ONLINE SAFETY ACT NETWORK

ANNEX B: Illegal Harms Consultation Response - Supplementary Evidence Table

This table is provided in support of our main consultation response document and follows the same thematic structure. It compares statements made by Ofcom in the illegal harms consultation documents with other available evidence, including Parliamentary transcripts and research evidence on harm, providing references and extracts where appropriate. We also refer again to the outcomes that Ofcom has said it hopes to deliver, particularly the first two below.

Specifically, we anticipate implementation of the Act will ensure people in the UK are safer online by delivering four outcomes (Figure 1):

- Stronger safety governance in online firms
- Online services designed and operated with safety in mind;
- Choice for users so they can have meaningful control over their online experiences; and
- Transparency regarding the safety measures services use, and the action Ofcom is taking to improve them, in order to build trust. (Ofcom approach document)

SECTION 1: Weak safety by design foundations

Issue

At a relatively late stage in the progress of the Online Safety Act, the Government inserted a new "clause 1" which set out the overall objectives of the legislation, including a duty on providers to ensure that services are "safe by design". The Act makes numerous other references to design, as well as "systems and processes" in relation to companies' risk assessment and safety duties. Indeed, Ofcom has adopted as one of its outcomes for implementation that "online services [are] designed and operated with safety in mind".

Much of our analysis in our covering consultation response is interlinked, providing evidence of the choices that Ofcom has made which – taken together – we believe will not deliver this stated outcome, notwithstanding the fact that these proposals are just one part of a jigsaw that will not be complete for a number of years. Similarly, much of the evidence to support this particular issue is found in other sections of this document.

No	Example	Consultation doc references	Evidence and commentary
1	The Act at section 1 says "OSA 1 (3) a Duties imposed on providers by this Act seek to secure (among other things) that services regulated by this Act are— (a) safe by design," But the measures recommended by Ofcom are limited to ex-post, take down approaches rather than systemic "by design" interventions.	Approach p5 "Our role is not to instruct firms to remove particular pieces of content or take down specific accounts, nor to investigate individual complaints. Our role is to tackle the root causes of online content that is illegal and harmful for children, by improving the systems and processes that services use to address them. Seeking systemic improvements will reduce risk at scale, rather than focusing on individual instances." BUT: there is very little focus in the supporting documents on the way in which design, or the business model of services, or their commercial imperatives contribute to causing harm	Introducing the new clause, Lord Parkinson said on Report day 1 6th July (column 1230) "Subsection 3 of the proposed new clause outlines the main outcomes that the duties in the Bill seek to secure. It is a fundamental principle of the legislation that the design of services can contribute to the risk of users experiencing harm online .. I am pleased to confirm that this amendment will state clearly that a main outcome of the legislation is that services must be safe by design. For example, providers must choose and design their functionalities so as to limit the risk of harm to users. … I hope this provides reassurance about the Government's intent and the effect of the Bill's framework." The Government published a series of documents in 2021 setting out how services could introduce "safety by design".

No	Example	Consultation doc references	Evidence and commentary
		The measures in the codes (see issue 6) below do not map back to the risks identified in volume 2. See supporting ANNEX A for more on this.	It introduced them by saying: “Safety by design is the process of designing an online platform to reduce the risk of harm to those who use it. Safety by design is preventative. It considers user safety throughout the development of a service, rather than in response to harms that have occurred. The government has emphasised the importance of a safety by design approach to tackle online harms. The government’s response to the Online Harms White Paper highlighted the importance of a preventative approach to tackling online safety, including through safer platform design. In response to this, the government committed to publishing guidance to help UK businesses and organisations design safer online platforms. By considering your users’ safety throughout design and development, you will be more able to embed a culture of safety into your service.” The “best practice” guides include ● private or public channels ● live streaming ● anonymous or multiple accounts ● search functionality ● visible account details or activity” **The Government’s 2022 Impact Assessment set out the following: “**While per business costs are expected to be higher for medium and large businesses, it is important to consider the possibility that some in-scope SMBs will have limited resources for compliance. To minimise burdens on SMBs, it will be vital for Ofcom to work with businesses and to ensure both requirements and enforcement are proportionate to the risk of harm and resources available to businesses. Proportionality in the context of effective safety measures must be balanced against the risk of harmful

No	Example	Consultation doc references	Evidence and commentary
			content being displaced to smaller and less well-equipped platforms. The government and Ofcom will work with SMBs to ensure that steps taken are effective in both reducing harms and minimising compliance costs. ***The government's Safety by Design framework and guidance is targeted at SMBs to help them design in user-safety to their online services and products from the start thereby minimising compliance costs.***" Australian e-safety Commissioner has produced Safety By Design principles– developed in conjunction with industry: "Rather than retrofitting safeguards after an issue has occurred, Safety by Design focuses on the ways technology companies can minimise online threats by anticipating, detecting and eliminating online harms before they occur. This proactive and preventative approach focuses on embedding safety into the culture and leadership of an organisation. It emphasises accountability and aims to foster more positive, civil and rewarding online experiences for everyone." We note that the Australian e-safety Commissioner provided evidence on safety by design for Ofcom's call on the illegal content duties but this is only referenced, briefly, twice in volume 4. The CMA has referred to online choice architecture/nudges in relation to competition and consumer harm. The National Cyber Security Centre has set out a series of "cyber security design principles" that focus on red-teaming design processes (described here) as a means to pre-empt problems. The Ministry of Defence has also produced a handbook on red-teams

No	Example	Consultation doc references	Evidence and commentary
			Beyond regulators and government bodies, IBM has looked at technology design principles that would address domestic violence and work has been done on abusability testing frameworks described here, with examples set out here and here.
2	Definition of takedown "duty" - problematic given this isn't the Act in these terms	Annex 10, A1. 14"As part of the illegal content duty at section 10(3)(b) of the Act, there is a duty for a user-to-user service to "swiftly take down" any illegal content when it is alerted to the presence of it" – this seems to disregard the fact that the obligation is to "operate a service using proportionate systems and processes designed to ..." - ie the obligation relates to the design of the systems and processes which must be operationalised – the obligation is not a stand alone take down obligation Annex 10, A1.16: "However, to make decisions for the purposes of the **takedown duty** or determining what search content is illegal content, services will need to take decisions about specific pieces of content. It is here that this guidance will be particularly useful."	Ofcom describes this correctly in the summary document - Vol 1 2.36 "A duty to operate the service using proportionate systems and processes designed to swiftly take down any (priority or non-priority) illegal content when they become aware of it (the 'takedown duty');" - which reflects s 10 of the OSA 10 (3) A duty to operate a service **using proportionate systems and processes** designed to— (a) minimise the length of time for which any priority illegal content is present; (b) where the provider is alerted by a person to the presence of any illegal content, or becomes aware of it in any other way, swiftly take down such content.
3	No detail on how to design a service to reduce illegal content - detail in guidance and measures in codes is all on the	Vol 1. 2.39 " Services are required to operate using proportionate systems and processes designed to take down all illegal content when they become aware of it and minimise the amount of time that priority illegal	See our supporting annex A for the lack of measures related to design in the codes.

No	Example	Consultation doc references	Evidence and commentary
	takedown	content is present on the service before being removed." Vol 1 2.46 "Search services must operate their service "using proportionate systems and processes designed to minimise the risk of individuals encountering" search content which is priority illegal content. It should be noted that this duty is qualified by the need to take proportionate measures to protect users. It does not amount to an absolute duty to prevent all priority illegal content from ever being present in or via search results."	
4	Detail on design (eg volume 2) doesn't feed through to codes	Eg re senior responsibilities (vol 3, 8.64) "Those key responsibilities would include ownership of decision-making and business activities that are likely to have a material impact on user safety outcomes. Examples include senior-level responsibility for key decisions related to the management of risk on the front, middle and back ends of a service. This would include decisions related to the design of the parts of a product that users interact with (including how user behaviour / behavioural biases have been taken into account), how data related to user safety is collected and processed, and how humans and machines implement trust and safety policies. **Depending on a service's structure, key responsibilities in online safety may fall under content policy, content design and strategy, data science and analytics, engineering, legal, operations, law enforcement and compliance, product policy, product management or other functions."**	See annex A analysis table re functionalities and mitigating measures
5	Design comes after decisions about content - this hardly	Vol 3 9.3 The illegal content risk assessment duties include a range of different elements. U2U services	We would refer Ofcom to the series of recent US court

No	Example	Consultation doc references	Evidence and commentary
	emphasises upstream mitigations and safety by design	must assess the risk of users encountering priority illegal content or other illegal content by means of the service, and the level of risk that the service may be used for the commission or facilitation of a priority offence. They must also assess the nature and severity of the harm which may be suffered as a result. 9.4 As part of the assessment, services must consider various characteristics of the service specified in the legislation – such as its user base, functionalities, business model, and systems and processes – and also take account of the relevant risk profile(s) produced by Ofcom. These two paragraphs are the wrong way round.	filings and whistleblower reports that have recently laid out what happens when a "safety by design" approach is not embedded in companies' culture and the impact of platforms' design choices on the harms that are caused to users, particularly children. What is relevant here is that these documents also demonstrate platforms' awareness – over a number of years – of the harms that are being caused by design and their apparent unwillingness to redesign their services to prevent them; this is the exact opposite of safety by design. In the UK, coroners' reports have also identified where platform design has had a direct role in creating the conditions in which individuals have decided to take their own lives. We list some of these documents here for Ofcom's reference and would recommend that these are urgently reviewed as part of their evidence base, not just for application to the measures recommended for addressing illegal content but for the development of the proposals for the children's codes. US court filings • New Mexico Attorney-General case against Meta - January 2024 • Bad Experience and Encounters Framework (BEEF) survey - Instagram internal research - unsealed as part of New Mexico court case - January 2024 • California Superior Court Opinion re dismissal of Fentanyl Case re Snap - January 2024 • Multistate Complaint re Meta - largely unredacted - Nov 2023 • Second amended complaint re Fentanyl and Snap - July 2023 • California Master Complaint in re Adolescent Social

No	Example	Consultation doc references	Evidence and commentary
			Media Addiction - May 2023 • Class action against Tinder et al – February 2024 Whistleblower material • Arturo Bejar testimony to Congress - November 2023 • Sophie Zhang oral evidence to Parliament & written evidence- October 2021 • Frances Haugen evidence to Congress & transcript - October 2021 • FB Archive - searchable repository of the Frances Haugen papers Coroners' reports • Prevention of Future Death Report: Chloe McDermott - December 2023 • Prevention of Future Death Report: Bronwen Morgan - November 2023 • Prevention of Future Death Report: Luke Ashton - July 2023 • Prevention of Future Death Report: Molly Russell - October 2022 • Prevention of Future Death Report: Joseph Nihill – September 2020 • Prevention of Future Death Report: Callie Lewis - December 2019
6	There is a reference to product testing but this is only an "enhanced expectation" for larger services	Table 9.5 vol 3 "We use 'product' as an all-encompassing term that includes any functionality, feature, tool, or policy that you provide to users for them to interact with through your service. This includes but is not limited to whole services, individual features, terms and conditions	Hansard Lords Report stage: Lord Parkinson said: This risk management approach is well established in almost every other industry and it is right that we expect technology companies to take user safety into account when designing their products and services. (Col 1320)

