WARNING: This consultation response contains language and/or material that may be distressing

Consultation response form

Please complete this form in full and return to protectingchildren@ofcom.org.uk.

Consultation title	Consultation: Protecting children from harms online
Organisation name	Online Safety Act Network

Your response

Question	Your response
Volume 2: Identifying the services children are using Children's Access Assessments (Section 4).	
Do you agree with our proposals in relation to children's access assessments, in particular the aspects below. Please provide evidence to support your view. 1. Our proposal that service providers should only conclude that children are not normally able to access a service where they are using highly effective age assurance? 2. Our proposed approach to the child user condition, including our proposed interpretation of "significant number of users who are children" and the factors that service providers consider in assessing whether the child user condition is met? 3. Our proposed approach to the process for children's access assessments?	**The responses to this and following questions are drawn from our full written response to the consultation, which we recommend is read in its entirety. For ease of analysis, however, we have extracted relevant material here in response to the specific sections of this consultation.**
Volume 3: The causes and impacts of online harm to children **Draft Children's Register of Risk (Section 7)**	
Proposed approach: 4. Do you have any views on Ofcom's assessment of the causes and impacts of online harms? Please provide evidence to support your answer. a. Do you think we have missed anything important in our analysis? 5. Do you have any views about our interpretation of the links between	**The responses to this and following questions are drawn from our full written response to the consultation, which we recommend is read in its entirety.** Response: While we commend Ofcom's assessment of the causes and impacts of online harms, we are concerned that these do not fully pull through to the codes of practice. We provide extracts from our response here and refer to the table we have submitted at Annex A and linked here **Recommendation**

Question	Your response
risk factors and different kinds of content harmful to children? Please provide evidence to support your answer. 6. Do you have any views on the age groups we recommended for assessing risk by age? Please provide evidence to support your answer. 7. Do you have any views on our interpretation of non-designated content or our approach to identifying non-designated content? Please provide evidence to support your answer. **Evidence gathering for future work:** 8. Do you have any evidence relating to kinds of content that increase the risk of harm from Primary Priority, Priority or Non-designated Content, when viewed in combination (to be considered as part of cumulative harm)? 9. Have you identified risks to children from GenAI content or applications on U2U or Search services? a) Please Provide any information about any risks identified 10. Do you have any specific evidence relevant to our assessment of body image content and depressive content as kinds of non-designated content? Specifically, we are interested in: a) (i) specific examples of body image or depressive content linked to significant harms to children, b. (ii) evidence distinguishing body image or depressive content from existing categories of priority or primary priority content. 11. Do you propose any other category of content that could meet the	Before discussing our concerns regarding Ofcom's approach to the risk profiles and evidence, we provide upfront here our recommendation, as per our response to the illegal harms consultation, that the following wording is inserted in the draft codes <u>for both illegal harms and protecting children</u>, between the section on governance and accountability and the section on content moderation. (This follows the order of areas in which measures should be taken identified in section 10 (4) and section 27 (4) (on illegal harms) and 12(8) and 29 (4) (child safety duties).) This recommended measure is referred to elsewhere in this proforma response. --- **"Design of functionalities, algorithms and other features** **Product testing** For all services, suitable and sufficient product testing should be carried out during the design and development of functionalities, algorithms and other features to identify whether those features are likely to contribute to the risk of harm arising from illegal content on the service. The results of this product testing should be a core input to all services risk assessments. **Mitigating measures** For all services, measures to respond to the risks identified in the risk assessment should be taken, including but not limited to, providing extra tools and functionalities, including additional layers of moderation or prescreening, by redesigning the features associated with the risks, by limiting access to them where appropriate or where the risk of harm is sufficiently severe by withdrawing the function, algorithm or other feature. Decisions taken on mitigating measures, as part of the product design process or as a response to is-

