Consultation response form

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:

UOG:

Virtual Reality hasn't been discussed as a functionality. Immersive headsets increase the psychological realism of what is experienced, and typically enact embodiment, meaning the user feels their virtual body is their own, engaging almost all the senses (visual, audio, phantom touch). Used with online immersive social VR platforms while still being able to maintain anonymity, this also opens the door to perceptually realistic abuses and harms e.g. see the recent media about virtual sexual assaults. Virtual worlds and immersive 3D platforms can also be seen as facilitating parallels of real-world spaces that would be seen as inappropriate for children to occupy, further exposing them to materials, behaviours, and experiences that would be considered "adult only" (e.g., virtual clubs, adults projecting graphic content). These platforms have definably different affordances to social media and streaming/video services that should be considered. And compared to "traditional" online spaces, they likely aren't as monitored/consolidated and therefore harm there may be harder to track and police.

See our papers about child safety in social VR: https://www.cristinafiani.com/papers - in particular:

Fiani, C., Bretin, R., MacDonald, S., Khamis, M. and McGill, M. 2024. *"Pikachu would electrocute people who are misbehaving": Expert, Guardian and Child Perspectives on Automated Embodied Moderators for Safeguarding Children in Social Virtual Reality*. In CHI Conference on Human Factors in Computing Systems (CHI 24), May 11–16, 2024, Honolulu, Hawaii. ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3613904.3642144

Fiani, C., Saeghe, P., McGill, M & Khamis, M. (2024). *Exploring the Perspectives of Social VR-Aware Non-Parent Adults and Parents on Children's Use of Social Virtual Reality*. Proc. ACM Hum.-Comput. Interact. 8, CSCW1, Article 54 (April 2024), 25 pages. https://doi.org/10.1145/3637331

Fiani, C., Bretin, R., McGill, M., & Khamis, M. (2023). *Big Buddy: Exploring Child Reactions and Parental Perceptions towards a Simulated Embodied Moderating System for Social Virtual Reality*. Interaction Design and Children (IDC '23), June 19–23, 2023, Chicago, IL, USA. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3585088.3589374

And our related works on the impact of perceptual realism - Graham Wilson and Mark McGill. 2018. *Violent Video Games in Virtual Reality: Re-Evaluating the Impact and Rating of Interactive Experiences.* In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '18). Association for Computing Machinery, New York, NY, USA, 535–548. https://doi.org/10.1145/3242671.3242684

YL:

It would be good to see more details on economic Impact. Although there is brief touch on fraud and financial loss but expanding on the broader economic impact of online harms, including costs to individuals beyond direct financial loss and to businesses in terms of reputation and trust (indirect loss).

[CONFIDENTIAL✂]

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

Response:

YL:

"Fraud is one of the UK's most common crimes, with around one in sixteen people falling victim a year. The social and economic cost of this crime is vast – the Home Office estimates that fraud costs individuals in England and Wales alone £4.7bn a year. What can sometimes be hidden by these huge sums however is the impact on individual victims, both financially and emotionally. It can be life changing."

https://www.lloydsbankinggroup.com/insights/recognising-fraud-as-an-online-harm.html

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

Response: No

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:

YL:

While discussing technologies/functionalities can pose threats to individuals facing to different kinds of illegal harm, there are some risk factors could be included in the discussion, such as Malware and Ransomware, Botnet, Social engineering (phishing emails) as well as the Identity theft. Generative AI has been mentioned but how the AI development will affect existing risk factors, for instance, from phishing email generation to verification, should be introduced and highlighted.

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:	

UOG:

Tracking and reporting of content should ideally be broadened to encapsulate reporting detected and self-reported harms and safety events on these services (e.g. user usage of safety/moderation tools). Again, from a social VR perspective, I'm not sure these recommendations are targeted enough to enable us to gain insight into what's happening on these platforms currently.

Transparency of what incidents are occurring on these platforms that can be shared for monitoring emerging abuses/harms? Currently these platforms operate as black boxes, and we have to rely on self-reports and ethnographic approaches to understand what is happening, and the extent of the problem. This is quite different to social media, such as Twitter/X for example, where at least researchers have the option of licensing or scraping largely public data to understand what's happening in terms of harms/abuses. I think this will increasingly allow social VR platforms to "fly under the radar" in our estimation of the extent of harms posed here, and what behaviours are occurring.

