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:

Refuge welcomes the opportunity to respond to this consultation. Our expertise is as a specialist domestic abuse organisation, providing support to thousands of women and children every day. In recent years, there has been a huge growth in technology-facilitated domestic abuse, and in 2017 Refuge established a tech-facilitated abuse team to directly support survivors of this form of abuse, and to stay abreast of the changing ways perpetrators use technology to abuse.

We wish to preface our response by stating that the length and complexity of the consultation has impacted our ability to respond in a comprehensive manner. We refer to the joint submission made by the VAWG Sector for further detail and strongly urge Ofcom to adopt the applicable recommendations made in the model VAWG Code of Practice (co-produced by a coalition of experts in women and girls' online safety) within the illegal Code of Practice.

Our overarching view of volume 2 is that the volume goes some way towards building an understanding the direct harms experienced by women and girls online and generally presents good evidence of the different forms of harm that make up the priority illegal offences. It is positive to see the impact of online VAWG on individuals reflected in volume 2, and we welcome the steps Ofcom has taken to outline research and evidence of these harms.

We recommend where specific sections of volume 2 can be strengthened below in section 1ii, but we believe further consideration needs to be given throughout the volume to the **wider societal harms** caused by online violence against women and girls (VAWG) and the costs associated with women coming offline and self-censoring. There is widespread evidence of the costs to society as a result of VAWG and online VAWG. Refuge research has shown that, as a result of tech-facilitated domestic abuse ('tech abuse'), many women are reducing their online presence, or coming offline entirely. 38% of survivors said they felt unsafe or less confident online as a result of the tech abuse on social media. 30% of survivors supported by our tech abuse team also said the tech abuse left them unable to use their devices. Women's Aid research has shown that children and young people exposed to misogynistic social media content like Andrew Tate's were almost 5x more likely than those not exposed to view hurting someone physically as acceptable if you say sorry afterwards. And as a recent study found, "Harmful views and tropes are now becoming normalised among young people…Online consumption is impacting young people's offline behaviours, as we see these ideologies moving off screens and into schoolyards." These forms of harm on a societal level are not sufficiently analysed in volume 2. In comparison, Ofcom outlines the risks associated with the glamourisation of firearms and that this is 'of particular concern.' Greater reflection of the wider harms to society caused by online VAWG in volume 2 - alongside examination of the impact on individuals already included - is required to illustrate the true scale and devastating impact of VAWG.

In addition, we are disappointed that Ofcom often cites in volume 2 that 'no specific evidence was found on how **business models** may influence risks of harm to individuals' from offences such as coercive control and intimate image abuse. A number of press investigations have found that

social media companies are profiting from a wave of misogynistic content creators. We therefore find it surprising that Ofcom has concluded there is no evidence in these cases. If such evidence does not meet Ofcom's threshold, this should be explained, and Ofcom should subsequently prioritise evidence-gathering on this issue via its information-gathering powers.

ii) Do you think we have missed anything important in our analysis? Please provide evidence to support your answer.

Response:

As referenced in response to question 1i, we recommend that there be further analysis of the societal costs and impacts of online VAWG, including the impact on women of self-censorship. Ofcom's own research has illustrated that women feel less able to share their opinions and feel free to be themselves online. Refuge's Unsocial Spaces report also found that 38% of women who experienced abuse on social media from a partner or former partner said they felt unsafe or less confident online as a result.

Whilst there is some recognition of the harms caused by **deepfake intimate images**, volume 2 must go further in outlining the harms and manifestations of deepfake intimate images. Research has shown that the most common deepfakes shared on the internet are non-consensual sexual depictions of women, yet the chapter on intimate image abuse is quite limited on this – in contrast to a lengthier explanation on the impact of deepfakes in the chapter on foreign interference offences. More broadly, we urge Ofcom to keep a close eye on developments in AI and genAI – the consultation notes that the Register may be expanded in future given advancements in genAI and that Ofcom has only 'considered this technology in a limited manner'. Unpublished data from **Refuge's upcoming UK National Tech Abuse Survey** reveals emerging trends relating to AI and deepfakes: **[CONFIDENTIAL✂]**