Question	Your response
definition of NDC under the Act at this stage? Please provide evidence to support your answer.	sues arising from the risk assessment, should be recorded. (Note: this would be included in the record keeping duties under section 23 (u2U) and section 34 (search).) **Monitoring and measurement** All services should develop appropriate metrics to measure the effectiveness of the mitigating measures taken in reducing the risk of harm identified in the risk assessment. These measures should feed back into the risk assessment." -------- The obligation here is to have a mechanism to consider how to mitigate, rather than requiring the use of particular technologies or the introduction of predetermined safeguards in relation to technologies. Significantly, and given the proposal is based on the duty of care, the measure of success is not wholly about output measures (though they may indicate whether an effective process is in place) but about the level of care found in outcome-oriented processes and choices. Assessment is about the features taken together and not just an individual item in isolation. Given that, the outcome may not be wholly successful; what is important, however, is the recognition of any such shortfall and the adaptation of measures in response to this. It may be that the language of the obligation should recognise that the measures proposed should be appropriate bearing in mind the objective sought to be achieved (in the sense that an arguable claim can be made about appropriateness rather than there being pre-existing specific evidence on the point). We note that Ofcom has proposed criteria for assessing the effectiveness of age verification criteria (technical accuracy, robustness, reliability and fairness) that are more about outcomes than specific outputs; it may be that analogous criteria could be introduced to assess the processes adopted to identify harms to select appropriate mitigation measures. Significantly, the extent of the testing and assessing obligation

Question	Your response
	should be proportionate, bearing in mind the provider's resources, reach and severity of likely impact on groups of users. The lack of reach and the less complex internal environment should of course mean that in any event the process will be less sizable for smaller providers than larger. **Ofcom's evidential threshold** This is a reiteration of the concerns we raised in response to the illegal harms consultation about the weight given by Ofcom to the amount of evidence already collected to support the proposals e.g. 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). To be clear, we are not suggesting that there should be obligations to take measures that are ineffective; rather that where there is some evidence of effectiveness but lots of evidence of harm, the precautionary principle should kick in. It would then be for the service to prove or disprove the appropriateness of the measures and for Ofcom to use this practical evidence to change the recommendation or add additional measures. Unfortunately, the approach taken by Ofcom reinforces the status quo, setting a "lowest common denominator" based on specific compensatory measures within a piecemeal, process-driven regime, rather than one that designs in safety and is focused on the outcomes described in the Act. **What the Act says** The Act makes no mention of the evidence on which Ofcom must base its recommendations for measures in the codes. There is a requirement that the measures must be technically feasible (Schedule

Question	Your response
	4, section 2 (c)) and age verification has some standards about effectiveness (Schedule 4, section 12 (3)). In terms of proactive tech, Ofcom is required to "have regard to the degree of accuracy, effectiveness and lack of bias achieved by the technology in question" and may refer to industry standards". (Schedule 4, section 13 (6)) **Parliamentary debate** The growing weight of evidence of the nature and prevalence of online harms was a significant driver in the Government's decision to legislate, announced in May 2018. The opportunities for evidence to be submitted – from industry as well as the academic and civil society research communities – to influence the scope of the policy development and the legislation were provided at many stages between 2017 (the publication of the Government's Internet Safety Strategy Green Paper) and Royal Assent. These included pre-legislative scrutiny by a Joint Committee in 2021 of the draft Online Safety Bill and then Committee stages during the Parliamentary passage of the Bill between 2022-2023. A summary of, and links to, the Parliamentary stages is provided here and related research and commentary during that period is summarised here. Numerous Parliamentary inquiries on related topics took place during this time, each one accumulating more evidence via written submissions and oral testimony. **Ofcom's proposals** Evidence has been crucial to the decisions Ofcom has made, both as regards the risk register in Volume 3 and the underpinning analysis for the codes of practice in Volume 5. Ofcom sets out in volume 5, para 14.11 that "Both the Illegal Content Codes and the Children's Safety Codes protect children. The illegal content safety duties protect children from illegal content and the children's safety duties protect children from harmful content other than illegal content. Accordingly, several measures proposed for the Children's Safety