No	Example	Consultation doc references	Evidence and commentary
	There is no justification for this not being a core expectation; in other sectors, product safety applies across the board regardless of size of service. Eg food safety standards; electrical standards etc. There is no definition of product safety either - for example, this could include testing to maximise user engagement revealing eg addiction problems. These are results even if carried out for product development purposes not expressly safety. The US court filings (see reference above) provide lots of examples where this kind of product development work demonstrated evidence of harm that was then not addressed. **See also section 7 re small vs large services**	(Ts&Cs), content feeds, react buttons or privacy settings. By 'testing' we mean services should be considering any potential risks of technical and design choices, and testing the components used as part of their products, before the final product is developed. We recognise that services, depending on their size, could have different employees responsible for different products and that these products are designed separately from one another" "Expectations for larger services: All else being equal, we will generally expect services with larger user numbers to be more likely to consult the enhanced inputs (unless they have very few risk factors and the core evidence does not suggest medium or high levels of risk). This is because the potential negative impact of an unidentified (or inaccurately assessed) risk will generally be more significant, so a more comprehensive risk assessment is important. In addition, larger services are more likely to have the staff, resources, or specialist knowledge and skills to provide the information, and are more likely to be the subject of third-party research." (Vol 3, 9.113)	New Mexico Attorney General material: "Meta launched Reels in order to attract teens who were transitioning to competitors, like TikTok, that already featured a video service. Internal Meta documents confirm that the launch of Reels was rushed in order to preserve engagement among Meta's teen users. One employee noted in a 2020 message: "The fact that we're shipping reels without a clear picture of the ecosystem impact is pretty mind boggling." Another employee echoed that sentiment: "it is scary the speed we are moving . . . we either do things WAY TOO FAST without Data. Or do things WAY TO[O] SLOW because of Design/Principles." These product designers were aware of the harm that could result from Reels, with one stating "I am worried that the cumulative effects are going to be bad."" (p163) Note that the ICO and CMA suggest that testing is to be done when designing choice architecture:Joint paper here

No	Example	Consultation doc references	Evidence and commentary
7	No overall requirement for using metrics from product testing to determine risk, just seen as an "input" in the enhanced category. Note risk of disincentivising product safety as part of design process.	"Any types of evidence listed under Ofcom's enhanced inputs (e.g. the results of content moderation, product testing, commissioned research) that the business already collects and which are relevant to the risk assessment, should inform the assessment. In effect, if the service already holds these inputs, they should be considered as core inputs" Volume 3, table 9.5	This is not enough - see the problems evidenced by Meta whistleblower:: "If the problems identified are not problems that the company's systems are designed to detect and measure, managers literally have no means to understand them. Zuckerberg is unwilling to respond to criticisms of his services that he feels are not grounded in data. For Meta, a problem that is not measured is a problem that doesn't exist." Testimony from Arturo Bejar
8	Risk assessment review after a significant change of service does not allow for testing/risk assessing at the time of design, rather suggests that the design should be implemented and then assessed, which may be too late. **See also risk assessment section 8 below.**	Inconsistency - this then doesn't allow for product testing (see above) Vol 3, Table 9.5 9.123 c) a duty to carry out a further suitable and sufficient illegal content risk assessment relating to the impacts of that proposed change before making any significant change to any aspect of a service's design or operation. This is at odds with: 9.135 "We opted for using a principle-led approach to give services flexibility as what amounts to a significant change can vary across the wide range of services in scope. We consulted with experts internally and externally to help understand the circumstances in which a change to a service may be significant enough to cause the risk assessment to become out of date and no longer provide a suitable and sufficient assessment of risk on the service." 9.138 "we understand that the larger and more complex a service may be, the more likely it is to have routine updates or system changes which we did not feel it was proportionate to capture under this duty.	The proposals do not however link to the risk mitigation measures, which are specific and which - at the bare minimum - is all that services need to comply with if they are to meet their duties under the Act in relation to the specific risks that they have identified. (Safe harbour) There is no "flexible" requirement on services to mitigate the harms they have identified via product testing or risk assessment.

No	Example	Consultation doc references	Evidence and commentary
		9.141 We provisionally conclude that our proposed "significant changes" to consider are necessary in order for services to be confident that they are complying with their legal duties, hence any associated costs are proportionate and are primarily based on the requirements of the Act, rather than on regulatory choices made by Ofcom. This is particularly given we have adopted a principle led approach (rather than directive) which affords flexibility to services to help them meet this duty as appropriate relative to its size, capability and specific circumstances that may affect risk. Overall, we think this approach is proportionate for services to help them meet a specific duty set out in the Bill." 9.135/6 suggests changes will happen and risk assessment will be out of date NOT that risk assessment should happen before change is mad	
9	Codes go straight from governance and accountability measures into content moderation - there is a gap where measures to deliver the duty relating to "design of functionalities, algorithms and other features" should be. (See section 10 (4) for U2U and section 27 (4) for search). **See also section 6 below re disconnect between volume 2 and volume 4**	"Compliance with these duties, in particular the duties to take down illegal content swiftly on becoming aware of it and to take appropriate action in response to complaints about illegal content, would be very difficult in practice absent some process for determining whether or not content ought to be taken down and implementing that decision as appropriate." (Vol 4, 12.8)	This plays into what the companies want - and presumably what they have told Ofcom. Eg In his recent evidence to Congress, Meta whistleblower Arturo Bejar said: "Meta's current approach to these issues only addresses a fraction of a percent of the harm people experience on the platform. In recent years, repeated examples of harm that has been enabled by Meta and other companies has come to light, through whistleblowing, outside research studies, and many stories of distressing experiences people have there. **Whenever such reports emerge, Meta's response is to talk about 'prevalence', and its investment in moderation and policy, as if that was the only relevant issue. But there is a material gap between their narrow definition of prevalence and the actual distressing**

No	Example	Consultation doc references	Evidence and commentary
			experiences that are enabled by Meta's products. However, managers including Meta CEO Mark Zuckerberg do not seem to seek to understand or actually address the harms being discussed. Instead, they minimize or downplay published findings, and even sometimes the results of their own research. They also try to obfuscate the situation by quoting statistics that are irrelevant to the issues at hand."
10	"Signals of emerging harm"	"Set and record internal content policies. These should set out rules, standards and guidelines about: what content is allowed and not allowed on the service, and how policies should be operationalised and enforced. In doing so, services should have regard to its risk assessment and signals of emerging illegal harm." Vol 4: Chapter 12 p19 Also Vol 4, Para 13.157 "Risk assessment and information pertaining to the tracking of signals of emerging harm - A service's risk assessment will be one of the key sources of information telling a service what risk of search content that is illegal content they have on their platform and would form the basis for internal content policies (see Measure 2). As moderators should be focused on enforcing the internal content policies, training should also be informed by the most recent illegal content risk assessment. In Chapter 8, we are also consulting on a proposed recommendation that services should track signals of emerging harm. If, following consultation, we remain of the view we should recommend this, this information would be one of the key sources of information about how illegal content manifests and it is therefore crucial services use this to inform their content moderation training and supporting materials."	Recent Facebook Oversight Board ruling has shown that, even where companies have content policies, these may be inadequate. Relying on the existence of these as a measure in itself therefore will not address the harm. The oversight board, in its ruling that - based on its "manipulated media" policy, FB was right to leave up a video that implied President BIden was a paedophile - said that the policy itself was "lacking in persuasive justification, is incoherent and confusing to users, and fails to clearly specify the harms it is seeking to prevent. In short, the policy should be reconsidered. The policy's application to only video content, content altered or generated by AI, and content that makes people appear to say words they did not say is too narrow. Meta should extend the policy to cover audio as well as to content that shows people doing things they did not do. The Board is also unconvinced of the logic of making these rules dependent on the technical measures used to create content. Experts the Board consulted, and public comments, broadly agreed on the fact that non-AI-altered content is prevalent and not necessarily any less misleading; for example, most phones have features to edit content. Therefore, the policy should not treat "deep fakes" differently to content altered in other ways (for example, "cheap fakes").

No	Example	Consultation doc references	Evidence and commentary
			The ruling is also instructive wrt to its finding that "in most cases Meta could prevent the harm to users caused by being misled about the authenticity of audio or audiovisual content through less restrictive means than removal of content. For example, the company could attach labels to misleading content to inform users that it has been significantly altered, providing context on its authenticity." See also the recent US Court filings for discussions on content moderation and problematic "prevalence" Eg In his recent evidence to Congress, Meta whistleblower Arturo Bejar said: "Meta's current approach to these issues only addresses a fraction of a percent of the harm people experience on the platform. In recent years, repeated examples of harm that has been enabled by Meta and other companies has come to light, through whistleblowing, outside research studies, and many stories of distressing experiences people have there. Whenever such reports emerge, Meta's response is to talk about 'prevalence', and its investment in moderation and policy, as if that was the only relevant issue. But there is a material gap between their narrow definition of prevalence and the actual distressing experiences that are enabled by Meta's products. However, managers including Meta CEO Mark Zuckerberg do not seem to seek to understand or actually address the harms being discussed. Instead, they minimize or downplay published findings, and even sometimes the results of their own research. They also try to obfuscate the situation by quoting statistics that are irrelevant to the issues at hand."
11	Operating on a "complaints only" basis - these are post hoc responses and put onus on users	Vol 4, para 12.17 "Some services, for example low risk/smaller services, may not have very much content to moderate (e.g.	Re deepfake story above: "Google allows deepfake victims to request the removal of such content from search results through a form, but it isn't proactively searching for and

No	Example	Consultation doc references	Evidence and commentary
	to report problems, not on services to identify and address risks of harm. And there may be a 'buy in' to the focus of the services so users don't spot problems too. How does this fit with "signals of emerging harm" from risk assessment (above)	because they receive few complaints, because proactive content detection technology is beyond their means, or because their business model is such that there is little likelihood of users uploading any illegal content without the service knowing about it). By contrast, larger and higher risk services may face significant challenges in terms of the volumes and diverse nature of the content they need to moderate, giving risk to questions about how to prioritise content for review, achieve consistency, quality and timeliness of decision-making, and plan their deployment of moderation resourcing so as to secure that users are appropriately protected."	delisting deepfakes itself. The takedown request page says, "We only review the URLs that you or your authorized representative submit in the form."

SECTION 2: the approach to the illegal content judgements guidance

The safety by design approach is central to the regime and should influence the implementation of both the illegal content safety duties and the children's safety duties, on which Ofcom will be consulting in phase 2 later this year. The illegal harms consultation, as the first component in the regime, should provide the framework on which these further consultations can build. Yet, the guidance focuses primarily on individual items of content and assessing whether they should be taken down – it even refers in the draft Guidance to the obligation being "to take content down" (Annex 10, A1.14), rather than, as s 10(3) says, to operate a proportionate system designed to have that effect. While there are parts of the consultation which reflect the obligation correctly - for example, in the "Overview" document where Ofcom says "A new legal requirement of the Act is for all services to swiftly take down specific illegal content when they become aware of it" – the Act's systemic language is generally ignored in the draft guidance itself. Choices about design happen before you get the content flowing across them. There is also no real consideration of scale - the sheer volume of information that is potentially involved. This then defines the scope of Ofcom's overall illegal harms approach, with a focus on ex-post measures, such as content moderation and take down, which we discuss in more detail below.

Furthermore, by requiring that a criminal offence has taken place each time content is posted (rather than acknowledging that content which has been deemed illegal remains illegal when shared as it is still connected with the original offence), an unnecessarily limited view of relevant content is baked into the proposals compounded by an approach that sets the standard of proof at a high threshold – in some instances close to the criminal level – at odds with what is a civil regulatory regime. Again this approach does not sit well with a systems-based approach. Moreover, this is especially problematic given that some criminal offences operate to protect individuals' fundamental rights; the rights balance here is, again, one-sided (see more general discussion here and in section 5 below and attached as a PDF). It is also unfortunate that Ofcom has not considered any of the existing non-priority offences, specifically s 127(1) of the Communications Act, which (unlike 127(2) Communications Act) has not been repealed.