Refuge warmly welcomed the addition of **coercive control** as a priority illegal offence in the Online Safety Act. Chapter 6G of volume 2 provides a good start to understanding this insidious offence and its manifestations online. However, we strongly recommend further analysis is outlined in this chapter, particularly on the gendered nature of risk, on identifying online coercive control, and on content which encourages and glamourises coercive control (such as content from Andrew Tate). We also encourage Ofcom to continue to review evidence of the impact of coercive control and other forms of online VAWG on survivors' physical safety. Online coercive control will often be occurring concurrently with 'offline' coercive control, and this continuum of abuse should be better recognised within volume 2. We welcome the use of Refuge evidence within chapter 6G, and suggest additional evidence that could be incorporated in this chapter below:

- Refuge Marked As Unsafe research found that, of the survivors we interviewed who had experienced tech-facilitated domestic abuse on social media, 59% said they had experienced coercive control via social media.
- An Opinium survey for Refuge found that the second most commonly experienced type of behaviour suggestive of coercive control experienced by young women was having social media accounts monitored (26%, sample size: 1,010 16–19-year-olds).
- Whilst the police do not collect data on the prevalence of online coercive control, we suggest that including police data on coercive control more generally, and the year-on-year growth in offences recorded by the police, would help paint a picture of the potential scale of online coercive control (noting that this is likely to be the tip of the iceberg given only 1 in 5 survivors of domestic abuse report to the police). In the year ending March 2023, there were 43,774 offences of controlling or coercive behaviour, compared to 41,039 in the year ending March 2022 (ONS).

- Further research to highlight the intersection between coercive control and intimate image abuse and threats/harassment can be provided: 43% of women surveyed for Refuge's Naked Threat report experienced coercive and controlling behaviour in addition to threats to share intimate images, and 39% experienced emotional abuse, highlighting the broader background of emotional manipulation and control abusers seek to instil.
- As a point of clarity, we recommend references to Refuge research are checked for the appropriate terminology, and that either 'domestic abuse survivor' or 'tech abuse survivor' are used. In some footnotes, it is incorrectly suggested that data is for all survivors of domestic abuse, rather than those who have experienced tech abuse (i.e. footnotes 746, 749, 760, 763, 768, 773).

We agree with Ofcom that user base demographics should be considered a risk factor for domestic abuse offences such as intimate image abuse, but we would like to see further consideration given to the heightened risks of harm among certain users, such as for survivors with disabilities, and South Asian women experiencing intimate image abuse, and further consideration of intersectionality.

As referenced in Q1i, further analysis should also be included on how providers' **business models** drive and enable online VAWG and misogyny to flourish. We also recommend Ofcom conduct an examination of the popularity and impact of misogynistic influencers such as Andrew Tate in encouraging and glamourising coercive control, and the interconnection with platform profits and recommender systems.

Further specific recommendations on additional research and evidence that should be considered are included below, categorised by chapter:

- 6E Harassment, stalking, threats and abuse offences: We welcome the recognition that perpetrators often use multiple fake, anonymous accounts to perpetrate abuse. This is a common tactic of domestic abuse perpetrators. In addition to the evidence outlined, we refer to an interview with the former head of Trust and Safety at Twitter (now X) who highlighted Twitter's awareness of multiple account detection and returning accounts: "*If you're a multiple-time violator, how do we make sure you stop? … instead, we will do nothing in that realm and instead come up with a new product feature." Because it was growth at all costs, and safety eventually*." It should also be recognised that survivors of domestic abuse will sometimes make use of anonymous account to protect their identify and safety. Robust action should be taken against fake accounts reported for abusive behaviour.
- 6M Intimate Image Abuse:
    - 6M.6 – This section references new offences under section 66(B) of the Sexual Offences Act brought into force by the Online Safety Act. The list should be amended to also include threats to share intimate images/films.
    - 6M.14 – Further analysis of 'collector culture', where groups of users anonymously exchange and swap intimate images of women without their consent, would be of benefit here, to ensure this growing trend is well-examined and understood by platforms.
    - 6M.16 – The statement that threats to share intimate images are predominantly perpetrated either as part of a wider pattern of domestic abuse or coercive control, or for financial gain, is binary. Our experience is that perpetrators of domestic abuse will also control survivors' finances and use intimate images to blackmail and threaten survivor's financial income. For example, perpetrators have threatened to share intimate images with a survivor's employer.
- 6S Cyberflashing: We recommend including research findings that suggests nearly half (48%) of those aged 18 to 24 received a sexual photo they didn't ask for in the last year alone.