Question	Your response
	Codes build on proposals in the Illegal Content Codes. In the areas of user reporting and complaints, governance and accountability, content moderation (U2U and Search), user support and terms of service, some of our proposed measures closely mirror proposals for the Illegal Content Codes." Given this repetition - and because we still feel that the approach to evidence is problematic across Ofcom's proposals - we repeat in full our analysis from the illegal harms consultation, updated with references to the children's consultation specifics. As in the illegal harms consultation, Ofcom sets out that it has considered the evidence by reference to certain criteria: "method, robustness, ethics, independence and narrative" (vol 3, para 7.35). It provides further information on these criteria, including the methodology of the studies, size and coverage, ethics (e.g. handling of personal data), whether stakeholder interests might have influenced findings and whether the commentary in the output matched the data found. By contrast, there is no such clear methodology for Volume 5 (and the methodology in Vol 3 is expressed so as only to apply to Vol 3). There is also a question as to whether the standards required for an academic research project should be the benchmark for policymaking in this area because so much has not been investigated, not been proven or cannot be proven due to complexity; moreover, studies tend to focus on functionalities in isolation rather than in context. Yet, if a problem is created or exacerbated by a combination of functionalities and how they are used, why would we expect one change to be a silver bullet? Again, we refer back to the merits of a "by design" safety obligation on companies to develop their own measures to address the risks it can see (via its own evidence) arising on their services. We note that the children's consultation document has a more considered approach to how much evidence is required in order for Ofcom to make a

Question	Your response
	judgement on whether to recommend a measure or otherwise in its code: For example: > "Working with imperfect evidence means that we face uncertainty when making our recommendations, with some decisions being finely balanced. Online services in scope of the Act, and the technologies they use, are evolving rapidly – and new harms may emerge. There is a need for prompt action to protect children online and a clear risk that children will not be protected if we only recommend measures where we have extensive and definitive direct evidence of effectiveness. Therefore, some of our proposed measures are based on an assessment of more limited or indirect evidence of impact, and reliance on logic-based rationales. We exercise regulatory judgement in prioritising measures which, on balance, we consider can materially improve children's safety online. In some cases, where we provisionally conclude that certain measures should not be recommended at this stage, or only recommended for some services but not others, we intend to consider this further as we review the responses to this consultation and as part of our future work." (para 14.34) However, there is a heavy reliance throughout the consultation document on statements from companies providing regulated services. "Best practice" examples are cited. But in many other areas, Ofcom refers to "limited" or "patchy" evidence for measures that work. This is particularly important given the increasing evidence from whistleblowers (e.g. Frances Haugen, Arturo Bejar) and from litigation in the States (provided elsewhere in this response) that some of the biggest social media companies have suppressed evidence and – it is claimed – sought to mislead both users and legislators. We include some of this evidence below. We appreciate that Ofcom has only recently received its information-gathering powers - though as noted above, we are surprised that they have not

Question	Your response
	yet been used (para 14.27). We note that the regulator intends to use them to expand its evidence base in order to inform future iterations of the codes. In volume 6 of the illegal harms consultation, Ofcom said "The statutory information gathering powers conferred on Ofcom by the Act give us the legal tools to obtain information in support of our online safety functions. These powers will help us to address the information asymmetry that exists between Ofcom and regulated services and to discover, obtain and use the information we need, including for monitoring and understanding market developments, supervising regulated services, and investigating suspected compliance failures." This is welcome. But we make two observations: firstly, it is not clear how Ofcom has determined how evidential thresholds had been satisfied, especially in relation to Volume 5 of this consultation. We also note that there are some concerns about whether solutions are proven to be effective, but we do not see a discussion of what the threshold is for that. For example: "As part of our scoping exercise, we considered the role of functionalities such as autoplay in amplifying the risk of harm but decided not to propose any specific recommendations at this stage given the more limited evidence on the role of autoplay in amplifying exposure of children to harmful content compared to other functionalities like recommender systems." (Vol 6, para 13.72) "At this stage, we do not have evidence that concerns about confidentiality are a barrier to complaining to providers of search services. We are therefore not proposing to recommend this measure for search services at this time." (18.124) "We note that some services offer users a range of comment control tools. These are beyond the options we have considered