We have published a detailed analysis on this issue by Prof Lorna Woods and provided this as a separate PDF (Annex D) which we refer Ofcom to as our evidence in this section.

SECTION 3: Burden of proof/evidence threshold

Much store is set in the consultation document narratives by the amount of evidence already collected to support the proposals eg the risk management approach, and on the "best practice" already provided by platforms to justify the approach. Conversely, where there is weak or limited evidence relating to the potential for a particular measure to address a particular outcome, this is given as a reason not to include it within the codes until more evidence becomes available (though this approach is not required by the Act). (See section 6 on measures and the codes below.) This approach reinforces the status quo, setting a "lowest common denominator" approach to a piecemeal, process-driven regime, rather than one that is focused on the outcomes described in the Act.

No	Example	Consultation doc references	
1	Burden of proof/lack of evidence There are lots of references throughout the consultation document to evidence lacking; the potential impact on market; that metrics should be down to companies and shouldn't be for Ofcom to define. Ofcom could instead, within the parameters of the Act, have chosen a position where it said "we don't have the right answer so we're not recommending a precise approach but we are asking companies to have a good faith attempt at it, in a way that is proportionate and appropriate to their service and its	Volume 4 11.16 & 11.17; Says there isn't evidence as to whether things will work / lack of precautionary principle "We recognise that identifying previously unknown content is an important part of many services' processes for detecting and removing illegal content. We do not yet have the evidence base to set out clear proposals regarding the deployment of technologies such as machine learning or artificial intelligence to detect previously unknown content at this time. As our knowledge base develops, we will consider whether to include other recommendations on automated content classification in future iterations of our Codes (Vol 4, 11.15, c) "Many of the measures we propose are for large services. This is often because we do not yet have enough information on the potential costs and benefits to know whether the measures are proportionate for smaller services at this point. As our understanding develops, it may be appropriate	Ofcom's letter to Peers in April 2023 reassured them that they were well advanced in relation to illegal harms because: **The Government's and Parliament's intentions about what they want platforms to achieve are clear**. We launched a call for evidence on illegal harms in July 2022, and **are well-advanced in gathering the necessary evidence, including on consumer experiences of those harms, drivers of risk, and the systems and processes available to services to address them." (here)** In previous work for Carnegie UK which set out the initial proposal for basing online harms regulation on a duty of care approach, Professor Lorna Woods and William Perrin set out the merits of the precautionary principle – already established within regulatory practice – as a means to address the risk of harm in areas of fast-moving innovation, where the evidence base may not nascent. The ILGRA published in 2002 a fully worked-up version of the precautionary principle for UK decision makers: The

No	Example	Consultation doc references	
	functionalities". This would also be in line with the precautionary principle.	in future iterations of the Codes to expand the range of services for which some measures are recommended." (Vol 4, 11.16) "Recognising that we are developing a new and novel set of regulations for a sector without previous direct regulation of this kind, and that our existing evidence base is currently limited in some areas, these first Codes represent a basis on which to build, through both subsequent iterations of our Codes and our upcoming consultation on the Protection of Children. In this vein, our first proposed Codes include measures aimed at proper governance and accountability for online safety, which are aimed at embedding a culture of safety into organisational design and iterating and improving upon safety systems and processes over time (Vol 4 11.14)" "Nevertheless, there is little available evidence on how services deploy this human resource across their content moderation systems to deal with illegal and/or harmful content. Where human reviewers are used, it is possible to have different teams for different types of harm, and/or different teams for different reporting channels (e.g. flags or reports from trusted flaggers could be channelled to different teams, or could be fed into one team). (Vol 4 12.26) Lack of evidence "At this stage, there is a lack of evidence and little consensus on the specific outcomes content moderation systems and	precautionary principle should be applied when, on the basis of the best scientific advice available in the time-frame for decision-making: there is good reason to believe that harmful effects may occur to human, animal or plant health, or to the environment; and the level of scientific uncertainty about the consequences or likelihoods is such that risk cannot be assessed with sufficient confidence to inform decision-making.' The ILGRA document advises regulators on how to act when early evidence of harm to the public is apparent, but before unequivocal scientific advice has had time to emerge, with a particular focus on novel harms. ILGRA's work focuses on allowing economic activity that might be harmful to proceed 'at risk', rather than a more simplistic, but often short-term politically attractive approach of prohibition. The ILGRA's work is still current and hosted by the Health and Safety Executive (HSE), underpinning risk-based regulation of the sort we propose. We believe that – by looking at the evidence in relation to screen use, internet use generally and social media use in particular – there is in relation to social media "good reason to believe that harmful effects may occur to human[s]" despite the uncertainties surrounding causation and risk. On this basis we propose that it is appropriate if not necessary to regulate and the following sets out our proposed approach." (Woods and Perrin, Online Harms: a statutory duty of care and regulator, Carnegie UK 2019; pp10-11)

No	Example	Consultation doc references	
		processes should be achieving, although we consider exceptions to this general position in in Chapter 14. (Vol 4 12.34) “At this stage we are inviting respondents to share any further information they may hold in relation to the following: - lists ACM, hashing, trusted flagger” … “We recognise that identifying previously unknown content is an important part of many services’ processes for detecting and removing illegal content. We do not yet have the evidence base to set out clear proposals regarding the deployment of technologies such as machine learning or artificial intelligence to detect previously unknown content at this time. As our knowledge base develops, we will consider whether to include other recommendations on automated content classification in future iterations of our Codes” (11.15 c) Limited evidence cited costs as a reason 14.12 “We are not proposing to recommend some measures which may be effective in reducing risks of harm. This is principally due to currently limited evidence regarding the accuracy, effectiveness and lack of bias of the technologies that the measures refer to. We recognise that some of these measures may be proportionate for certain services to take, and welcome further innovation and investment in safety technologies to support ACM. We plan to consider further ACM measures for future versions of our Codes.”	It is not in regulated companies’ interest to provide evidence to fill gaps for Ofcom in order then to be regulated on it. Civil society organisations are being asked to fill the gaps - but with minimal resources and without access to the information that is held within platforms. Also, recent court cases in US have revealed the amount of evidence of harm (and knowledge of it) that has been suppressed by companies. Eg in the multistate complaint filed by 48 Attorney-Generals last November, it summarises the case against Meta: “Meta’s scheme involved four parts: (1) through its development of Instagram and Facebook, Meta created a business model focused on maximizing young users’ time and attention spent on its Social Media Platforms; (2) Meta designed and deployed harmful and psychologically manipulative product features to induce young users’ compulsive and extended Platform use, while falsely assuring the public that its features were safe and suitable for young users; (3) Meta concealed and suppressed internal data showing the high incidence of user harms on its Social Media Platforms, while routinely publishing misleading reports boasting a deceptively low incidence of user harms; and (4) despite overwhelming internal research, independent expert analysis, and publicly available data that its Social Media Platforms harm young users, Meta still refuses to abandon its use of known harmful features—and has instead redoubled its efforts to misrepresent, conceal, and downplay the impact of those features on young users’ mental and physical health.” Companies’ own evidence gathering functions - and how

No	Example	Consultation doc references	
		Ofcom have evidence on “beacon platforms” but then they say it’s not enough to recommend. (Vol 4. 14.221) So a) why say it? Or b) why not come at issue systemically or based on outcomes etc? “However, at this stage, we consider we require further evidence in order to propose a recommended measure tackling the harm created by the dissemination of these links, in particular about the following areas” Also vol 3 8.131 re independent audit “ we do not consider there is currently enough information on the effectiveness of other possible measures to be able to recommend them in Codes at this stage.	this affects “best practice” material provided to Ofcom - is not clear. For example, in the New Mexico Attorney General case it notes - re CSAM material - “Meta’s reliance on user reports to identify unlawful, dangerous, or inappropriate conduct demonstrates the failure of its own efforts to detect and remove these materials.” “User reports of potentially violative content, including commercial sexual activity and CSAM, are discouraged and do not reflect the kinds of abuses children encounter or experience, and are often met with no response, delayed response, or, shockingly, a response indicating that material clearly violative of Meta’s Community Standards was not, in fact, a violation” p 94 The New Mexico AG final has a long section detailing all the evidence that has emerged from Meta in recent years that demonstrates that awareness of the harm caused on their platforms - and the impact of their design choices and decisions on that harm - was well known within the company. Ofcom has not taken this into account in their evidence gathering. (See XIV. META WAS ACUTELY AWARE OF THE HARM TO YOUTH WELL-BEING RESULTING FROM ITS DESIGN CHOICES, BUT FAILED TO DEVOTE SUFFICIENT RESOURCES TO ADEQUATELY ADDRESS THE HARM TO YOUTH pp168 onwards) “At the same time that Meta was making these design choices, internal documents confirm that Meta was aware of the harmful effects that its products were having on the wellbeing of children and teenagers. Meta performed numerous studies and analyses concerning teen usage and the effects resulting therefrom, but systematically ignored internal red flags in favor of chasing profits.”

SECTION 4: The approach to proportionality

Ofcom's approach to proportionality is primarily economic: to avoid imposing costs on companies. While the OSA requires regulated services take a "proportionate" approach to fulfilling their duties, and recognises that the size and capacity of the provider is relevant, the Act also specifies that levels of risk and nature and severity of harm are relevant. This focus on costs and resources to tech companies is not balanced by a parallel consideration of the cost and resource associated with the prevalence of harms to users (for example, on the criminal justice system or on delivering support services for victims) and the wider impacts on society (particularly, for example, in relation to women and girls and minority groups, or on elections and the democratic process). The assumption in the proportionality analysis that "small" means "less harm" due to less reach, and "single risk" means "less impact" due to it being obvious, is also an issue, particularly given that it downplays the severe harm that can occur to minoritised groups on targeted, small sites - which we discuss further below. We look below in section 7 at how the principle of proportionality plays into Ofcom's differentiated approach to small and large companies.

No	Issue	Consultation doc references	Evidence and commentary
1	Costs on companies	Refers to Comms Act re Ofcom's duty to protect citizens - but narrative throughout is focused on protecting businesses (Vol 1, 1.5): The Communications Act 2003 ('the CA 2003') places a number of duties on us that we must fulfil when exercising our regulatory functions, including our online safety functions. Section 3(1) of the CA 2003 states that it shall be our principal duty, in carrying out our functions: • To further the interests of citizens in relation to communication matters; and • To further the interests of consumers in relevant markets, where appropriate by promoting competition. (Vol 4, p4) "We consider larger services will tend to be better able to bear the costs of the more onerous measures than smaller services.Not about commercial viability of companies but about harms	Arturo Bejar testimony "I have specific recommendations for regulators, to require any company that operates social media services for teenagers to develop certain metrics and systems. These approaches will generate extensive user experience data, which then should be regularly and routinely reported to the public, probably alongside financial data. I believe that if such systems are properly designed, we can radically improve the experience of our children on social media. The goal must be to do this without eliminating the joy and value they otherwise get from using such services. I don't believe such reforms will significantly affect revenues or profits for Meta and its peers. These reforms are not designed to punish companies, but to help teenagers. And over time, they will create a safer environment. " Choice of word "onerous" - that itself has a value judgement Dictionary.com provides this definition: *"burdensome,*

No	Issue	Consultation doc references	Evidence and commentary
		(including human rights violations)" Microbusinesses: 11.47 "We are required under the Act to consider the impact of our proposed measures on small and micro businesses" 11.53: "We consider it can be prudent to exempt smaller services from incurring those costs (where appropriate provided they are not high risk), as there will often be significant uncertainty in any assessment of benefits and costs, and we want to reduce the possibility of imposing financially damaging costs on businesses when the magnitude of benefits expected to result from the measure is uncertain." 12.88 "Services that do not currently have internal content policies would incur the costs of developing them. This could take a small number of weeks of full-time work and involve legal, regulatory, as well as different ICT staff, and online safety/ harms experts. In some cases, services may use external experts which could increase costs. Agreeing new policies may also take up senior management's time which would add to the upfront costs. For most services we expect these costs to be in the thousands of pounds, although larger/riskier services may require more complex content policies which may increase costs. In addition there may be some small ongoing costs to ensure these policies remain up to date over time. "	*oppressive, or troublesome; causing hardship: onerous duties; having or involving obligations or responsibilities, especially legal ones, that outweigh the advantages: onerous agreement*". It is an inappropriate choice of word by a regulator charged with implementing a regime that is about reducing harm to individuals (including human rights obligations) and not about preserving the profitability of companies.