- 6T Search services: We recommend analysis of the risk of harm of VAWG from search services be included in this chapter. There is currently very little evidence presented on this. For example, we would expect analysis relating to the role of search engines in promoting deepfake and nudification websites and providers such as **[REDACTED ✂].**
- 6U Governance, systems and processes:
    - 6U.42 Recommender systems – We recommend this section is expanded to also consider how platform-made choices about recommender and algorithmic systems impacts the spread of online misogyny, considering influencers such as Andrew Tate. For example, ISD research has suggested that platforms give greater visibility to abusive hashtags over non-abusive hashtags, and that 'Tradwives are able to "adapt their content" to exploit algorithmic feeds…while promoting anti-feminist and anti-LGBTQ+ belief systems.'
    - 6U.56 Speed of action – We have further evidence to support the statement that where users perceive a lack of action on a report, or an absence of any action at all, they are less likely to report in the future. Our Marked As Unsafe research found that after their experience of reporting content, 41% of interviewed survivors of tech abuse on social media said they were unlikely to report content again. Fewer than 1 in 5 survivors (18%) said they would be very likely to report again in future'.

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

Response: Yes – unpublished data from Refuge's UK National Tech Abuse Survey.

Question 2:

i) 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.

Response:

We would like to see further analysis of the interconnection between intimate image abuse, controlling and coercive behaviour, harassment, stalking, threats and abuse, within a domestic abuse context. Whilst references are made within different chapters of volume 2 to other chapters sections – for example via a statement in chapter 6G on coercive control that 'a case of CCB might include cyberstalking, harassment and threats of violence, intimate image abuse' – more could be done to clearly outline the links between these forms of harm. It is Refuge's experience that perpetrators of domestic abuse will use any means to control and abuse the survivor, and they will frequently not limit themselves to one form of abuse. In addition, we often find that tech companies have a limited understanding of tech-facilitated domestic abuse, and our specialist tech team spend a substantial amount of time explaining the severity, impact and nuances of tech abuse to tech companies, for example via content moderators, trust and safety teams and trusted flagger pathways. To the untrained eye, tech abuse can often be hard to recognise without an understanding of the broader context of domestic abuse and coercive control. It is therefore vital that volume 2 provides tech companies with as in-depth an understanding of tech abuse and other forms of online VAWG, and their interconnection, as possible.

ii) 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 3:	
i)	Do you agree with our proposals in relation to governance and accountability measures in the illegal content Codes of Practice?
Response:	
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:	

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

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

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

Service's risk assessment

Question 7:	
i)	Do you agree with our proposals?
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:	

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

Question 9:	
i)	Are the Risk Profiles sufficiently clear?
Response:	
ii)	Please provide the underlying arguments and evidence that support your views.
Response:	
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:	

Record keeping and review guidance

Question 10:	
i)	Do you have any comments on our draft record keeping and review 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:	

Question 11:	
i)	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?
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:	

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:

Whilst we appreciate that the Codes of Practice will be iterative and that Ofcom has developed this first Code without its information-gathering powers, we believe that the draft illegal content Code of Practice must go much further. The Code needs to change industry practice and deliver a noticeable impact for platform users, and it is a missed opportunity to just recommend and set low, **base-level minimum standards**, which most providers seem to comply with already. In the development of the Codes, emphasis seems to have been placed on recommending measures which are already in place and in use by platforms, perhaps in order to minimise cost and burden on providers. For example, the Code recommends that certain providers should ensure that hash matching technology is used effectively to analyse relevant content to assess whether it is Child Sexual Abuse Material (CSAM). However, we understand that many platforms already use such technology and have done so for many years. In addition, the only specific recommended measure that large platforms at medium or high risk of coercive control occurring on their sites should introduce is to offer every user options to block or mute other user accounts on the service. Again, such tools are already in common use across many platforms, leading us to question how far such measures will go to create tangible improvements in safety for survivors of tech abuse. We question whether this minimal approach to the Codes will lead to the "strong protection for women and girls" Ministers pledged the Act would deliver. One of the reasons the Online Safety Act is so crucial is because the measures that industry are currently using have not sufficiently protected users. Therefore, an overdue focus on measures already in use by large swatches of the tech industry will be unlikely to deliver the systemic change required to deliver the objectives of the Act.