Question	Your response
	here. While we are supportive of these tools as a means of empowering users to exercise more control over comment functionalities, at this stage we have limited evidence around more granular controls, and have concerns given the risk of unintended consequences with regard to uneven impacts on freedom of expression and likely higher implementation costs." (21.108) [NB this last extract is in relation to functionality that is *already being offered by some services*; the fact that Ofcom does not then go on to recommend as something that all relevant services should do in order to build the evidence base on their effectiveness, it could - given the "safe harbour" status of the codes - mean that those services that currently offer comment control tools withdraw them.] While there is a clear rationale for not recommending proven ineffective measures, this approach is worrying where there is some evidence of effectiveness. Moreover, absence of evidence is not evidence of ineffectiveness and responses in respect of which there is no evidence should not be excluded from the field of possible measures. More worryingly, Ofcom has also used lack of evidence in relation to its assessment of costs to justify the non-inclusion of tools in relation to smaller services. This begs the question as to why they have created this threshold for themselves when it so clearly prevents the recommendation of mitigation for a known, evidenced harm. Not only is there a question as to the appropriate evidence threshold, but the problem could have been avoided had Ofcom started from the premise that companies should address the issues arising from their risk assessment systemically or based on outcomes, rather than via a specific measure, and by a focus on safety by design as well as the relevant action required by the Act in relation to designated content, whether illegal or as covered under the children's duties. This issue seems to have been a result of the approach taken to the sort of measures recommended. See

Question	Your response
	also our discussion on the measures in the codes of practice. This approach is likely to significantly limit the likelihood that there will be much material change in the online safety of users when these first codes of practice are published. Indeed, as we suggest above, it could potentially lead to a rowing back of some measures already deployed by services because they do not need to continue to resource them in order to comply with the codes. In this context, we were concerned to hear an Ofcom Principal describe, on a webinar addressed to businesses during the illegal harms consultation phase, how Ofcom's evidence threshold was in effect a bar to them codifying measures which are already accepted by regulated companies as "good practice" and how voluntary principles were all that they could rely on in many areas as a result. > "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." (our emphasis) (WE Communications webinar: Navigating Tech Regulation in the Wake of the Online Safety Act – 31 January 2024; this extract is at 36 minutes in) A further point that has been omitted entirely from consideration is that absence of evidence of a proposition is not proof that that proposition is not true. We also note that where there is presumptive harm, especially harm which is serious in nature and wide

Question	Your response
	reaching – as has been clearly evidenced by Vol 3 – that both Parliament in its debate and the overarching duty of care principle would dictate a more precautionary approach. Ofcom's position here is therefore not what would have been anticipated: > "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." (Illegal harms consultation; Vol 4 11.14) **Evidence** We refer back now to the work we quoted from extensively in the illegal harms consultation on the merits of the precautionary principle to help make regulatory interventions in a fast-moving environment where evidence might be lacking or as yet unambiguous, including work for Carnegie UK and the ILGRA paper on the precautionary principle. As we set out above and in our previous submission, there is also plenty of evidence from recent court filings and whistle-blower material that the big platforms have ample internal evidence on the harmful design of their products and the decisions that would/should be taken to mitigate that. While Ofcom may not feel that it has – at present – evidence to support the recommendation of specific measures for all in-scope services to mitigate these harms, it is very likely that the biggest companies do but have chosen not to develop, test or deploy these measures. (Indeed, as far back as 2017, one of Facebook's co-founders, Sean Parker, admitted that they knew when developing the site that the objective was "How do we consume as much of your time and conscious attention as possible?" It was this mindset that led to the creation of features such as the "like" button that would give users "a little do-

Question	Your response
	pamine hit" to encourage them to upload more content. It's a social-validation feedback loop … exactly the kind of thing that a hacker like myself would come up with, because you're exploiting a vulnerability in human psychology." (Reported in the Guardian) If the codes (as we discussed above) do not compel companies to comply with anything beyond the specific measures recommended therein, then there is no regulatory imperative and therefore no consequence for those services if they don't. This underlines the importance of having an upfront catch-all measure in the codes on illegal content that requires companies to act on the knowledge they may already have about the harmful design effects of their products, notwithstanding the need also to adopt the evidence-based measures that Ofcom includes in the rest of the codes. **Evidence, risk and the precautionary principle - a case study: Generative AI** There are many studies that identify the risks posed to children by GenAI and immersive technologies. Indeed, Ofcom recognises this and provides the following summary in volume 3, with links to research studies: "There is evidence which shows that GenAI can facilitate the creation of content harmful to children, including pornography, content promoting eating disorders, and bullying content, which is then shared on U2U services. Evidence shows there has been a pronounced increase in the availability of AI-generated pornography online, particularly on pornography services which are dedicated to AI-generated pornography and which could be accessed by children. We have found evidence showing that GenAI models can create eating disorder content, which has in some instances been shared on U2U services such as eating disorder discussion forums. There is also evidence of GenAI