Records of events that can be taken forward for e.g. legal action. This particularly matters for immersive platforms where events are treated as communication rather than content, meaning there is no digital archival/record/surveillance of what occurred in that space unless the user explicitly chooses to record/capture the session. This means that evidencing harms that have occurred (e.g. for legal action) is problematic.

[CONFIDENTIAL✂]

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: **ME** Yes
ii) Please provide the underlying arguments and evidence that support your views.
Response: **ME** I think it is clear, accessible and proportionate. I would consider providing more detailed breakdown of sub-steps under each of the headings as a way to further organise the guidance.
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:
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:	
ii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:	

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: **TN:** Ofcom's measure of onerousness is not well explained – this appears to be based on unquantified assumptions of the costs of regulatory compliance in respect of the measures. It is noted that a cost benefit analysis is provided for under Annex 14 for certain (but not all) measures. In order for the argument to be sustained, Ofcom ought to set out, on balance, each measure which has been disapplied, and how the test of onerousness has been considered on the basis of: - The anticipated cost of the obligation/measure - The anticipated benefits of applying the obligation/measure - The anticipated risks of disapplying the obligation/measure In addition, there does not appear to be clear analysis of the risks of disapplication of measures, and how such risks, in Opcom's view, may be adequately mitigated or continuously reviewed and monitored	
ii)	Please provide the underlying arguments and evidence that support your views.

Response:

TN:

As an example, measures proposed for U2U services disapplies the obligation of risk assessments and monitoring and assurance for all smaller service providers.

Where risk assessments are required for large service/low risk, and large service/specific risk service providers, internal monitoring and assurance obligations have also been disapplied.

Furthermore, obligations to establish a dedicated reporting chanrisknel for fraud has been disapplied for all service providers except providers of large service/specific risk.

This is a deviation from best practice and expectations from the financial services industry, where, for example, assessment of the inherent risk of business activities and the financial crime risks to which firms may be exposed has been the norm for many years. It is also a legal obligation under the Money Laundering Regulations 2017. In addition, firms must also have in place internal systems and controls to monitor and assess risks on a continuous basis.

As successive studies have noted, online platforms are a key vector for fraud, the proceeds of which are integrated into the banking sector. It therefore appears perverse that service providers are not subject to equal/similar standards of behaviour.

Ofcom's proposal to disapply specific measures and the underlying reasoning, in light of best practice from the financial industries, should be subject to further disclosure and scrutiny.

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

Response:

Question 14:

i) Do you agree with our definition of large services?

Response:

TN:

No

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

Response:

TN:

The binary distinction between large services (average user base greater than 7 million per month) and smaller services (average user base lower than 7 million per month), and the different obligations applying to each, potentially creates a cliff edge and adverse incentives:

- Small to medium size service providers may be incentivised to segregate service offerings so as to avoid exceeding the 7 million average user base, and the additional obligations this attracts under the Code
- Service offerings by small to medium size service providers may become more attractive as a platform to illegal content due to perceived or actual reduction in oversight, compared with larger firms

A disapplication of specific obligations under the Code, based on average user size applied in a binary manner, remains problematic. The eco-system view should be undertaken, rather than an assumption that entities operate on an individualised basis with only localised effects.
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: ME Partially
ii) Please provide the underlying arguments and evidence that support your views.
Response: ME There appears to be an ontological flaw in the split between low, specific and multi. A service medium or high on one risk that did not fall into one of the nominated specific categories would find that none of the categories would apply to them.
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:
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: **UOG**

No discussion of the safeguarding of staff moderating this content **YL** Yes, moderation can be used to effectively mitigate the online harm. To make the proposal more useful, it can be extended with the details of policies on what are categorised as “harmful contents” or “unharmful contents”, where are the boundaries set out as the guidance of content moderation.
ii) Please provide the underlying arguments and evidence that support your views.
Response: **YL:** Having the concrete (qualitative) details can help improve the transparency of the policy.
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:	
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 (User to User)