There also appears to be a **disconnect** between the measures presented in the Codes and the severity, prevalence and risks associated with online VAWG harms presented in volume 2. A good level of detail and evidence on the types of functionalities that cause harm for women and girls are included in volume 2, but this does not seem to feed through to recommendations in the Codes. In addition, no methodology is provided for the development of the Codes and the evidential threshold Ofcom appears to have set to make recommendations seems very high. It also appears throughout volume 4 that preferential weighting is given to evidence collected from industry and "best practice" from tech companies. Governance and risk assessment proposals take at face value evidence from providers that they are "doing much of this already." We urge Ofcom to make the Codes much more aspirational and to seek to raise safety standards higher across the tech industry via the illegal content Code.

In addition, we would like to see further measures included in the Code that focus on **safety by design** and prevention of online VAWG. Despite clause 1 of the Online Safety Act highlighting that duties imposed on providers by the Act seek to secure services that are safe by design, the focus of the illegal content Code appears to be on mitigation after harms have occurred and on takedowns. Whilst this is a crucial piece of the puzzle – and survivors of tech abuse often state that a prompt response from social media companies when abusive content is posted by their

partners or ex-partners is a priority – tech companies must make further 'upstream' changes to their platforms to truly ensure the design and production of safer online spaces. We question whether an approach which focuses mostly on takedowns, user control over default settings and downstream mitigation of harms will be sufficient to ensure user safety becomes a basic design consideration in the development of any new product/tool/feature. The government's own principles of safer online platform design emphasise the importance of a preventative approach. Whilst this might generate pushback from certain parts of the tech industry as more 'disruptive', such an approach would be much more in keeping with the approach promised during the passage of the Online Safety Bill through Parliament.

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:

We have responded to Q13i and Q13ii jointly.

We do not agree with the proposed approach which differentiates heavily between large/multi-risk services and small single-risk services. Only very light measures are placed on single-risk small services, yet such services can include extremely harmful sites which have been solely created for a single harmful purpose i.e. platforms dedicated to 'collector culture' which enable and support the collection of intimate images shared without consent and platforms which are dedicated to deepfake abuse. Many of the measures – including board or governance oversight of risk management – only apply to 'large' companies. Ofcom has pointed to the few Code measures that will be imposed on smaller services, but these are limited to CSEA and use of 'hash-matching' to detect such images. We do not believe that scale of use is the sole indicator of risk. Most platforms can be used for domestic abuse purposes, and perpetrators can use less 'obvious' platforms to fly under the radar. Platforms such as Telegram are often used as forum where perpetrators share intimate images and post requesting for more images of certain women. We are concerned that this regulatory loophole may encourage the creation of new, small platforms for harmful purposes, which effectively avoid more stringent regulation. To not even recommend smaller services (low risk and specific risk) adopt basic safety measures such as training for and appropriate resourcing of staff working in content moderation is disappointing.

As stated in our response to the call for evidence on categorisation, the perpetrator of the 2021 Plymouth shootings, Jake Davison, is known to have visited smaller incel forums after he was banned from Reddit in the days preceding the shooting. Whilst our Unsocial Spaces research found that the majority of tech-facilitated domestic abuse ('tech abuse') survivors experienced abuse from a partner or former partner on a Facebook (now Meta)-owned platform, which are among the largest platforms in terms of user base, we have supported women who have been subject to abuse on much smaller platforms with a relatively small user base.

In addition, we point to Glitch's Digital Misogynoir report which highlights concerns about smaller, high harm platforms such as 4chan and Gab, and how 'hateful rhetoric and jargon is trickling from the alternative platforms (Gab, 4chan) to the mainstream ones (Twitter [now X], Instagram, and Facebook)'. Glitch's analysis of 4chan and Gab deemed these sites to be some of the most 'toxic'

in terms of misogynoir and noted that hateful jargon from alternative platforms steadily moves to more mainstream sites.

Ofcom's approach therefore potentially lets some harmful or risky small companies escape much-needed regulatory obligations, whilst seeming to prioritise company revenue. Volume 4 specifically states that 'more onerous expectations on smaller services' may have implications for cost, competition and innovation.