Question	Your response
	models being used to create content to bully and threaten individuals including 'fakes' of individual's voices, which is shared on U2U services and could be encountered by children. There is also emerging evidence indicating that GenAI models can create other kinds of harmful content which could be shared on U2U services and encountered by children. For example, audio and language GenAI models can produce racist, transphobic, violent remarks and religious biases ('abuse and hate') and engage in self-harm dialogue, even where unsolicited ('suicide and self-harm')" Prior to setting out this summary, Ofcom had noted that "children are early adopters of new technologies, and GenAI is no exception". So, one would expect that there would be a measure requiring companies that use GenAI in their products and services, or that host content that may have been created by GenAI, to take account of their risk assessment relating to the harms that this might cause and take appropriate steps - especially as this would be a new feature and not already built in. But there is no such measure. Instead, despite the evidence of harm that Ofcom has already provided, it says that "the evidence base for children's interaction with harmful AI-generated content on U2U and search services will be limited". It goes on "We are also aware that the risks associated with GenAI models may not yet be fully known. However, given the rapid pace at which the technology is evolving, we must not underestimate the expected risks associated with GenAI for children. As new evidence emerges over the coming years, we will update this Register appropriately." There is evidence of harm occurring now but Ofcom suggests doing nothing until new evidence emerges over "the coming years". This

Question	Your response
	is absolutely where a precautionary approach - as proposed by our recommended code of practice measure - would be appropriate, putting the responsibility on the services where GenAI might create harm to children to take measures to prevent that harm. This approach would in itself, then help to create an evidence base from which Ofcom could draw on to develop best-practice recommendations for future codified measures, resulting in a positive feedback loop focused on improving safety, rather than a void in which harm will continue to proliferate and evolve until such time as Ofcom has defined the appropriate response. Not only would this limit harm but also save Ofcom time and resources down the line. **Recommendation** We believe that, based on the analysis above, the addition of the proposed additional measures – as set out at the start of this question, would address the problems we have identified. This approach avoids the risk of Ofcom effectively requiring something of companies that is ineffective and inefficient and is in line with the "precautionary principle" approach to regulation in other sectors where there are safety risks.
Draft Guidance on Content Harmful to Children (Section 8)	
12. Do you agree with our proposed approach, including the level of specificity of examples given and the proposal to include contextual information for services to consider? 13. Do you have further evidence that can support the guidance provided on different kinds of content harmful to children?	Confidential? N **The responses to this and following questions are drawn from our full written response to the consultation, which we recommend is read in its entirety.** We refer Ofcom to the section in our full response on safety by design and our concerns about Ofcom's content-focused nature of the assessment of risk

Question	Your response
14. For each of the harms discussed, are there additional categories of content that Ofcom a) should consider to be harmful or b) consider not to be harmful or c) where our current proposals should be reconsidered?	and harm. In our response to the illegal harms consultation, we included analysis which we had published as a standalone blog on Ofcom's approach to the illegal content judgements guidance. We don't intend to rehearse or repeat the arguments again here but make a couple of observations about how far this may have influenced - in a way that is not required by the Act - Ofcom's approach to PPC, PC and NDC in the children's duties and the decisions it has made in relation to design-based measures in the codes. While the Act itself is problematic, in its designation of content in those three categories, it refers in slightly different ways across PPC, PC and NDC or "content of a kind" (Act ref). Ofcom, conversely, refers to "*examples of kinds of conten*t" (eg para 8.20) which is a much more specific description bringing service's attention to individual pieces of content rather than "types" of content. This inevitably leads to an ex-post perspective on harm - eg, does this individual piece of content fit one of the categories in the Act and how was it dealt with by the service provider? Rather than, how does the service design lead to the creation, promotion and engagement with "content of a particular type" in a way that is harmful to children? This perspective is the one which the ICJG proposes in relation to criminal offences. This may have been understandable in the context of the ICJG and the concerns about the mental element (though we still have concerns about the precise approach adopted) but there is no similar requirement for mental element here. Instead the emphasis is on the likely impact on users, which is looking at the prediction of harm arising from classes of material. Furthermore, in our view it is an approach that is not appropriate given that the taking down individual pieces of harmful content is not a requirement for compliance. We would also draw Ofcom's attention to the point made in relation to VAWG:

Question	Your response
	As part of its general duties under s 3(4) Communications Act, Ofcom has considered the position of people beyond children who are vulnerable but the regulator provides no details as to which groups were considered and how that consideration affected Ofcom's output - especially given the different experience of men and women online (taken generally). (see Vol 5 14.23) Ofcom - in Volume 3 (the causes and impacts of harms to children) - also recognises in many instances that there is a gendered risk of harm and that girls are disproportionately more likely to be impacted by some harms than boys. For example: > "Most evidence suggests that girls are at higher risk than boys of being targeted by bullying content online, especially by certain kinds of bullying content. A recent study by Internet Matters, among 13-16-year-old girls, found that they had received and observed 'hateful comments' on popular social media platforms. These were in response to both content they had posted and content posted by others, and typically targeted girls' appearance such as clothes, weight or bodies, which participants said impacted on their wellbeing. The participants attributed the comments to men and boys and noticed a lack of similar comments on boys' videos." (Vol 3, 7.54) Ofcom also recognises the fact that those in other minoritised groups and with intersecting characteristics are also likely to experience some harms and that indirect harm can be caused to women and girls through the proliferation of misogynistic views (6.4, 7.4.26-29, 7.4.38 et seq, 7.6.38), including the specific issue of harmful sexual behaviours and attitudes (7.1.19). We question, however, whether the measures pick all the problematic issues up. There is a notable omission of misogynistic content in the section on abuse and hate (section 8.6) given that Andrew Tate is mentioned elsewhere and his influence is having an increasing impact on attitudes towards girls and female teachers in schools and a

Question	Your response
	wider societal culture of hatred towards girls and women. The focus on age-gating porn (and other primary priority content) may deal with one clearly relevant set of content-based issues but this leads to heavy reliance on a single point of possible failure - ie the effectiveness of the age verification/estimation technology used to keep children off the platform - rather than addressing some of the underlying issues that arise from the design of the platform itself and how its features and functionalities exacerbate the risk of content-based harm. (See also the reference in 15.173 to the fact that violent content (designated as "priority content", with services required by use of age assurance measures "to ensure that children are protected from encountering" it) "can include violence against women and girls which does not meet the threshold of illegality.") More broadly, we would defer to experts from across civil society – including the children's sector, VAWG sector, mental health and suicide prevention experts etc – on the specifics of Ofcom's consultation questions here.
Volume 4: How should services assess the risk of online harms? **Governance and Accountability (Section 11)**	
15. Do you agree with the proposed governance measures to be included in the Children's Safety Codes? a) Please confirm which proposed measure your views relate to and explain your views and provide any arguments and supporting evidence. b) If you responded to our Illegal Harms Consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response.	Confidential? N **The responses to this and following questions are drawn from our full written response to the consultation, which we recommend is read in its entirety.** The governance and risk assessment proposals draw heavily on the same approach in the illegal harms consultation. Given the influence of the literature on corporate governance and risk assessment, we remain concerned about whether this is orientated towards safety by design and - as previously - the

Question	Your response
16. Do you agree with our assumption that the proposed governance measures for Children's Safety Codes could be implemented through the same process as the equivalent draft Illegal Content Codes?	absence of learnings from product safety approaches. There remains a significant reliance in Ofcom's proposals on what platforms are already doing in terms of what they assess might be possible and/or should be recommended. It is not clear that Ofcom has determined that what these platforms are doing is a) effective; and b) enough to deliver their duties under the OSA. This links to the burden of proof point we set out above in response to questions 4-7. As in the illegal harms risk guidance, some of the outcomes for the steps in the children's risk assessment draft Guidance (annex 5) seem to go to process (e.g. you will have read this document) rather than objectives of the process (have you identified the relevant risks)? Again, it is predicated (along with governance proposals) on the basis that companies are doing this already and therefore won't need to incur more costs. Governance structures, along with robust risk assessment processes, are fundamental to influencing product design choices with a view to reducing the risk of harm. So, Ofcom's proposals here are crucial to the overall effectiveness of the Online Safety Act regime. **Parliamentary debate** The prominence of the risk assessments in the Government's intentions for the regulatory regime are seen in, for example, Lord Parkinson's statement at Lords Report on 6 July 2023: > "That is why the legislation takes a systems and processes approach to tackling the risk of harm. User-to-user and search service providers will have to undertake comprehensive mandatory risk assessments of their services and consider how factors such as the design and operation of a service and its features and functionalities may increase the risk of harm to children. Providers must then