No	Issue	Consultation doc references	Evidence and commentary
2	Costs for companies are cited as a factor in undertaking proper risk management - without counterbalancing the costs of harm to society	Vol 3 9.66 “In addition, our proposed methodology is intended to be flexible depending on service’s risk levels, size and resources in order to minimise the cost burden. We intend that it could be integrated into existing risk management practices to improve the effectiveness of online safety risk assessments and minimise additional costs”	Does focus on costs suggest that if a company doesn’t have risk management in place, the costs of implementing it are not justifiable? A “flexible” approach should mean that companies should incur more costs if they are starting from a lower base, not that a lack of resources should take into account a more minimal approach to risk. Proportionality assessment does not take into account significance of harm - and impact on users, costs to society. The Government’s 2022 Impact Assessment (IA) quantified the cost to society of a number of illegal and other harms (including CSEA, hate crime, drugs, modern slavery and cyberstalking) and estimated that these added up to £5 billion/year. The IA went on to say that “these estimates are likely to underestimate the full extent of online harms for several reasons · It has only been possible to quantify the cost of a subset of all online harms in scope: there are a number of harms that are encountered by a significant number of adults and children in the UK, but for which there is no evidence on which to make an estimate of their cost. These include encouraging terrorism and radicalisation online, which 5% of adults and 6% of children in the UK have encountered, and encouraging self-harm, which 5% of adults and 10% of children have encountered. · For those harms that have been quantified, a conservative approach has been undertaken. For example, for illegal harms analysis is based on the number of recorded offences with an online element, which is likely to understate the true prevalence (as some crimes will go unreported - although this is adjusted in part by the use of multipliers where

No	Issue	Consultation doc references	Evidence and commentary
			appropriate) · Crimes may feature an online element but not be flagged as online: currently, whether a crime is recorded as having an online element is reliant upon police recording practices and how police forces apply the online flag. This, again, will reduce the reported prevalence of a given harm, and lead to an underestimate of its cost." (p80) The Australian e-Safety Commissioner recently reported on information provided to her office by X/Twitter via a transparency report including the decision to cut staff working on safety globally, which demonstrates what happens when costs rather than risks are the primary driver of company decision-making: " X Corp. said Twitter/X's global trust and safety staff have been reduced by a third, including an 80 per cent reduction in the number of safety engineers, since the company was acquired in October 2022. The company also said the number of moderators it directly employs on the platform have been reduced by more than half, while the number of global public policy staff have also been reduced by almost 80 per cent." What is notable in this report are the findings of the impacts on the platform's safety. In the same period since the acquisition by Elon Musk: ● there had been a 20% slowing in the median time to respond to user reports about Tweets and a 75% slowing in the median time to respond to direct messages. eSafety notes that prompt action on user reports is particularly important given that Twitter

No	Issue	Consultation doc references	Evidence and commentary
			solely relies on user reports to identify hateful conduct in direct messages. ● As of May 2023, X Corp. reported that no tests were conducted on Twitter recommender systems to reduce risk of amplification of hateful conduct. However, X Corp. stated no individual accounts are artificially amplified, and that its enforcement policies apply to Twitter Blue accounts in the same way as other accounts. ● As of May 2023, automated tools specifically designed to detect volumetric attacks or "pile-ons" in breach of Twitter's targeted harassment policy were not used on Twitter. ● As of May 2023, URLs linking to websites dedicated to harmful content are not blocked on Twitter. ● From 25 November 2022 (the date it was announced)1 to 31 May 2023, 6,103 previously banned accounts were reinstated by Twitter, which eSafety understands relates to accounts in Australia. Of these, 194 accounts were reinstated that were previously suspended for hateful conduct violations. X Corp. stated that Twitter did not place reinstated accounts under additional scrutiny.
3	Large services, large risk (numerical judgement) Small services, small volumes. **See also small vs large companies in section 7**	Vol 3, 8.86 "However, because large services have high reach and the potential to affect a lot of users, we consider that failures in oversight of risk management would have wider impacts on user" Vol 3, 8.86 12.98 "We are not proposing to recommend this measure for smaller and lower risk services. We consider the benefits of an internal content moderation policies are likely to be materially	No evidence is given by Ofcom for these judgements. GIven it is fundamental to the way the regime has been designed, it would be helpful to see where this evidence on impacts on users has come from.

No	Issue	Consultation doc references	Evidence and commentary
		smaller for services which are neither large nor face material risks. They are unlikely to face large volumes of content they need to assess. So even though the costs of this measure are low, we do not propose to recommend it for such services."	
4	Single risk sites deemed to be less harmful than multi-risk sites therefore a "proportionate" response is not to recommend measures for them. See also section 7 on small vs large sites (below)	Vol 4, 11.44: "We intend these measures to apply to services that face significant risks for illegal harms in general. There is a question over what it means for a service to have such risks. One option would be to recommend these measures to services that have identified as medium or high risk of at least one kind of illegal harm. However, where services only identify a risk of a single kind of illegal harm, the benefits of these measures to address all harms will be lower. This is partly because if services have only identified a single area of risk, the extent of harm will tend to be lower compared to if they have identified a range of kinds of offence where they are high risk. It is also partly because many of these measures are about enabling services to have a good understanding of their risks and of the content moderation policies needed to address those risks. If a service was only of medium or high risk for a single kind of illegal harm, the risk is more likely to be well understood across the organisation, such as the risk of fraud for some marketplace services. This tends to mean the benefits of these measures in terms of improving understanding and consistency of approach are smaller than if there were multiple areas of risk. The case for the measures to address all harms being proportionate therefore tends to be stronger if we only apply them to services that have identified multiple kinds of illegal harm"	This assumption that a single risk site causes less harm than a multi-risk site and the "benefits" of addressing it are therefore lower is not borne out by the specific harm that some small dedicated sites can cause to individuals. · groupings of providers that do not have a distinct legal form or are shell companies and therefore can reconstitute themselves as different sorts of legal entities with different URLs or websites (eg marketplaces for suicide methods that are repeatedly taken down and re-emerge, evading regulatory intervention; here and here); · small sites that have a single purpose that is extremely harmful to some groups, often with targeting of individuals - eg revenge porn collector sites (for example, here and here); · dedicated hate and extremism sites, such as those researched in relation to inceldom by CCDH here and covered in this Parliamentary submission; far-right ideologies investigated by Hope Not Hate here and here; and extremism in this ISD report.

No	Issue	Consultation doc references	Evidence and commentary
5	Proportionality analysis on costs measured against overall risk management	Vol 3 8.89 “Given the benefits of ensuring senior level responsibility and oversight for online safety, and small costs associated with this measure, we consider it proportionate to provisionally recommend to large services (with the exception of large vertical search services) and services which identify as multi-risk (including vertical search services which are multi-risk). Although for small risky services the cost impact will tend to represent a higher share of total revenue, our view is that such a measure is proportionate given the evidence that clearly defined roles and responsibilities at a senior level helps improve overall risk management processes. We consider this an important aspect in ensuring the effective management and mitigation of all illegal harms.” (smaller and low risk - what about smaller and high risk)	
8	Tracking evidence of new and increasing harm	Vol 3 8.147 “We have identified ongoing costs associated with these recommendations. We anticipate that these costs are likely to scale with service size, whereby larger services will likely face higher costs related to implementation. However, we recognise that these costs are likely to be a larger proportion of revenue for smaller services	Talks about scaling - but why not apply across board now? Eg if small services integrate them into their processes while small, then they can scale them up as they grow, rather than waiting for the problem to become significant once they reach the large numerical threshold that Ofcom has identified.
9	Staff training - multidisciplinary teams	8.161 8.162 - generic 8.167	No minimum considerations are offered. Standards would be welcome here to ensure there is a culture change within organisations as well as the necessary regulatory effects. Ofcom use scant evidence to say this is already being done but are unwilling to use scant evidence for alternatives
10	Metrics Relying on information from	Vol 3 9.34 “There are clear differences between large services which often provide detailed information about the metrics they gather to	New Mexico A-G court filings show the problem with trusting Meta’s metrics (p199-201):

No	Issue	Consultation doc references	Evidence and commentary
	companies means that you don't find the risks - risk assessment doesn't require evidence	assess safety on their services, and smaller services, with fewer UK users, which have often never engaged in risk assessment nor considered why it could be important in their industry. For instance, among smaller services whose business models are likely to result in higher levels of risk, such as those hosting adult content, some state that they circumvent the need for a risk assessment by moderating every piece of content which appears on the platform."	Meta's efforts to publicly portray its platforms as safe and largely free of illicit content extends to quarterly Community Standards Enforcement Reports ("CSER") which "provide metrics on how we enforced our policies . . . and estimates on the amount of violating content (Prevalence) on Facebook and Instagram." Meta's May 15, 2018 press release announcing the formation of these reports made clear that the reports were and are intended to allow the public to see "how much bad stuff is out there," and thereby permit the public to "judge our performance for yourself." Meta positioned itself as a company invested in eliminating illicit content from its platforms: "We believe that increased transparency tends to lead to increased accountability and responsibility over time, and publishing this information will push us to improve more quickly too. This is the same data we use to measure our progress internally – and you can now see it to judge our progress for yourselves." Each and every one of these reports underreport the existence of objectionable or violative conduct on Facebook or Instagram because they all rely on Meta's flawed "prevalence" standard. A May 23, 2019 blog post described "prevalence" as "[o]ne of the most significant metrics we provide in the Community Standards Enforcement Report." Meta reported that "we consider prevalence to be a critical metric because it helps us measure how violations impact people on Facebook. We care most about how often content that violates our standards is actually seen relative to the total amount of times any content is seen on Facebook." It compared this metric to "measuring concentration of pollutants in the air we breathe" and claimed that "[p]revalence is the internet's equivalent – a measurement of what percent of times someone sees something that is harmful." Meta's CSERs consistently reported low prevalence of

No	Issue	Consultation doc references	Evidence and commentary
			human trafficking, CSAM, bullying and other problematic materials. For example: a. The CSER released in November 2019 claimed that prevalence was an "upper limit [of] 0.04%" of views for content violating Meta's policies prohibiting "child nudity and sexual exploitation of children, regulated goods, suicide and selfinjury, and terrorist propaganda." b. The December 2020 CSER claimed that "less than 0.05% of views were of content that violated our standards against Child Nudity and Sexual Exploitation" and that "less than 0.05% of views were of content that violated our standards against Suicide and Self-Injury." c. The Q3 2021 CSER reported "that between 0.14% to 0.15% of views were of content that violated our standards against Bullying & Harassment" and that "less than 0.05% of views were of content that violated our standards against Suicide & Self-Injury." Individually and collectively, each of these reports conveyed the impression that Meta aggressively enforced its Community Standards on both Facebook and Instagram, and that its efforts were succeeding in keeping the platforms relatively free of harmful content. For example, a November 13, 2019 news release announcing release of the fourth CSER includes the claims that the purpose of the report is to "demonstrate our continued commitment to making Facebook and Instagram safe and inclusive.". Nowhere do the CSERs explain how much sexualized content remains on the platforms and accessible to children; the ability of adult strangers to identify, groom, and seek sexualized content and activity from children; or the widespread sale of CSAM, among other commercial sexual exploitation of children. Moreover, as explained above, the prevalence metric consistently underestimated the amount of problematic and illicit content displayed on Facebook. The prevalence metric contradicted the findings of Meta's own BEEF study, which