We recommend Ofcom reconsider its overarching approach to categorisation to ensure smaller harmful platforms do not avoid much-needed safety measures. If Ofcom proceed with the current proposals, we wish to emphasise the importance of a robust and speedy mechanism to assess risk-level of each service and to 'demote' or 'promote' services as necessary into different categories.

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: No	

Question 14:	
i)	Do you agree with our definition of large services?
Response: We have responded to Q14i and Q14ii jointly. We do not agree with the high threshold used in the definition of large services of an average user base greater than 7 million per month in the UK, which is approximately equivalent to 10% of the UK population. The 7 million user base threshold would not include platforms such as Fortnite and Roblox. Whilst a platform may not be used widely by the general population, and so therefore may not meet the large service definition, the platform may be highly used by one particular group, such as women or children.	
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: 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:	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 16:	
i)	Do you have any comments on the draft Codes of Practice themselves?
Response: Please see our response to question 12 and individual responses to proposals from question 18 onwards.	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 17:	
i)	Do you have any comments on the costs assumptions set out in Annex 14, which we used for calculating the costs of various measures?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Content moderation (User to User)

Question 18:	
i)	Do you agree with our proposals?
Response: We have responded to Q18i and Q18ii jointly. As many survivors experience issues with content moderation and reporting/complaints procedures concurrently, and given the two systems interact closely, our response to this question overlaps with our response to question 28i. Whilst the broad principles of the proposals on content moderation for User to User services are welcome, Refuge believes the measures need to be much more tightly defined to produce tangible changes. For example, the requirement for content moderation systems or processes to be designed to take down illegal content swiftly is vague. Ofcom's decision not to set timelines	

within which content should be removed will lead to a lack of clarity, which tech companies could exploit. The term 'swiftly' with respect to designing processes or systems which are designed to take down illegal content at pace will leave room for a range of interpretations. For example, we understand from Glitch that in response to their Digital Misogynoir report, social media companies informed them that they had responded to horrific cases of misogynoir and misogynistic content by 'de-amplifying' content, rather than removing it. We would ideally like to see minimum standards set for timelines for acknowledging reports and responding to reports. Many survivors are currently waiting weeks, months or even years for a response to flagged domestic abuse-related content. Over half (53%) of survivors interviewed for our Marked As Unsafe report did not even receive a response from the platform to their report. Whilst we acknowledge that some degree of flexibility may be needed to ensure content moderation decisions are not made incorrectly due to time pressures, a continuation of the current situation would be untenable. At the very least, systems should be put in place to inform users of the status of their report. Our tech team report that survivors rarely receive an update when platforms have decided not to progress a report, leaving them in limbo, unaware that no action will be taken on content they have reported.

We also urge Ofcom to provide further specificity on what the performance targets for content moderation functions should look like for different types and sizes of platforms. Allowing platforms to decide their own targets is meaningless. At a minimum, we would strongly encourage Ofcom to look at robust processes when developing guidance on transparency reporting to ensure such targets are reviewed, monitored closely and published.

Similarly, we would like to see further direction from Ofcom on the training and materials that staff working in content moderation must receive. What forms of harm will be included in this training? Will this be delivered in-house or sourced externally from experts? We recommend training is delivered externally by experts, and that training should include slang and alternate harmful meanings for seemingly innocuous words, as well as terms used in different languages that are harmful, but may not be immediately understood as so. Many platforms rely on third party moderation services by outsourcing content moderation work – it is vital that such outsourced processes and systems are also included within regulatory requirements on content moderation and elsewhere. A specific statement to this effect within the Code of Practice would be useful.

We welcome the inclusion of a measure relating to prioritisation process for reviewing content. Domestic abuse and online VAWG must be included as key content which should be prioritised by content moderators. It may also be pertinent to include the estimated age of the user/depicted person as a measure. The assessment of 'severity of content' to determine prioritisation could be clarified – 'severity' may leave room open to interpretation. Is this intended to measure potential harm to the user? In addition, we do not necessarily agree that 'virality' of content should be a prioritisation metric above a potential 'harm' metric. For example, for content or images shared that may raise the risk of so-called 'honour'-based abuse, content may only need to be shared a couple of times within the community to place the survivor at a risk of harm. Yet such content may not be prioritised by the virality measure.