Question 20:	
i)	Do you agree with our proposals?
Response: **[CONFIDENTIAL✂]** **UOG** Content Moderation, and specifically Automated Content Moderation, is very much focussed on text/imagery as the content - this does not take into account user behaviour, interactions and intentional, directed verbal/non-verbal communication as the "content" such as in social virtual reality (e.g., VRChat, RecRoom) and immersive online platforms. The potential for automation here can go well beyond URLs and keyword search or transcripts analysis, towards looking at the non-verbal social signals in the space (proxemics, speech analysis) to e.g., infer stress/distress, longitudinal records of behaviour/interactions (particularly for platforms with some authentication). However, the associated cost of detection also increases, and the complexity of interpreting non-verbal behaviour is high, and would require validation. See paper: https://doi.org/10.1145/3531073.3534492 **YL** Yes, the use of automated technology can help the content moderation. Some additional considerations may help to improve this part.	
ii)	Please provide the underlying arguments and evidence that support your views.
Response: **YL** Similar to the previous section on Content Moderation, I believe the introduction of algorithm design and usage is necessary. Besides, how the dataset of “harmful contents” can be built and maintained as the moderation is going on is critical to ensure the quality of the moderation work. In addition, having the user feedback on automated moderation decisions can help improve the accuracy and fairness of these systems. Therefore, the design of the interface (UX/UI) and	

features (user input and processing) to enable wide user engagement should be highlighted. Meanwhile, here should include user training of this mechanism to report errors in automated content moderation. **[CONFIDENTIAL✂]**
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:	
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:	

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

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:

UOG:

Particularly for immersive Social VR platforms, **Children need training and education on how to use the existing safety measures effectively.** In immersive platforms experiences can be particularly overwhelming because they are in 3D and surround the user, with e.g. group harassment being particularly problematic here.

There's no consideration given for **parental/guardian insight/oversight of usage of platforms**. The nature of VR experienced through a head-mounted device makes it challenging for parents to oversee the child's usage. Parents also lack awareness on existing parental controls and risks social VR platforms may have. This has been a common, strongly recurring theme in our work. Consideration needs to be given to how to keep parents in-the-loop e.g. via automated journalling driven by existing harm detection approaches and the child's own use of platform safety tools.

Particularly where automated moderation is concerned, there needs to be a degree of **visibility, transparency, and explainability around actions taken for children in particular**. Without visibility and transparency, automated moderation won't set suitable boundaries in terms of the expected behaviours on the platform, and without explainability children won't develop trust in how this moderation is applied. The social VR app landscape is still very heterogeneous, and there is no industrial standard for logging, recording, and making available events and in particular non-verbal conversational/communicational interactions to authorities. This poses a challenge to the above proposals on oversight, moderation or journaling. While automated moderation can be promising, it is still under-developed and would need human-involved checks to some extent.

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 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: **UOG** Common features that should be supported on immersive platforms include personal space boundaries that can be easily and quickly set – perhaps a recommendation that there should be additional platform-appropriate user controls to minimize risk of harm, with suitable defaults (e.g. personal space boundaries in social immersive platforms; restrictions to speech usage, aged-group matching or private settings that can limit child to interact with their friends' list) Response: **YL** Yes users should be giving some levels of management of the content visibility and interaction with others. Some additional considerations may help to improve this part.

ii)	Please provide the underlying arguments and evidence that support your views.
Response: **YL** Considering the users may not have sufficient knowledge about privacy and online safety, it is important to illustrate the consequences of decisions made. To ensure users play a more active role in enhancing their online experience, providing insight into how automation algorithms determine the content they see and offering controls to adjust these settings is essential. The proposal could also include requirements for using the control services, provided with different sets of educational resources and tips on how they can protect themselves from illegal content and online harms. This could cover topics such as recognising common threats, using platform safety features, digital literacy etc.	
iii)	Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No	