No	Issue	Consultation doc references	Evidence and commentary
			showed a much greater “prevalence” of bad experiences involving illicit, questionable or violative conduct on Meta’s platforms. Arturo Bejar re changing the metrics/data that is collected/required: “The most effective way to regulate social media companies is to require them to develop metrics that will allow both the company and outsiders to evaluate and track instances of harm, as experienced by users. This plays to the strengths of what these companies can do, because data for them is everything. If something cannot be evaluated by data analysis, it is generally very difficult for Meta and other such companies to understand the problem or take action. Process-based or policy-based regulations are essential for security and privacy. In order to effectively regulate the safety of a social media environment, the focus should be on metrics based on user experience. “

SECTION 5: The approach to human rights

The OSA directs Ofcom to consider freedom of expression (Art 10 ECHR) and privacy (Article 8 ECHR), but these are not the only relevant rights – as indeed Ofcom notes. All the rights protected by the Convention should be considered when considering the impact of the regime – or the lack of it. So, as well as the qualified rights of freedom of expression (Article 8 ECHR), the right to private life (Article 11 ECHR) and rights noted by Ofcom – e.g. the right to association (Article 11 ECHR) – we should consider other rights including the unqualified rights – the right to life (Article 2 ECHR), freedom from torture and inhuman and degrading treatment (Article 4 ECHR) as well as the prohibition on slavery and forced labour (e.g people trafficking) (Article 4 ECHR). Note also that rights can include positive obligations as well as an obligation to refrain from action; a public body can infringe human rights by failing to protect as well as by interfering itself in an individual's rights.

Article 14 ECHR constitutes the requirement for people not to be discriminated against in the enjoyment of their rights; all people (and not just users of a particular service) should be considered. This reflects the general principle of human rights that all people's right should be treated equally – and indeed that the starting point is that no right – for example, freedom of expression – has automatic priority over another. It also means that the European Court has adopted a specific methodology for balancing rights of equal weight (see e.g. *Perinçek v. Switzerland* (27510/08) [GC] 15 October 2015, para 198; *Axel Springer AG v. Germany* (39954/08) [GC] 7 February 2012, paras 83-84 on the balance between articles 8 and 10) rather than its typical approach where a qualified right may suffer an interference in the public interest but that interference must be limited. This difference in methodology reaffirms the significance of seeing all the rights in issue when carrying out balancing exercises. A failure to carry out a proper balance by national authorities has itself led to a finding of a violation of the procedural aspects of the relevant right. – the precise factors taken into account in the balance will vary depending on the underlying facts in a case and the rights involved.)

Note also that Article 17 prohibits the abuse of rights so that "any remark directed against the Convention's underlying values would be removed from the protection of Article 10 by Article 17" (*Seurot v France* (57383/00), decision 18 May 2004). While this applies only to a narrow sub-set of speech, it is nonetheless a factor that should form part of the balancing exercise where relevant. Areas where Article 17 might be relevant include threats to the democratic order (*Schimanek v Austria* (32307/96), dec 1 February 2000); racial hatred (*Glimmerveen and Hagenbeek v NL* (8348/78 8406/78), dec 11 October 1979); holocaust denial (*Garaudy v France* (65831/01), dec 24 June 2003); religious (*Belkacen v Belgium* (34367/14), dec 27 June 2017) or ethnic (*Ivanoc v Russia* (35222/04), dec 20 February 2007) hate; hatred based on sexual orientation; incitement to violence and support for terrorist activity (*Roj TV A/S v Denmark* (24683/14), dec 18 April 2018). The Court has not considered CSAM material but it is submitted that it, likewise, would fall outside the protection of Article 10.

We have published a detailed analysis on this issue by Prof Lorna Woods and provide it as a PDF at annex C as our evidence in this section.

SECTION 6: Disconnect between approach to risk identification and risk mitigation (codes)

We have concerns that the approach set out in volume 2 and 3 - the identification of risks and the material for the risk register, and the approach to risk management - does not follow through to the measures that are described in the codes. Even when limited to content moderation (not addressing systemic and functionality mitigation measures), small/single-risk services are let off hook based on their size and the proportionality assessment - exemption from measure 2 in volume 4 leads to further exemptions for measure 3 and 4 that are not risk-based. We refer to our large evidence table at annex A which compares the functionalities identified in volume 2 with the measures (or lack thereof) to address them in volume 4. The extracts below provide further context to this.

No.	Issue	Consultation doc references	Evidence and commentary
1	The approach set out at the start of the consultation by Ofcom says that the onus is with providers to be the judge but the codes are presented as a "tick-box" list. The Act does not specify that this approach has to be taken.	"The Act is clear: first and foremost, the onus sits with service providers themselves, to properly assess the risks their users may encounter, and decide what specific steps they need to take, in proportion to the size of the risk, and the resources and capabilities available to them" (p4) Approach doc, p4 Volume 4, para 11.7: "Services that choose to implement the measures we recommended in our Codes of Practice will be treated as complying with the relevant duty. This means that Ofcom will not take enforcement action against them for breach of that duty if those measures have been implemented."	While the s 41 expects Ofcom to describe measures and s 49 introduces a comply or explain approach, measures can be described with greater or less degrees of precision. Measures also need not be technical but could incorporate for example an safety by design obligation (see definition of 'measure' in s 236(1) and list in s 10(4)(a) and (b) and s 27(4)(a) and (b)) which fall within the list of 'measures' that could be taken)
2	Ofcom presents lots of detail and evidence on types of functionalities that can cause harm but this does not then feed through to codes. The evidential threshold for Ofcom to	Volume 2 This section on "suitable and sufficient" illustrates the problem: 9.22 (b): "Given the purpose of the risk assessment duty, we propose that a suitable and sufficient risk	See separate analysis of some examples of functionalities and whether these are covered in the codes at Annex A.

No.	Issue	Consultation doc references	Evidence and commentary
	make recommendations seems (unnecessarily) high. See section 3 on the burden of proof/evidence thresholds.	assessment should be relevant to the specific characteristics of the service in question and should accurately reflect the risks. It is important that the risk assessment provides services with an adequate understanding of the risks to implement appropriate measures in response. c) We therefore propose that risk assessments should, as far as possible, be based on relevant evidence on the risk of harm on the service. In particular, services should consider evidence on the risk arising from the characteristics of the service specified in under Sections 9(5) and 26(5). The quality of the evidence and analysis underpinning the risk assessment is a key component of ensuring it is suitable and sufficient." Volume 3 Eg the business model is mentioned in some domains but this might also apply to other types of problem content Eg "Our goal is that services prioritise assessing the risk of harm to users (especially children) and run their operations with user safety in mind. This means putting in place the insight, processes, governance and culture to put online safety at the heart of product and engineering decisions." (Vol 3, 9.8)	
3	Codes are presented as a base on which to build BUT This is potentially the lowest common denominator. There is also no timescale for escalating issues	Vol 3: 8.16 "Our first Codes are aimed at establishing robust governance and accountability processes and represent a basis on which to build. We anticipate making further updates to our Codes through a process of iteration as our evidence base evolves."	Schedule 4 of the OSA sets out that: OFCOM must ensure that measures described in codes of practice are compatible with pursuit of the online safety objectives. Section 4 sets out "the online safety objectives for regulated

No.	Issue	Consultation doc references	Evidence and commentary
	for the first iteration, or influencing the drafting of the next versions. See also section 3 above on burden of proof.	Vol 4 11.14 "Recognising that we are developing a new and novel set of regulations for a sector without previous direct regulation of this kind, and that our existing evidence base is currently limited in some areas, these first Codes represent a basis on which to build, through both subsequent iterations of our Codes and our upcoming consultation on the Protection of Children. In this vein, our first proposed Codes include measures aimed at proper governance and accountability for online safety, which are aimed at embedding a culture of safety into organisational design and iterating and improving upon safety systems and processes over time"	user-to-user services are as follows— (a) a service should be designed and operated in such a way that— (i) the systems and processes for regulatory compliance and risk management are effective and proportionate to the kind and size of service, (ii) the systems and processes are appropriate to deal with the number of users of the service and its user base, (iii) United Kingdom users (including children) are made aware of, and can understand, the terms of service, (iv) there are adequate systems and processes to support United Kingdom users, (v) (in the case of a Category 1 service) users are offered options to increase their control over the content they encounter and the users they interact with, (vi) the service provides a higher standard of protection for children than for adults, (vii) the different needs of children at different ages are taken into account, (viii) there are adequate controls over access to the service by adults, and (ix) there are adequate controls over access to, and use of, the service by children, taking into account use of the service by, and impact on, children in different age groups; (b) a service should be designed and operated so as to protect individuals in the United Kingdom who are users of the service from harm, including with regard to—

No.	Issue	Consultation doc references	Evidence and commentary
			(i) algorithms used by the service, (ii) functionalities of the service, and (iii) other features relating to the operation of the service.
4	Risk profiles suggest there may be a cumulative impact of features and some societal impact but the codes do not allow for this.		Expectation re how women and girls were to be protected via the illegal content codes (prior to the Govt concession on VAWG guidance) suggests Govt expected they should be doing more than just explaining measures for takedown. Lord Parkinson at Lords Report stage said: "On Amendments 94 and 304, tabled by my noble friend Lady Morgan of Cotes, I want to be unequivocal: all service providers must understand the systemic risks facing women and girls through their illegal content and child safety risk assessments. They must then put in place measures that manage and mitigate these risks. Ofcom's codes of practice will set out how companies can comply with their duties in the Bill. I assure noble Lords that the codes will cover protections against violence against women and girls. In accordance with the safety duties, the codes will set out how companies should tackle illegal content and activity confronting women and girls online. This includes the several crimes that we have listed as priority offences, which we know are predominantly perpetrated against women and girls. The codes will also cover how companies should tackle harmful online behaviour and content towards girls." Parkinson went on to say: My noble friend Lady Morgan suggested that the Bill misses out the specific course of conduct that offences in this area can have. Clause 9 contains

No.	Issue	Consultation doc references	Evidence and commentary
			provisions to ensure that services “mitigate and manage the risk of the service being used for the commission or facilitation of” an offence. This would capture patterns of behaviour. In addition, Schedule 7 contains several course of conduct offences, including controlling and coercive behaviour, and harassment. The codes will set out how companies must tackle these offences where this content contributes to a course of conduct that might lead to these offences.Lords Committee stage 16 May 2023 column 205 See also the separate submission to this consultation from the VAWG sector.
5	Is the definition of suitable and sufficient enough?	There is no definition in the OSA. However, “we consider this to be an important requirement which has two main components: a) Services must ensure they complete all the relevant elements of a risk assessment specified in the Act; and b) Services must carry out each of these individual elements to a standard that is suitable and sufficient for their service in the context of its obligations under the regime as a whole.” (Vol 3, 9.22)	
6	Risk profiles - does the *type* of service present a risk?		Some of this is covered by the lists of services that are associated with particular offences (eg chatrooms for suicide/self harm) but this does not follow thorough to the codes. Some services may be wilfully blind to their risk so is focus on functionalities the right way? Is it a rational approach to the reality of crap services?