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: No	

Content moderation (Search)

Question 19:
i) Do you agree with our proposals?
Response: We have responded jointly to Q19i and Q19ii. Whilst we do not comment directly on the proposals for content moderation on Search services, we wish to highlight a concern relating to domestic abuse, online VAWG and search platforms to Ofcom. **[REDACTED ✂]**
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: No.

Automated content moderation (User to User)

Question 20:
i) Do you agree with our proposals?
Response: We have responded to Q20i and Q20ii jointly. Given the use of hash matching has been recommended to proactively identify CSAM material, we question why hash matching has not also been recommended for identifying adult intimate image abuse content. StopNCII is a free global tool to protect intimate images from being shared online by perpetrators of intimate image abuse. Major social media companies are already using StopNCII, including Facebook, Instagram, TikTok, Bumble, OnlyFans, Reddit, Aylo, Threads and Snap Inc. Since its launch in 2021, StopNCII.org has received over 434,000 hashes and over 182,000 adults across the world have created cases to protect their intimate images from being shared by perpetrators (see evidence here), proving it is a viable and effective tool. We urge Ofcom to recommend platforms implement hash matching for intimate image abuse, as well as CSAM.
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: 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:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

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

Question 23:	
i)	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;
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:	

Question 24:	
i)	Costs of applying our CSAM URL detection measure to smaller services, and the effectiveness of fuzzy matching for CSAM URL detection;;
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:	

Question 25:	
i)	Costs of applying our articles for use in frauds (standard keyword detection) measure, including for smaller services;
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:	

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

Automated content moderation (Search)

Question 27:	
i)	Do you agree with our proposals?
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:	

User reporting and complaints (U2U and search)

Question 28:	
i)	Do you agree with our proposals?

Response:

Due to the nature of the experiences of the survivors we have supported, our response to this question overlaps with our response to question 18i. We have responded jointly to questions 28i and 28ii.

Refuge argues that there should ideally be a dedicated reporting channel for violence against women and girls, similar to the proposed recommendation for a dedicated channel for fraud. This channel should be staffed by trained specialist employees who are skilled in understanding and identifying the many different and nuanced forms of online VAWG and in supporting survivors of VAWG. Such a reporting channel could also help capture emerging harms and trends in perpetrator behaviour, which would better enable tech companies to build in further safety measures and ensure training for staff is kept up to date. However, we recommend Ofcom consult with specialist VAWG organisation in developing guidance for the creation of such reporting channels. Such channels must be appropriately resourced – to avoid the re-creation of micro "gender" or "sexual abuse" teams within large companies, who have to answer all organisational

questions on VAWG with very limited capacity. Lessons can also be learnt from the function of trusted flagger pathways – which can be valuable, but still have long waiting times.

Consideration of the link between fraud and VAWG vis a vis reporting channels should also be given, given the rise in romance fraud.

Whilst it is promising to see the recommendation that complaints systems and processes should be easy to find, easy to access and easy to use, this measure again lacks specificity. Refuge's tech team has experienced barriers to contacting social media companies on behalf of domestic abuse survivors, due to a failure by companies to provide easily accessible contact details. For example, Snapchat did not provide contact details or transparent information about where users can find support, making it very difficult to report harmful content and escalate reported content. We recommend that further recommendations are put in place by Ofcom to ensure reporting/complaints systems are easy to use in a range of languages, to ensure processes are truly accessible and in line with best practice for accessibility (see Government Web Content Accessibility Guidelines). In addition, we would like to query what consideration Ofcom has given to recommendations relating to different methods of contact for making complaints for users that may not be able to engage with a particular process/format. For instance, could alternative methods be put in place for users to speak on the phone with a content moderator/complaints handler if the individual struggles to use forms on websites, due for example, to disability?