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: **JC &IK** The Ofcom Guidance is significant in that it organizes the details necessary for the fulfilment of the obligation to take measures against illegal information in accordance with the Online Safety Act 2023. In order for Internet service providers to fulfil their legal obligations, such as by way of deletion of content and blocking of processing on the ground of illegality, the Ofcom Guidance sets out (a) what a service should consider to determine if it has 'reasonable grounds to infer that content is illegal content', and (b) what may constitute information that is 'reasonably available' to services when making an illegal content judgement. However, it seems that the following points should be taken into account: 1. The Guidance takes a dualist path, allowing services to either follow the process set out therein to determine when a piece of content is illegal or to draft their own terms and conditions in a way that complies with the ICJG. The Guidance rightly assumes that most services will opt for the second (or for a hybrid) path (26.19). A mapping of the terms and conditions of a number of platforms (both large and small) against the ICJG would provide more precise guidance as to where there are possible needs for adjustment. 2. The Guidance states that 'Services are free to take down content above and beyond what is illegal under the Act, so long as they make this clear in their terms of service, and that their content moderation practices result in the timely removal of illegal content as set out in the illegal content safety duties' (26.18). Furthermore, it emphasises that any limitations on the right to freedom of expression as a result of the Act must be prescribed by law, pursue a legitimate aim and be necessary in a democratic society. The implication is that content restrictions carried out on the basis of the platforms' terms of service do not need to comply with fundamental right guarantees, in particular Art. 10 ECHR. However, courts in other jurisdictions accept an indirect third-party effect of fundamental rights (see e.g. OLG München, judgement of 24.08.2018 – 18W 1294/18; OLG Karlsruhe, judgement of 25.6.2018 – 15 W 86/18). It would be useful to provide certain benchmarks that would ensure the compliance of platforms' terms of service with Art. 10 ECHR. 3. The Guidance states that 'it is important to note that "reasonable grounds to infer" is a new legal threshold and is different from the "beyond reasonable doubt" threshold used by the criminal courts' (26.14). Given that service providers are obliged to take content down when they have reasonable grounds to infer that the content is illegal, it would be useful to provide them with specific examples of a)

	'beyond reasonable doubt' cases recognised by the criminal courts and b) lower threshold 'reasonable grounds to infer' cases to ensure the predictability of the Guidance.
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:	

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

Response:

JC &IK

1. The Guidance acknowledges that 'inferring the state of mind or 'mental' element of a piece of content is a particularly difficult challenge at scale' (26.13), and that such inferences 'are particularly difficult in online situations, where contextual clues are often not apparent and, for example, what would be an obvious joke or piece of sarcasm in an offline context might not appear so obvious when online' (26.43). Inferring the state of mind is also especially challenging where content has been shared, forwarded or reposted. As recognised in the Guidance, the same material can be shared by one user in a way that is supportive of the message conveyed, and by another to challenge or criticise the same content (26.46). Further guidance as to how to infer the state of mind in the case of reposting would be desirable. If in doubt, the threshold for assuming illegality in such cases should be higher in view of the limited engagement of 'a significant minority of users' with the content they forward (26.54). Overall, the advice on inferring users' mental state is somewhat conflicted. On the one hand it is stated that 'neither Ofcom nor in scope services can put aside the state of mind or "mental element" requirement as this is a part of the "reasonable grounds to infer" threshold, as established by the Act'. At the same time, the Guidance advises that it is possible to sidestep this complex requirement by applying services' own terms and conditions, community guidelines or publicly available statements rather than the law (26.43). In order to increase the accessibility of the Guidance, also for services with limited access to legal expertise, and to ensure that freedom of expression is protected, it is advisable to provide specific examples from case law and further guidelines as to how to assess the mental element required for specific offences, beyond what is contained in the Annex.
2. The Guidance states that in view of the significant overlap between laws in the United Kingdom's three legal jurisdictions, the practical impact of jurisdictional differences is limited, but that there are isolated cases in which a priority offence in one part of the United Kingdom is different from the other jurisdictions (26.78). This is particularly evident in the area of threatening and abusive behaviour, where the

Scottish offences are the broadest and easiest to prove, whilst offences from other jurisdictions cannot be ignored as there is only partial overlap (26.134). The guidance on these different levels of culpability is particularly difficult to access for services with limited access to legal expertise. It might be advisable to set a grace period for this differentiated application of the Online Safety Act, and to require service providers (with limited access to legal expertise), in the first instance, to only comply with the common denominator for the standard for illegality in the three jurisdictions in question.