No.	Issue	Consultation doc references	Evidence and commentary
7	Lower expectations re role of codes, "patchy" status quo acknowledged. **See also section 7 on small vs large platform**s	"In this vein, our first Codes aim to capture existing good practice within industry and set clear expectations on raising standards of user protection, especially for services whose existing systems are patchy or inadequate. Each proposed measure has been impact assessed, considering harm reduction, effectiveness, cost and the impact on rights." Chapter 11, p3 "Work to capture existing good practice, not to raise bar"	Problem with industry's representation of its existing good practice and/or its reassurances that it is doing all it can, set out here in this list of "red herrings" related to the recent Congressional hearings: https://www.techpolicy.press/red-herrings-to-watch-for-at-the-senates-child-safety-hearing/ In recent webinar (36 mins onward), one Ofcom representative said: "Tech Uk are a really close partner with us … voluntary principles are already in place across a number of harms that a number of us have helped to formulate over the years .. and actually, to be candid, for quite a while some of those voluntary principles are going to go further than we're going to be able to go on the codes until we're able to catch up … It's going to be easier to recommend something as a voluntary principle than it is to have to meet the bar of evidence to codify that in a code of practice. So there will be some time where voluntary principles go further until we catch up .. a lot of those voluntary principles contain some really good practice things about what companies can be doing." If this is already agreed by industry as good practice - and that is what Ofcom is building the codes on - why aren't these voluntary principles already in the codes? New Mexico AG filings summarise all the internal documentation that has emerged in recent years demonstrating the awareness within platforms of how their services cause harm, which often went unaddressed. Re lowering the bar, Meta's awareness of how their platforms encouraged and enabled the discovery of suicide content goes back at least as far as 2019, when discussion on how to handle media responses to the Molly Russell case included the following (as summarised in the New Mexico AG filings):

No.	Issue	Consultation doc references	Evidence and commentary
			"Although the coroner's inquest took several years, Meta employees were acutely aware of the lack of safeguards built into Instagram and expressed their concerns in emails following the Guardian's outreach to Meta for comments on Ms. Russell's death. In a January 26, 2019 email thread addressing Meta's response to a forthcoming media story profiling "30 families of suicide victims accusing Instagram of killing their children," one Meta employee wrote: "We are defending the status quo when the status quo is clearly unacceptable to media, many impacted families, and when revealed in press, will be unacceptable to the wider public." Recipients of the thread included Zuckerberg, Sandberg, and Mosseri. Another Meta employee responded to echo the theme that Instagram protocols were insufficient: "our present policies and public stance on teenage self harm and suicide are so difficult to explain publicly that our current response looks convoluted and evasive . . . The fact that we have age limits which are unenforced (unenforceable?) and that there are, as I understand it, important differences in the stringency of our policies on IG vs Blue App [Facebook] makes it difficult to claim we are doing all we can." Sandberg eventually chimed in, asking whether Meta could improve its policies or whether it was a question of enforcement and confirmed "We can definitely say that we need to improve our enforcement of our policies." (p173 Revealing Reality report on Snapchat:" This research suggests Snapchat's design features not only enable the sharing of unpleasant and illegal material, but in some cases shape the behaviour that leads to its creation".
8	Measure 2 on content policies only applies to large or multi-risk services, and as a result, additional	Annex 7, p 64 Eg "A service is at medium or high risk of a kind of illegal harm specified in the table if the risk assessment of the service identified a medium	Act and Parliamentary debates didn't take variegated approach to CSEA and terrorism.

No.	Issue	Consultation doc references	Evidence and commentary
	measures that flow from this are not recommended. So, even though vol 4 says they will be included in their illegal harms and CSEA codes, they are not for all services. CSEA and terrorism duties not covering all services? Is that what is intended?	or high risk (as the case may be) in relation to the offences (taken together) specified in the table in relation to that harm, including (where relevant) as further specified in the table." This includes terrorism and CSEA and suggests that *both* have to be present to be deemed medium or high risk. Measure 2: "we consider that services that follow this measure are more likely to operate effective content moderation systems. As we have shown, the evidence suggests that effective content moderation plays a hugely important role in mitigating the risk of harm to users meaning the measure would have important benefits. As with measure 2, these benefits will be greatest for services that are either large or multi-risk … We are not proposing to recommend this measure for smaller and lower risk services, because it is less clear the benefits are great enough given the lower volume of content such services need to assess." (12.114-116) Then, this flows on from that assessment repeated at : "this measure is predicated on services having the internal content policies of Measure 2 above and the performance targets we propose in Measure 3, so it makes sense for this measure to apply to the same set of services as those proposed measures are recommended for" Eg 12.171, 13.141 Eg recommender measure then only applies to services " that meets both of the following conditions: a) the provider conducts on-platform testing of recommender systems on the service; and b) the service is at medium or high risk of at least two of the following kinds of illegal harm"	

No.	Issue	Consultation doc references	Evidence and commentary
9	Measure 4: Large or multi-risk services should have and apply policies on prioritising content for review. In setting the policy, the provider should have regard to at least the following factors: virality of content, potential severity of content, the likelihood that content is illegal, including whether it has been flagged by a trusted flagger.	12.165 “We are aware of a small service which needed to increase spending for online safety by several hundred thousand per annum to deal with problematic content on its service, some of which was illegal.This illustrates the potentially substantial scale of the costs even small services may face where they are high risk.”	Ofcom assumes that content teams will always be underfunded - should there be a minimum standard of resourcing that is proportionate to size of platform. It seems unacceptable to suggest that services can indefinitely postpone dealing with “minor” illegal content if it’s illegal Evidence from eSafety Commissioner on X shows what happens when costs in content moderation and other teams are cut, even at one of the largest platforms: The Australian e-Safety Commissioner recently reported on information provided to her office by X/Twitter via a transparency report including the decision to cut staff working on safety globally, which demonstrates what happens when costs rather than risks are the primary driver of company decision-making: “ X Corp. said Twitter/X’s global trust and safety staff have been reduced by a third, including an 80 per cent reduction in the number of safety engineers, since the company was acquired in October 2022. The company also said the number of moderators it directly employs on the platform have been reduced by more than half, while the number of global public policy staff have also been reduced by almost 80 per cent.” What is notable in this report are the findings of the impacts on the platform’s safety. In the same period since the acquisition by Elon Musk: > there had been a 20% slowing in the median time to respond to user reports about Tweets and a 75% slowing in the median time to respond to direct messages. eSafety notes that prompt action on user reports is particularly important given that Twitter solely

No.	Issue	Consultation doc references	Evidence and commentary
			relies on user reports to identify hateful conduct in direct messages. As of May 2023, X Corp. reported that no tests were conducted on Twitter recommender systems to reduce risk of amplification of hateful conduct. However, X Corp. stated no individual accounts are artificially amplified, and that its enforcement policies apply to Twitter Blue accounts in the same way as other accounts. As of May 2023, automated tools specifically designed to detect volumetric attacks or “pile-ons” in breach of Twitter’s targeted harassment policy were not used on Twitter. As of May 2023, URLs linking to websites dedicated to harmful content are not blocked on Twitter. From 25 November 2022 (the date it was announced)1 to 31 May 2023, 6,103 previously banned accounts were reinstated by Twitter, which eSafety understands relates to accounts in Australia. Of these, 194 accounts were reinstated that were previously suspended for hateful conduct violations. X Corp. stated that Twitter did not place reinstated accounts under additional scrutiny.

SECTION 7: Small vs large platforms

No	Issue	Consultation doc references	Evidence and commentary
1	Ofcom divides up measures between those that apply to all services and those that only apply to large and multi-risk services. Have Ofcom got a remit within the Act to differentiate in this way?	See proportionality extracts above:	Parliamentary debates on small vs large focused on category 1 = but this categorisation doesn't apply to illegal harms so why are Ofcom differentiating so early? Lord Parkinson Committee stage refused amendments that would have exempted smaller services; "The current scope of the Bill reflects evidence of where harm is manifested online. There is clear evidence that smaller services can pose a significant risk of harm from illegal content, as well as to children … Moreover, harmful content and activity often range across a number of services. While illegal content or activity may originate on larger platforms, offenders often seek to move to smaller platforms with less effective systems for tackling criminal activity in order to circumvent those protections. Exempting smaller services from regulation would likely accelerate that process, resulting in illegal content being displaced on to smaller services, putting users at risk. … the Bill has been designed to avoid disproportionate or unnecessary burdens on smaller services. All duties on services are proportionate to the risk of harm and the capacity of companies. This means that small, low-risk services will have minimal duties imposed on them. Ofcom's guidance and codes of practice will set out how they can comply with their duties, in a way that I hope is even clearer than the Explanatory Notes to the Bill, but certainly allowing for companies to have a conversation and ask for areas of clarification, if that is still needed. They will ensure that low-risk services do not have to undertake unnecessary measures if they do not pose a risk of harm to their users." (Col 1153)

No	Issue	Consultation doc references	Evidence and commentary
			Also, when batting away Bns Morgan's amendment re category 1 platforms at Report Stage, Parkinson was emphatic: I will say more clearly that small companies can pose significant harm to users—I have said it before and I am happy to say it again—which is why there is no exemption for small companies. The very sad examples that my noble friend Lady Morgan gave in her speech related to illegal activity. All services, regardless of size, will be required to take action against illegal content, and to protect children if they are likely to be accessed by children. This is a proportionate regime that seeks to protect small but excellent platforms from overbearing regulation. However, I want to be clear that a small platform that is a font of illegal content cannot use the excuse of its size as an excuse for not dealing with it.
2	"Bad actors" use large and small platforms - how does this map onto what they propose.	Summary p8 and also volume 2 commentary (see next row). Eg Terrorism volume 2 "Services with a small user base and less reach can also be used by terrorist actors, but for different reasons. For example, while services with a large user base may be used to attract and draw individuals into the group through influence tactics and dissemination of propaganda, smaller services can be used by perpetrators to undertake more sensitive activities, such as recruitment, planning and fundraising. " 6B.31 Eg Fraud volume 2 6O.45"While larger services are a particular target for fraudsters, services with small user bases may also be targeted, depending on the type of fraud. The NFIB has found that	This isn't reflected in volume 4 and in draft codes where small platforms are exempted. See large analysis document at annex A.