Lastly, we would like to see further measures put in place to ensure tech companies are cooperating more closely with the police in order to obtain digital evidence. Survivors supported by Refuge have often said that they wish social media companies would have been more forthcoming with providing evidence of domestic abuse to the police, particularly given this is often data concerning the survivor and therefore data to which they are entitled to view under GDPR Right to Access. Tech companies should work hand-in-hand with the police and with other companies to protect their users and hold perpetrators to account by expediting collection of data and evidence to prosecute perpetrators. Greater collaboration with the police should not been seen by tech companies as a way to absolve themselves of taking action on online harm. For example, we have experienced issues with WhatsApp when reporting the non-consensual sharing of intimate images, the response from WhatsApp is frequently that a local police enforcement needs to be contacted. This can mean the images continue to be distributed whilst awaiting a police response, which is often much slower. Training and support on online VAWG which enables staff to respond effectively to tech abuse may also support greater cooperation with the police on, as staff are likely to be more easily able to identify and understand VAWG offences.

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: No	

Terms of service and Publicly Available Statements

Question 29:	
i)	Do you agree with our proposals?
Response:	
We have responded jointly to Q29i and Q29ii.	
Refuge welcomes the proposed measures regarding clarity and accessibility of terms of service/Publicly Available Statements and requirements for terms to outline how users are protected from illegal content, information on proactive technology used for compliance with the illegal content safety duties, and the policies and processes that govern the handling and resolution of complaints. We regard the inclusion of this information in an easily accessible and understood format to be the very minimum companies should be doing to ensure their users are aware of the safety measures on their platforms. Providers will need to set out in their terms of service how they will minimise the length of time priority illegal harms are present on their sites. We recommend platforms should also be required to set expectations for users on the approximate timelines associated with this measure.	
Transparency around the use of AI in content moderation is key. Often, the harmful content that perpetrators of domestic abuse share/send to survivors is contextual, nuanced, and may require additional explanation and understanding of domestic abuse. Refuge's tech team will provide this context as part of a two-way conversation with social media platforms via trusted flagger pathways. Our concern is that AI often misses nuances in relation to domestic abuse, tech-facilitated domestic abuse, and culturally specific cases of VAWG, and a solely AI-driven content moderation system would not respond appropriately to content and reports/complaints. It is likely that AI moderators would fail to identify the nuances and contextual nature of tech abuse given its highly subjective nature. Refuge therefore recommends that there should always be an element of human oversight in content moderation.	
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: No	

Question 30:	
i)	Do you have any evidence, in particular on the use of prompts, to guide further work in this area?
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:	

Default settings and user support for child users (U2U)

Question 31:	
i)	Do you agree with our proposals?
Response: We welcome the requirement that supportive information is provided to children using a service to make informed choices about risks, access to safeguarding processes and support in a timely and accessible manner. This measure needs to work hand-in-hand with robust and improved age verification processes.	
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: No	

Question 32:	
i)	Are there functionalities outside of the ones listed in our proposals, that should explicitly inform users around changing default settings?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 33:	
i)	Are there other points within the user journey where under 18s should be informed of the risk of illegal content?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Recommender system testing (U2U)

Question 34:	
i)	Do you agree with our proposals?
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:	

Question 35:	
i)	What evaluation methods might be suitable for smaller services that do not have the capacity to perform on-platform testing?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

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.

Question 36:	
i)	Are you aware of any other design parameters and choices that are proven to improve user safety?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Enhanced user control (U2U)

Question 37:	
i)	Do you agree with our proposals?
Response:	

We agree that users should be give more control over their account and be able to block accounts and disable comments. However, there should not be an undue reliance or onus placed on users taking action against online abuse – as stated in response to Q12i, further requirements must be placed on tech companies to implement upstream prevention and adopt a safety by design approach to their entire platform. Features such as blocking and disabling comments are also already commonly offered by major social media companies. Yet in order for these measures to work effectively, these features need to be accessible. Navigating settings menus, or even locating where to start with this, particularly on a smartphone/app view with reduced screen space, can be challenging. For example, the settings and privacy tabs on Facebook are different for Android, iPhone and web browsers, with some menu options being listed under completely different headings. Providers should also be required to ensure settings relating to user control features are are easy to find and easy to use, and provide clear, straightforward guidance in multiple languages and in accessible formats about what each feature will do and how to activate/deactivate.