Question	Your response
	put in place measures to manage and mitigate these risks, as well as systems and processes to prevent and protect children from encountering the categories of harmful content." (Hansard 6 July 2023 col 1384) Also, "the list of functionalities in the Bill is non-exhaustive. There may be other functionalities which could cause harm to users and which services will need to consider as part of their risk assessment duties. For example, if a provider's risk assessment identifies that there are functionalities which risk causing significant harm to an appreciable number of children on its service, the Bill will require the provider to put in place measures to mitigate and manage that risk." (Hansard 6 July col 1382) **Note that this last statement specifically puts the obligation on service providers - not Ofcom - to work out which measures are appropriate for mitigation.** Elsewhere, in part of a debate on end-to-end encryption, Lord Parkinson referred to the fact that "companies will need to undertake risk assessments, including consideration of risks arising from the design of their services, before taking proportionate steps to mitigate and manage these risks. Where relevant, assessing the risks arising from end-to-end encryption will be an integral part of this process". He went on to say that the risk assessment process used in "almost every other industry" and said that "it is right that we expect technology companies to take user safety into account when designing their products and services" (Col 1320). **Ofcom's proposals** We refer Ofcom to our previous submission and our broad concerns about the risk assessment proposals, which we do not intend to repeat in full here, except to note the same marked reliance on "best practice" in risk management (largely focused on corporate governance and reputational risk, not product safety and harm minimisation) and on industry evidence as to what they do already/what

Question	Your response
	works already with no qualitative assessment as to whether it is effective and/or sufficient. We would however want to emphasise the following points that are specific to the children's consultation. • The Risk Assessment guidance itself has been restructured so as to be more accessible in this consultation than in the illegal harms consultation. What this has done is expose further how much the process of risk assessment is - in Ofcom's approach - a tick-box exercise. The list of things to cover are literally presented as tasks to complete, not outcomes to aim for in terms of improvements to the service or the mitigation of risks. There is no requirement for product testing, red teaming, safety-by-design interventions or the consideration of evidence taken from R&D operations. The guidance allows services to record that they've done something but not what is the actual measure/outcome/change that flows from it. • We would question why Ofcom does not feel that the approach to risk assessment relating to children's protection should be different and/or more robust than the approach set out for illegal content. While we understand that consistency between the two processes is desirable, to reduce burdens on services, it is unfortunate that there has been no specific tailoring to the specific way in which risks arise on platforms relating to children. For example, Ofcom uses the same examples relating to safeguarding that have been drawn from other sectors; these are relevant to managing the risks of harms to children within organisations but not to the risks of harms to children arising from the services or products that are created *by* those organisations. Significantly in this regard, there is no route for people (like Arturo Bejar when he was working for Instagram) who are seeking from *within* organisations to flag risks to children's safety arising from their services or products - this seems to be a gap in protection mechanisms.

Question	Your response
	• Ofcom seems to confuse (in 11.140) horizon-scanning with capturing evidence of new/emerging harms after they have already happened (e.g. via complaints, or information relating to the death of a child). This isn't forward-looking enough for harms that can become prevalent very quickly, particularly when - elsewhere - Ofcom refer to the fact that children are early adopters of new technologies. The OSA's requirement for a higher level of protection for children than adults is not being met when the risk assessment expectations for both sets of users are the same and largely predicated on a retrospective approach to governance oversight - reviewing the *process* of risk management, rather than acting on what the risk management information is telling the Board. • Similarly, while 11.147 sets out the need to have a "mechanism to notice new trends", there is no related governance responsibility for this nor any measures in the codes to do anything about the information that the company might collect through this mechanism. It is also unclear why small, single-risk services are exempt from this tracking - the very tracking mechanism that might highlight to them that they are *no longer* single risk, particularly when they will be under a duty to notify Ofcom of NDC. Given the simplicity of the service implied by single risk it is also likely that tracking trends should be comparatively straightforward. **Evidence** We refer Ofcom again to the paper submitted at annex F of our previous response: a paper prepared by Peter Hanley and Gretchen Peters that argues for Ofcom to shift its approach to a "product assured safety management" approach which would "encourage safety rather respond to risk, and stop problems before the emerge rather than cleaning them up afterwards". This builds on their expertise and experience in other sectors and is in line with the principles that underpin the UK's Health and