No	Issue	Consultation doc references	Evidence and commentary
		some fraudsters look for more niche services in the UK, if these are widely used by certain communities or professions which they can target. For instance, romance fraudsters will join user groups centred around dating or making friendships such as widower groups or singles groups, and comment on their availability, compliment others, and seek to communicate privately. Fraudsters will also target investment groups; they often send mass messages, a practice which is less likely to be adopted by legitimate users looking for a personal connection. “	
3	Lots of functionalities listed in vol 2 are used by small services that might not now be caught by codes.	Grooming . “Perpetrators will move victims from larger to smaller services depending on their objective. “ Vol 2 CSAM “evidence suggests that perpetrators also often use small and less-mature services to share CSAM, as these services may be less likely to have CSAM detection technology and processes in place. “ (Vol 2 p62) CSAM “some services with a smaller user base offer users specific functionalities which may not be available on services with larger user bases, such as the ability to post content without a registered account. Perpetrators may target these services in order to exploit such functionalities” (Vol 2 p71)	See large analysis document at annex A for comparison between functionalities and codes.
4	Volume 2 specifically identifies business model risk of small/early stage companies re eg terrorism, grooming CSAM - but these services	Eg Vol 2 p63 “Low-capacity services, and services that are earlier in their business development lifecycle, will be at greater risk of being used by perpetrators to share CSAM. Early-stage services	

No	Issue	Consultation doc references	Evidence and commentary
	are then ruled out of many measures due to cost burdens	are less likely to have established processes or resources to detect and/or remove CSAM from their services."	
5	Governance proposals do not take account of the scale and virality of small platforms - can escalate very quickly or be deliberately designed to be risk/catch attention Ref to services not being "mature" but Ofcom does not want "stifling innovation"	Not recommending annual review of risk management for small companies (vol 3 8.45) "For services that are not large, including smaller services that identify some higher risks for users, we are not proposing to recommend this measure at this time. The benefits of imposing this on smaller services are likely to be lower because these services tend to be simpler and easier for management to ensure coordination and consistency in approach." But then this at 8.79 "We consider that this measure could also provide indirect benefits for some services. For example, by ensuring they have adequate risk management and governance frameworks in place from an early stage, which can evolve and expand as the business grows, smaller firms can address any online safety issues early and even save costs overall." "Moreover, it is likely, particularly for smaller services which find high risks to users, that an organisation is not mature enough to have a fully developed governance body. This is especially the case for micro and start-up businesses, or small-scale non-commercial services. This measure would imply significant staff and resource costs, and a change in the overall structure and dynamic of the service for these types of organisations. This could stifle innovation." (8.46) NB microbusinesses specifically identified as a risk	Examples of new companies ripping off model - not pushing innovation forward, using tech to develop services that are controversial or harmful.

No	Issue	Consultation doc references	Evidence and commentary
		in volume 2 “Different research conducted by Tech Against Terrorism concludes that smaller and newer services are most at risk of exploitation, as terrorists and violent extremists such as ISIS may use them. This includes micro-services that may be run by a single individual. This is largely due to targeting and a lack of technical and financial resources for effective moderation.” (6B.76)	
6	Different expectations for larger services - “enhanced” measures are only applying to them	“Expectations for larger services: All else being equal, we will generally expect services with larger user numbers to be more likely to consult the enhanced inputs (unless they have very few risk factors and the core evidence does not suggest medium or high levels of risk). This is because the potential negative impact of an unidentified (or inaccurately assessed) risk will generally be more significant, so a more comprehensive risk assessment is important. In addition, larger services are more likely to have the staff, resources, or specialist knowledge and skills to provide the information, and are more likely to be the subject of third-party research.” (Vol 3, 9.113 e)	
7	Internal logic of distinction then continues to let small companies off hook on content moderation - if they don’t have content policies, or performance targets, they don’t have to have adequate resources etc	This is obligation within the Act - degree of thoroughness distinguishes between service; don’t just say you don’t do them, bearing in mind resources “We are not at this point proposing extending the proposal to services that are not large and are not multi-risk. The amount and diversity of content such services need to moderate is likely to be materially lower and the benefits would therefore be materially smaller, making it questionable	

No	Issue	Consultation doc references	Evidence and commentary
		whether the potentially substantial costs of the measure were always justified for such services. Moreover, this measure is predicated on services having the internal content policies of our proposed Measure 2 above and the performance targets we propose in Measure 3, so it makes sense for this measure to apply to the same set of services as those proposed measures are recommended for." (Vol 4, 12.171) Same for search measures 2, 3, 4 for search	

SECTION 8: Governance and risk assessment

No	Issue	Consultation doc references	Evidence and commentary
1	Is their mitigation of risk the same for u2U and search?		We would refer Ofcom here to the paper we have provided at Annex F from Peter Hanley and Gretchen Peters.
2	How is risk assessment supposed to fit into the review process?	"Regular review of risk management and regulatory compliance by a governance body is required for appropriate oversight over internal controls. Evidence supporting this principle can be found in corporate governance good practice principles and codes. It will be important for governance bodies within services to have a full understanding of risks as identified in an illegal content risk assessment, measures that a service has put in place to mitigate and manage those risks, and how a service intends to deal with developing areas of risk. This requires that governance bodies are made aware of relevant information regarding risk management in a service (provided, for example, by internal assurance functions) and have appropriate reporting lines with senior management." (8.25 8.26 vol 3) 8.29 (is Google a good example?) 8.58 (ditto X)	Emergency review? Ex-post learning from crisis What's the responsibility on the business to take account of review findings? Is this to fit into risk process overall or just annual review? How does a service deal with developing areas of risk?
3	Abusability testing - how does this fit with risk assessment? No general obligation	"Ensuring that services track evidence of new kinds of illegal content, and unusual increases in particular kinds of illegal content, including but not limited to evidence derived from reporting and complaints processes, content moderation processes, referrals from law enforcement and information from trusted flaggers and any other	Who are the expert groups?

No	Issue	Consultation doc references	Evidence and commentary
		expert groups, and report these new kinds of illegal content or unusual increases in illegal content through relevant governance channels to the most senior governance body" (Vol 3 8.97)	
4	Don't raise issue of likelihood of harm happening and not just users, but also non-users	"The illegal content risk assessment duties include a range of different elements. U2U services must assess the risk of users encountering priority illegal content or other illegal content by means of the service, and the level of risk that the service may be used for the commission or facilitation of a priority offence. They must also assess the nature and severity of the harm which may be suffered as a result" (Vol 3 9.3)	IN the OSA "harm" is s 234 – it refers to individuals not users, content is s 236(1) and very broad
5	Risk assessment best practice - this is focused on reputational risks/external risks to the company not product safety and design risks created by their own products and services	Table 9.1, 9.44 "comprehensive risks faced by an organisation"	There are plenty of existing frameworks for rights-based risk assessments that Ofcom can use to improve its approach and methodology. Professor Lorna Woods, under the auspices of Carnegie UK, developed a four-stage model for risk assessment and mitigation on social media platforms that draws on best practice processes through a code-based approach. We would refer Ofcom to her Model Code of Practice as evidence but also provide here extracts from the Ad Hoc Advice to the United Nations Special Rapporteur on Minority Issues which focus on risk assessment. (pp 7-11) This advice was a precursor to the advice to inform the development of his guidance on hate speech as a precursor to developing the Model Code. There is a wealth of high-level guidance on risk assessment that social media companies do not appear to be following. (See Sanjana Hattotuwa, "Making Facebook's New Human

No	Issue	Consultation doc references	Evidence and commentary
			Rights Policy Real", Institute for Human Rights and Business 20 April 2021). Social media companies coming to risk assessment for the first time should evaluate its existing risk management practices and processes, practices in relation to human rights impact assessments generally, and data protection/ privacy impact assessments to evaluate any gap or tensions in those practices and processes and ensure that there is appropriate. Particular attention should be paid to reliance on techniques driven by machine learning and artificial intelligence and the well-known questions around the design and deployment of ML/AI46. (see Council of Europe 'Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems The risk assessment process should be based on data and, where available, research, rather than a hopeful expectation that bad stuff is not happening or, if it is, that it is not the problem of the social media provider. It involves the recognition that the use of technology, including AI, does not in and of itself necessarily ensure human flourishing. (See UNESCO First Draft of the Recommendation on the Ethics of Artificial Intelligence). It should cover an assessment of actual and potential impacts. This involves gathering data in a systemic manner as to what is happening on the service (e.g. what sorts of user complaints are coming, how are they dealt with), as well as the results of any testing on the product (see below), to understand the nature of the problem, as well as its scale, context and triggers and to acknowledge that information, not bury it. For example, hate speech tends to spike for 24-48 hours after key national or international events such as a terror attack, and then rapidly fall. (See Matthew Williams and Mishcon de Reya, Hatred Behind the Screens: A Report on the Rise of

No	Issue	Consultation doc references	Evidence and commentary
			Online Hate Speech). Systems should be responsive to foreseeable public events (e.g. major sporting championships), and the due diligence process and mitigations should reflect this. Companies should also bear in mind wider industry experience (e.g. whether certain features – for example live streaming – are particularly risky) and good practice. Where human rights are involved in risk assessment and risk management, their special nature should be recognised, as the OECD due diligence guidance recognises. Companies should respect the need for diversity and inclusion in a risk assessment process so that issues – especially those which particularly affect minorities – are not overlooked or under-valued. This may be particularly relevant when products designed for operation in one state are then deployed in others.
6	Design missing from the risk assessment process - is "understanding the harms" both the offences and the functionality?? Or just the harm? Looks at likelihood and impact but doesn't focus on functionalities	"In our draft detailed guidance on methodology, we have proposed a process which reflects these four steps: i) understand the harms; ii) assess the risks; iii) decide measures, implement and record; and iv) report, review and update the risk assessment. We also include key common concepts from best practice which align to the risk assessment duties, such as: a) Assessing risk through a matrix of likelihood and impact; b) Assigning a risk level for each harm; and c) Considering residual risk after mitigating measures have been applied." (vol 3 9.52)	The New Mexico Attorney General court filings demonstrate clearly how design of Meta's platforms have allowed CSAM to flourish and how Meta has made a series of decisions not to deal with it. For example, see section VII "THE HARMFUL CONTENT ON META'S PLATFORMS REMAINS AND IS PROLIFERATED BY META'S ALGORITHMS" (para 174 onwards) which - in addition to documenting failures in age verification - finds that eg
7	Statement of larger users therefore larger impact - that's not risk-based	"As part of the risk level table, we also provide draft guidance on the effect of a service's user numbers on its level of risk. In general, all else being equal,	Glitch Digital Misogynoir Report (and other research cited there on this subject)

No	Issue	Consultation doc references	Evidence and commentary
	See also small vs large above, section 7.	the more users a service has, the more users can be affected by illegal content and the greater the impact of any illegal content. We have therefore proposed that services which reach certain user numbers should consider the potential impact of harm to be medium or high." (Vol 3, 9.59) **BUT this contradicts vol 2, 6F.31 re evidence on hate offences "However, there is evidence that niche online services can contain far more abuse, including hateful activity, than mainstream services, despite these services attracting far fewer users"** "We are clear in the Service Risk Assessment Guidance that in some instances the number of users may be a weak indicator of risk level. They need to be considered alongside other risk factors. It is possible for a large service to be low risk, and for a small service to be high risk, depending on the specific circumstances of each service" (vol 3, 9.62)	https://glitchcharity.co.uk/wp-content/uploads/2023/07/Glitch-Misogynoir-Report_Final_18Jul_v5_Single-Pages.pdf
8	Overview document says services don't have to assess risk of every possible offence occurring on service - but if they have evidence, they should consider this. Thoroughness of risk assessment – is this a tension with suitable and sufficient requirement?	"Services do however need to assess the risk of harm from relevant non-priority offences appearing on the service … this does not mean assessing the risk of every possible individual offence that is not a priority offence occurring on your service. However, if you have evidence or reason to believe that other types of illegal harm that are not listed as priority offences in the Act are likely to occur on your service, then you should consider those in your risk assessment." (Summary document Vol 1 2.33)	Is this what the Act says? While there are some distinctions between priority and non-priority offences, and 'other illegal content' is dealt with together, it does not in principle exclude categories of illegal content (see e.g. s. 9(5)(d) which seems to expect consideration across the board)

No	Issue	Consultation doc references	Evidence and commentary
9	Effective governance re all priority harms - what about non-priority?	Governance and accountability underpin the way that a service manages risk and ensures that efforts to mitigate them are effective. We consider that these processes are essential components of a well-functioning system of organisational scrutiny, checks and balances, and transparency around risk management activities. Effective governance and accountability processes should be effective in tackling all priority illegal harms Vol 3 8.13 “Effective governance and accountability processes should be effective in tackling all priority illegal harms” (Vol 3 8,13) Annex 10 Para A1: 30 “In recognition of the quantity and complexity of offences which could be included within the scope of the definition of ‘other’ offences, Ofcom has chosen to provide specific guidance on ‘other’ offences where they have been created by the Online Safety Act and do not wholly overlap with any priority offences.”	Non-priority offences that Ofcom covers here are: epilepsy trolling, self-harm, cyberflashing, false communications, threatening communications. But these are the offences introduced by the Act, not necessarily a complete list of those most likely to be relevant Does this mean that all non-priority offences are effectively excluded from the duties?
10	Independence of monitoring and assurance No specification of third party involvement, reliance on “evidence” from tech companies re what is going on already, and makes specifying this an issue re costs	Vol 3 8.102:”We do not envisage independence as requiring services to engage an independent third party (such as an external auditor) to confirm effectiveness of mitigations, although services may choose to do so” 8.106 “Mindgeek specified that internal audit included work related to process workflows, technical audit, and gap identification in compliance. “	How are they understanding internal controls? Should they map on to third party audit standards? Ref to MindGeek as good practice - has Ofcom checked how effective this is?? Or just been told by them?