We also wish to highlight the need for platforms to support users with disabilities. We are not aware of any platforms that offer voice recognition options for changing settings, for example, to support users without sight.	
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:	

Question 38:	
i)	Do you think the first two proposed measures should include requirements for how these controls are made known to users?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

Question 39:	
i)	Do you think there are situations where the labelling of accounts through voluntary verification schemes has particular value or risks?
Response:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

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

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

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

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

Service design and user support (Search)

Question 44:	
i)	Do you agree with our proposals?
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:	

Cumulative Assessment

Question 45:	
i)	Do you agree that the overall burden of our measures on low risk small and micro businesses is proportionate?
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:	

Question 46:	
i)	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?
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:	

Question 47:	
i)	We are applying more measures to large services. Do you agree that the overall burden on large services proportionate?
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:	

Statutory Tests

Question 48:	
i)	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?
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:	

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

The Illegal Content Judgements Guidance (ICJG)

Question 49:
i) Do you agree with our proposals, including the detail of the drafting?
Response: We have responded jointly to Q49i and Q49ii. We also refer to the points outlined in the joint submission made by the VAWG sector regarding the Illegal Content Judgements Guidance (ICJG), in particular those made by Professor Lorna Woods regarding the current approach's predominant focus on individual pieces of content and the need to take a more systems-based approach in the ICJG: *"I think the signals for inference in the system's context have to be broader … than they might be if you're making a decision whether or not to take content down. So if you're thinking more broadly, about how you weight your algorithm, for example, whether you have nudges about whether this is nice behaviour or not, or whether you have content revenue sharing system, these don't tie in very well with the idea of understanding 'mens rea' and defences very, very tightly in the same way they would do, as in the case of a takedown…I think there's an issue there that in the focus that we see in the consultation on this being more or less about takedown and not about upstream safety by design."* We remain very concerned by the suggestions posed in the ICJG that minimal or no action need be taken by platforms in relation to the reposting of non-consensual sharing of intimate images/videos and cyberflashing. Our understanding is that there will be no obligation on platforms to remove reposts of non-consensual intimate image abuse material i.e. where other users have re-shared the original content/post, and no obligation to act on cyberflashing due to the need to prove the perpetrator's intention. The suggestion made at 26.45 in volume 5 is that the state of mind/mens rea of all users posting shared, forwarded or reposted content would need to be determined 'for the content to be an offence,' and therefore presumably subject to the takedown duty. In addition, Ofcom states that it believes the state of mind requirement is 'unlikely to be able to be reasonably inferred in most cases' of cyberflashing, but does not provide an explanation as to why. There is not sufficient evidence to assume that most perpetrators committing cyberflashing without consent are doing so for non-harmful reasons (and we refer to the separate submission to the consultation made by Professor Clare McGlynn on this point). Whilst Appendix 10 at A10.34 notes that if content is known to be posted without consent 'it should be taken down', we do not believe this is sufficient. Platforms should be obligated to remove all intimate images that are known to be non-consensual. We also wish to highlight questionable wording used in volume 5 26.263 that, in relation to cyberflashing, 'the recipient can, after all, delete it' (the image). This appears to minimise the harms of cyberflashing. We would also suggest that further detail on inchoate offences relating to coercive control and domestic abuse are included. Given the rise of misogynistic content creators such as Andrew Tate, it seems pertinent for Ofcom to provide guidance to platforms on where content that encourages domestic abuse meets the threshold for inchoate offences of encouraging of assisting an offence.

Finally, we welcome the inclusion of paragraph A3.98 in the ICJG in relation to harassment offences, and the recognition that content which does not immediately appear to be problematic/illegal may in fact be domestic abuse. We agree that information and context from the targeted user or on behalf of the user may be needed – however, we would like to see further guidance/requirements for platforms on how this is to be gathered, to ensure survivors are not repeatedly having to explain the context of domestic abuse related content and to relive their trauma. We also suggest it would be most beneficial if contextual information could be raised directly with the platform and that platforms do not solely rely on the trusted flagger pathways and independent advocacy from VAWG specialist organisations for such context, as not every survivor will be receiving support from such an organisation.
ii) What are the underlying arguments and evidence that inform your view?
Response:
iii) Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: refer to confidentiality agreements for the joint submission from the VAWG sector.

Question 50:
i) Do you consider the guidance to be sufficiently accessible, particularly for services with limited access to legal expertise?
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:

Question 51:
i) What do you think of our assessment of what information is reasonably available and relevant to illegal content judgements?
Response:
ii) Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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:	