3. Ofcom provides limited guidance as regards the offences of 'foreign interference' and 'false communication' in view of the relative lack of body of case law or academic discussion. These are offences where the risk of disproportionate restriction of freedom of expression is high, as acknowledged in Annex 10 (A13.24). The false communications offence criminalises the sending of a message known to be false with the intent of causing non-trivial emotional, psychological or physical harm to a likely audience without reasonable excuse[1].Online platforms subject to considerable fines will be inclined to take down content perceived to be false without evidence of harm but on the mere basis that this content might cause harm to an indeterminable audience [2] The Act does not define what amounts to 'non-trivial emotional, psychological or physical harm', raising the possibility of unduly stifling freedom of expression if the bar is set too low. In addition, it does not specify what would constitute a reasonable excuse. While in the case of the abolished harmful communications offence the contribution to a matter of public interest was a possible, though not necessarily absolving excuse, the same does not necessarily apply to false communications [3]. The Act does not explain what a possible excuse might be. A citizen journalist reporting on President Trump's suggestion to use disinfectant as a cure for Covid-19 might hence find themselves criminally liable. They might not have the intent to cause harm. However, if intent to harm is inferred from the communication act itself, as exemplified in a fact-sheet that tested the operation of the Bill by way of a series of case-studies, then the door is opened for legitimate content to cross the threshold of criminal liability [4].

1 Online Safety Act, s. 179.

2 See I. Katsirea, Press Freedom and Regulation in a Digital Era: A Comparative Study (OUP, forthcoming May 2024), Ch. 5.

3 Law Commission, 'Harmful Online Communications: The criminal offences. A consultation paper', 11 September 2020 < https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2020/09/Online-Communications-Consultation-Paper-FINAL-with-cover.pdf> 136.

4 Department for Culture, Media and Sport, 'Online Safety Bill: Communications offences factsheet', 19 April 2022 < <https://www.gov.uk/government/publications/online-safety-bill-supporting-documents/online-safety-bill-communications-offences-factsheet>.

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: **JC &IK** 1. The Guidance states that reasonably available information, which could provide the foundation for 'reasonable grounds to infer', includes such that is obtained from law enforcement (26.26). The Annex states further that services may choose to maintain bilateral relationships with law enforcement, and that nothing in this Guidance should be taken as discouragement to maintain such relationships (A1.47). It is, however, necessary to bear in mind that such authorities are not functionally independent from the government as is required in the case of agencies regulating audiovisual media content (see VG Köln, decision of 1 March 2022 – 6L1277/21). There is a concern of collateral censorship as Internet service providers may be more likely to delete or block content due to complaints filed by law enforcement agencies. One area where this concern is pronounced is that of inchoate offences related to immigration and human trafficking. In view of the complex nature and high evidentiary threshold of these offences, the Guidance proposes to rely on information made available by law enforcement authorities (26.268). This could lead to the proactive removal of images of immigrants out of political expediency in a way that could stifle democratic discourse around immigration. Therefore, it is necessary to review closely the inclusion of law enforcement agencies among the third parties. 2. A further source that it used as the basis for judging the illegality of content is 'user profile information' and 'user profile activity' (26.26). This can easily induce users to self-censor, and restrict their freedom of expression online. It is therefore necessary to include strict parameters in the Guidance for the use of such information for the purpose of inferring illegality. 3. The Guidance, in line with s. 192 of the Online Safety Act, further includes information made available by way of automated systems and processes, as well as such that is provided by automated systems and processes together with human moderators, as information that is reasonably available to internet service providers (26.33ff). When human moderators verify information provided through an automated system or process to infer illegal information, it is necessary to stipulate essential procedural requirements and moderators' qualifications to ensure their objectivity and expertise. Furthermore, the proposed automated content detection technology (26.34) will help Internet service providers judge illegal information. However, those who distribute illegal information tend to avoid illegal information regulation by changing URLs (e.g. AAA.COM → AAA.COM1). It is therefore necessary to consider recognizing URLs as a basis for judging illegal information not only when they are perfectly matched, but also when URL changes are abusively performed, in order to avoid the same information distribution. At the same time, safeguards need to be put in place to confirm the identity of illegal information distribution whilst avoiding the problem of overregulation and/or regulation with malicious intent. It is also necessary to add security guidance to prevent the leakage of URL lists distributing illegal information to the outside world.
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:	