No	Issue	Consultation doc references	Evidence and commentary
		Linking to costs 8.121: “The costs of this measure would be considerable, with the main cost being the ongoing staff costs to run the monitoring and assurance function. There may also be additional costs associated with wider training and awareness raising of the remit of an internal assurance function among existing teams who would be expected to feed into the work of the function.”	

SECTION 9: Violence Against Women and Girls

There are a number of new criminal offences proposed that address online VAWG, which are welcome. But the impact of all the strategic and policy decisions taken by Ofcom above will do little to shift the dial in terms of their overall safety online. Indeed the UN Special Rapporteur on Violence Against Women and Girls who, at the time of writing, has just finished a visit to the UK, said:

"While the enactment of the Online Safety Act is a welcome development, gaps remain, specifically around the issues of violence within the pornography industry, the influence of pornography on individual and societal attitudes towards VAWG and the impact of legal pornography on perpetration of child sexual abuse, both online and offline. We need to move away from companies self-regulating towards a legally enforced duty of care on tech companies across the distribution chain to ensure that they have adequate infrastructure to prevent tech abuse and to support survivors." (UNSR Summary of Preliminary Findings after visit to UK – 21 February 2024

Until the Government conceded on Baroness Morgan's amendment in the latter stages of the Bill's Parliamentary passage, the Government promised that the new offences would go a long way to improving protections for women and girls and that a separate code of practice was unnecessary. The opposite is true – and Ofcom's guidance on VAWG, which was the Government's concession, will not be consulted on for at least another year.

Further, as evidenced in Glitch's Digital Misogynoir Report, Black women continue to be disproportionately impacted by online abuse, and the online abuse directed towards Black women is interconnected with other forms of hate online, like antisemitism, Islamophobia and transphobia. While the OSA accounts for intersectionality, it remains to be see how those vulnerable to harm because of their intersectional identities will be protected; nor is it clear how Ofcom plans to develop and implement frameworks for ensuring Black women – and many other multiply-marginalised communities – do not fall through regulatory and legal gaps.

More detail on these concerns is provided in a detailed joint submission from organisations and experts in the VAWG sector, which we also support. We would also draw Ofcom's attention to the submission from Professor Clare McGlynn, from Durham University, which looks in detail at the consultation's proposals in relation specifically to the intimate image abuse, cyberflashing and extreme pornography offences.

SECTION 10: Gaps in protections

	Issue	Consultation doc references	Evidence and commentary
1	Section 127 and obscenity missing from harms - these will perform a mopping up role (eg abuse of footballers) But detailed guidance given by Ofcom is only on priority harms and random non-priority. **See full response section 2 on the illegal content judgements guidance.**	"Our initial Code of Practice on Illegal Harms will recommend services adopt protections to address all types of illegal content covered by the Act. " Approach p5	Section 127 of Comms Act: offences of "sending a message or other matter that is grossly offensive or of an indecent, obscene or menacing character" and "for the purpose of causing annoyance, inconvenience or needless anxiety to another" sends "a message that he knows to be false". Detailed guidance given by Ofcom is only on priority harms and random non-priority Section 127 of Comms Act: offences of "sending a message or other matter that is grossly offensive or of an indecent, obscene or menacing character" and "for the purpose of causing annoyance, inconvenience or needless anxiety to another" sends "a message that he knows to be false". NB OSA Schedule 6 refers to s 2 Obscene Publications Act (but in reference to children only), the fact that it is mentioned there means complete disregard later is the more noticeable. CPS guidance on that is here and, on relevance of section 127, see here from CPS
2	Search Clicking through thumbnails to harmful content is identified in risk profile document in a few places but then in the codes, there is no mention of a "one-click" limit	Vol 2 para 2.29 6U38: "Service design may in some instances facilitate the risk of illegal content being encountered and shared and therefore increase the risks of harm to users on U2U or search services. Offence-specific risks of harm associated with service design are outlined in different	Evidence recently demonstrated how deepfake porn was found just one click away via Google and Bing and Ofcom's own recent research has found similar with regard to self-harm content (research commissioned to inform the child safety code but which has direct relevance to design choices relating

	Issue	Consultation doc references	Evidence and commentary
		chapters of this Register, and the most prominent examples are in chapter 6D: Encouraging or assisting suicide or serious self-harm and chapter 6L: Extreme pornography. Such examples relate to how vulnerable users may be recommended content that is increasingly harmful and potentially illegal. Similarly, users may be led to illegal content within a few clicks from their query on a search service (for further information, see chapter 6T on risks of harm to individuals on search services)." 6U.50 "Further information as to how services can implement service design effectively on search services, and mitigate the risks described here, can be found in the Codes of Practice " Vol 4 13.5 "It is important to recognise that content is to be treated as 'encountered via' search results where it is encountered as a consequence of interacting with results (for example by clicking on them). This means that search content includes content on a webpage that can be accessed by interacting with search results. The safety duties, and the measures we recommend for the purposes of complying with them below, should be considered in this context."	to illegal content too.) Harm may be indirect. This also may be a particular issue for landing pages or review sites which make the route to illegal content clear; adverts for/discussion of tools (eg nudification apps) which are then used for illegal purposes. Harm may be indirect. This may be a particular issue for landing pages or review sites which make the route to illegal content clear; adverts for/discussion of tools (eg nudification apps) which are then used for illegal purposes.
3	Overview document says services don't have to assess risk of every possible offence occurring on service - but if they have evidence, they should consider this. Thoroughness of risk assessment – is	"Services do however need to assess the risk of harm from relevant non-priority offences appearing on the service … this does not mean assessing the risk of every possible individual offence that is not a priority offence occurring on your service. However, if you have evidence or reason to believe that other types of illegal harm	Is this what the Act says? While there are some distinctions between priority and non-priority offences, and 'other illegal content' is dealt with together, it does not in principle exclude categories of illegal content (see e.g. s. 9(5)(d) which seems to

	Issue	Consultation doc references	Evidence and commentary
	this a tension with suitable and sufficient requirement?	that are not listed as priority offences in the Act are likely to occur on your service, then you should consider those in your risk assessment." (Summary document Vol 1 2.33)	expect consideration across the board)
4	Excluding supply chain from risk assessment - very limited references to risks of supply chain/third party involvement despite recognition that many services will rely on third party software (or moderation services) in their business	Vol 2 8.97 "Requiring services to have measures to mitigate and manage illegal content risks audited by an independent third-party; d) Requiring due diligence of third-party contractors or providers of services involved in the mitigation and management of illegal content risks to assure their approaches lead to good online safety outcomes" Vol 3 12.22 "If they have automated technology at all it is likely to be trained by a third-party (i.e. 'off-theshelf' tools), rather than bespoke and/or specially trained automated technology." Vol 3 14.50 "We understand that third-party entities support perceptual hash matching, and it forms the basis of many in-house solutions developed by larger service providers. Some services discuss their use of perceptual hash matching technology and solutions publicly, such as through transparency reporting".	
5	Animal protections	Vol 2, 5.21 "At a fairly late stage in its consideration of the Bill which became the Online Safety Act, the offence in section 4(1) of the Animal Welfare Act 2006 (unnecessary suffering of an animal) was added to the list of priority offences. We will consult in due course on how we propose to include that offence in our Register. "	Does this mean that service providers effectively have no obligations with regard to this priority offence?

	Issue	Consultation doc references	Evidence and commentary
6	Lack of focus on Gen AI and metaverse	Summary p9 One ref para 3.60 Ability to edit visual imagery is included as a risk factor in the risk profiles	Harms are here already - what's timescale for including them? However the Government during the passage of the Bill was keen to emphasise how the approach was "technology neutral" and harms arising from these new technologies would be covered if it was user-to-user in nature. See, for example, Lord Parkinson in the Lords Committee stage debate on 25 May: "The Bill has been designed to be technology-neutral in order to capture new services that may arise in this rapidly evolving sector. It confers duties on any service that enables users to interact with each other, as well as search services, meaning that any new internet service that enables user interaction will be caught by it … the Bill is designed to regulate providers of user-to-user services, regardless of the specific technologies they use to deliver their service, including virtual reality and augmented reality content. This is because any service that allows its users to encounter content generated, uploaded or shared by other users is in scope unless exempt. "Content" is defined very broadly in Clause 207(1) as "anything communicated by means of an internet service". This includes virtual or augmented reality. The Bill's duties therefore cover all user-generated content present on the service, regardless of the form this content takes, including virtual reality and augmented reality content. **To state it plainly: platforms that allow such content—for example, the metaverse—are firmly in scope of the Bill.**" (Hansard 25 May col 1010) There is plenty of evidence already of harm from both technologies in the here and now. There was a particularly

	Issue	Consultation doc references	Evidence and commentary
			graphic debate in the Lords at Committee stage of the Online Safety Bill (indeed, so graphic that a group of school children were ushered out of the public gallery) on the sexual abuse of children within VR environments. And there have been numerous recent reports: see for example, the IET report on harm arising in virtual spaces; the NSPCC's detailed report on "Child Safeguarding and Immersive Technologies" and the recent news report of a virtual gang-rape of an under-16 in the metaverse. On Gen AI, Europol reported last year on its exploitation by criminals and Taylor Swift has recently been a very high-profile victim of deepfake porn. Yet Ofcom gives no timescales for how they are going to respond to this in future iterations of the codes and again, without the "catch-all" measure we recommend above, there is no obligation on services to take steps to address these harms in order to comply with their regulatory duties.
7	Equality Act - Welsh language posts not specifically mentioned (Ofcom obligation)? What about other minority languages in the UK?	As regards Welsh language, Ofcom says this in Annex 13 at A13.8 "More generally, we are proposing that services should have regard to the needs of their user base in considering what languages are needed for their content moderation, complaints handling, terms of service and publicly available statements. To this extent, we consider our proposals are likely to have positive effects or increased positive effects on opportunities to use Welsh and treating Welsh no less favourably than English."	New Mexico attorney general finding re Meta and CSAM: As with images, Meta's identification and blocking of terms associated with trafficking and CSAM are too narrow and rigid, and easily evaded, and do not adequately screen communications and terms in Spanish or other languages (Page 96) Also this report looks at the role of Facebook and Telegram in allowing incitement and online hate to spread in countries where a lack of moderators in the local languages was a factor. There will be a need for content moderation in multiple languages in UK - but there is no mention of this. This is surprising given that Ofcom has insight into some of the challenges here from its broadcasting role, including the tensions between different communities within the UK; for example, here and here.