; |

Response:

| | |
|---|---|
| ii) | Please provide the underlying arguments and evidence that support your views. |

Response:

Hash matching is a key part of our content moderation programme. However, whilst the technology is effective, it has notable shortcomings. Threshold sensitivity is a particular challenge. Hash-matching technology relies on sensitivity thresholds to detect "perceptual" matches - aka "fuzzy matches" - which aim to find images close to the original image but which in practice can - and do with sufficient regularity - result in false positives. This is why human review is a vital component of our content moderation programme. Human review is critical to ensuring the accuracy of automated tools, protecting the privacy rights of users, and vetting external or user reports of potentially violative content.

Whilst automated tools can prove highly useful to help content safety teams identify illegal content and prioritise action, they do not provide a complete solution, nor will they be equally effective for every company. The application and success of such tools is different for every company and based on a range of factors such as size, business model and product. Furthermore, the costs associated with a detection programme (both automated and human) can be extensive and ongoing, covering everything from technical costs (both third-party and in-house), to acquisition and quality control of ingested hash sets.

As outlines in more detail in response to question 18, it is also incumbent on services which rely on content moderators to invest in adequate support to maintain the resilience, mental and emotional wellbeing of the reviewing team. Maintaining wellness has a cost associated to establish and maintain the programme, including technical tools to reduce the impact of certain images during the reviewing process and the provision of trauma-informed wellness consultations with external licensed providers.

iii) Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: no

| **Question 26:** | |
|---|---|
| i) | 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. |
| Response: | |
| ii) | Please provide the underlying arguments and evidence that support your views. |
| Response:<br><br>As Ofcom is already aware, the challenge of assessing context with regards to potentially illegal content is formidable. The additional layer of context needed to assess the legality of a particular piece of content is almost never available to us. Our ability to review is limited to the content itself, so it is very difficult, if not impossible, to identify the reason why a user would have a given piece of content in their account. For instance, an academic or journalist might want to store, for perfectly legitimate reasons, material that might otherwise be regarded as illegal. Given the lack of context we have, we note the risks that over-removal could pose to users' privacy and freedom of expression.<br><br>There are no perfect solutions to this problem, however we believe that a combination of hash matching, human review, and effective reporting and complaints systems (including for complaints from users who feel their content has been unfairly removed) can help services to strike the delicate balance between protecting user rights and detecting illegal content. We note however that the lack of global definition around what terror content is can mean the available terrorism hash sets can be either quite limited - or limiting - depending on a company's policy definition of terror content.<br><br>As with CSAM detection, the costs associated with a detection programme (both automated and human) can be extensive, as they must cover both technical and team-related costs. | |
| iii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: no | |

## Terms of service and Publicly Available Statements

| **Question 29:** | |
|---|---|
| i) | Do you agree with our proposals? |
| Response: | |
| ii) | Please provide the underlying arguments and evidence that support your views. |
| Response: | |

| |
|---|
| Ofcom's proposals here are sensible. Our Terms of Service, Privacy Policy and Acceptable Use Policy are easily available to the general public on our website. All of our communications, including our Terms of Service, Acceptable Use Policy, and materials on the our Help Center, are written to suit the lowest possible reading age to ensure clarity and accessibility for all. We partner with a third-party accessibility testing service to ensure we deliver on our commitment to inclusion at all times, measuring against a set of standards set by the Web Content Accessibility Guidelines (WCAG). |
| iii) Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: no |

## User access to services (U2U)

| **Question 40:** |
|---|
| i) Do you agree with our proposals? |
| Response: |
| ii) Please provide the underlying arguments and evidence that support your views. |
| Response:<br><br>We agree with the general principles of these proposals. However, with regards to the proposed measure to address accounts operated by or on behalf of a terrorist group or proscribed organisations in the UK, we would like to reiterate our concerns around the difficulty of assessing context with regards to potentially illegal content. As outlined in question 26, such context is often not readily available to services like ours, where content is located within a private space without the additional contextual information provided by features such as personal profiles, comments, or reposts. This lack of context makes it extremely difficult to assess the risk without jeopardising users' privacy and freedom of expression, so it is important that services seek to strike the right balance between effective content moderation and protecting user rights via robust detection and complaints processes. |
| iii) Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: no |

***Do you have any supporting information and evidence to inform any recommendations we may make on blocking sharers of CSAM content? Specifically:***

| **Question 41:** |
|---|
| i) 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)? |
| Response: |

| ii) | What are the advantages and disadvantages of the different options, including any potential impact on other users? |
|---|---|

Response:

Our Terms of Service clearly state that we can review the conduct and content of our users for compliance with our Acceptable Use Policy, which prohibits a range of activity including the storing, publishing, or sharing of illegal material such as CSAM and terrorist content. Those documents also explain to our users that, in response to violations of our policy, we will take appropriate action, including as removing or disabling access to content, suspending a user's access to our services or terminating an account. In the case of apparent CSAM, for example, we take immediate action by freezing the user's account and disabling all shared links that user has created.

We believe that these actions are appropriate in response to violation of our Acceptable Use Policy, however we also have systems in place to account for potential false positives. Users who feel their content has been unfairly removed can appeal to explain why the content does not violate our terms and conditions, or why they have a legitimate reason to possess it. Depending on the nature of the content, the potential restrictions placed on their account, and the information available to us, the content may be re-reviewed and reinstated.

Our robust set of policies and processes, including disabling users who have been found to be in violation of our Acceptable Use Policy, helps to ensure that, whilst there is some risk of illegal content on our service, this risk is minimised at every possible junction and addressed quickly and effectively when encountered. Please see question 16 for further information on our policies and processes which minimise illegal content.

| iii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
|---|---|

Response: no

| **Question 42:** | |
|---|---|
| i) | How long should a user be blocked for sharing known CSAM, and should the period vary depending on the nature of the offence committed? |

Response:

Child sexual exploitation and abuse has no place on our service, and when we become aware of it, we swiftly take action to disable the account and prevent the content from being shared.

| ii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
|---|---|

Response: no

***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.***

| **Question 43:** | |
|---|---|
| i) | 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? |

Response:

Whilst automated tools can prove highly useful to help content safety teams identify illegal content and prioritise action, they do not provide a complete solution, nor will they be equally effective for every company. This is why human review is a critical component of content moderation programmes, to ensure accuracy of automated tools, protect the privacy rights of users, and vet external or users reports of potentially violative content. However, even with human review it can be incredibly difficult to establish context when reviewing potentially illegal content which can pose a threat to users' privacy and freedom of expression. This is why we also have systems in place whereby users who feel their content has been unfairly or incorrectly removed can appeal to explain why the content does not violate our terms and conditions, or why they have a legitimate reason to possess it.

There are no perfect solutions to this problem, however we believe that a combination of automated detection, human review, and effective reporting and complaints systems (including for appeals) can help to strike the right balance between safety and preserving user rights. As outlined in question 42, in cases of apparent CSAM on our service, the user's access to the account is disabled. We believe an immediate action to disable the user's account protects victims more effectively than a strike-based system given the severity of the crime committed, and the risk that the user would reoffend if allowed back onto the service.

| | |
|---|---|
| ii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |

Response: no

# Volume 5: How to judge whether content is illegal or not?

## The Illegal Content Judgements Guidance (ICJG)

| **Question 51:** | |
|---|---|
| i) | What do you think of our assessment of what information is reasonably available and relevant to illegal content judgements? |
| Response: Please see our response to question 26. | |
| ii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: no | |

# Volume 6: Information gathering and enforcement powers, and approach to supervision.

## Information powers

| **Question 52:** | |
|---|---|
| i) | Do you have any comments on our proposed approach to information gathering powers under the Online Safety Act? |
| Response: | |
| ii) | Please provide the underlying arguments and evidence that support your views. |
| Response: | |
| iii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: | |

## Enforcement powers

| **Question 53:** | |
|---|---|
| i) | Do you have any comments on our draft Online Safety Enforcement Guidance? |
| Response: | |
| ii) | Please provide the underlying arguments and evidence that support your views. |
| Response: | |
| iii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: | |

# Annex 13: Impact Assessments

| **Question 54:** | |
|---|---|
| i) | Do you agree that our proposals as set out in Chapter 16 (reporting and complaints), and Chapter 10 and Annex 6 (record keeping) are likely to have positive, or more positive impacts on opportunities to use Welsh and treating Welsh no less favourably than English? |
| Response: | |
| ii) | If you disagree, please explain why, including how you consider these proposals could be revised to have positive effects or more positive effects, or no adverse effects or fewer adverse effects on opportunities to use Welsh and treating Welsh no less favourably than English. |
| Response: | |
| iii) | Is this response confidential? (if yes, please specify which part(s) are confidential) |
| Response: | |