Question	Your response
	Safety at Work Act 1974. We also published a blog on Ofcom's approach to governance in the light of a Wired interview with Del Harvey - the former head of Trust and Safety at Twitter (now X). In it, Harvey talks about some of the things that concerned her during her time in her role. She gives the example of trying to escalate within the company the potential threat from a DM she had received suggesting that Twitter's offices should be bombed: there was no route within the company to do this for such tweets. Harvey says: > "It was the same issue that it always has been and always will be, which is resourcing. I made requests in 2010 for functionalities that did not get implemented, in many instances, till a decade-plus later." She also gives the following example: "Multiple account detection and returning accounts. If you're a multiple-time violator, how do we make sure you stop? Without going down this weird path of, "Well, we aren't sure if this is the best use of resources, so instead, we will do nothing in that realm and instead come up with a new product feature." Because it was growth at all costs, and safety eventually." Finally, and crucially, she says: "When trust and safety is going well, no one thinks about it or talks about it. And when trust and safety is going poorly, it's usually something that leadership wants to blame on policies. Quite frankly, policies are going to be a Band-Aid if your product isn't being designed in a way that actually doesn't encourage abuse. You've got to plan there, guys." [emphasis added] 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

Question	Your response
	Code of Practice as evidence and the Ad Hoc Advice to the United Nations Special Rapporteur on Minority Issues which focuses on risk assessment. (pp 7-11), which we provided extracts from previously. **Recommendation** While Ofcom has carried out an extensive review of the literature on risk assessment, we would recommend that further advice is sought on the many experts available who understand how best to carry this out – particularly with regard to product safety testing – in sectors that have a similar obligation with regard to the safe design and operation of their products and services. We also suggest – as per our upfront recommendation re adding a new measure to the codes - that product testing should be a mandatory part of the risk assessment process, even if discretion is given to services on the way in which they undertake this.
Children's Risk Assessment Guidance and Children's Risk Profiles' (Section 12)	
17. What do you think about our proposals in relation to the Children's Risk Assessment Guidance? a) Please provide underlying arguments and evidence of efficacy or risks that support your view. 18. What do you think about our proposals in relation to the Children's Risk Profiles for Content Harmful to Children? a) Please provide underlying arguments and evidence of efficacy or risks that support your view. Specifically, we welcome evidence from regulated services on the following: 19. Do you think the four-step risk assessment process and the Children's Risk Profiles are useful models to help	Confidential? –N **See response to questions 4-11 and 15-16 above which covers our analysis of both the risk profiles and risk assessment guidance.**

Question	Your response
services understand the risks that their services pose to children and comply with their child risk assessment obligations under the Act? 20. Are there any specific aspects of the children's risk assessment duties that you consider need additional guidance beyond what we have proposed in our draft? 21. Are the Children's Risk Profiles sufficiently clear and do you think the information provided on risk factors will help you understand the risks on your service? a) If you have comments or input related to the links between different kinds of content harmful to children and risk factors, please refer to Volume 3: Causes and Impacts of Harms to Children Online which includes the draft Children's Register of Risks.	
Volume 5 – What should services do to mitigate the risk of online harms **Our proposals for the Children's Safety Codes (Section 13)**	
Proposed measures 22. Do you agree with our proposed package of measures for the first Children's Safety Codes? a) If not, please explain why. **Evidence gathering for future work.** 23. Do you currently employ measures or have additional evidence in the areas we have set out for future consideration? a) If so, please provide evidence of the impact, effectiveness and cost of such measures, including any results from trialling or testing of measures.	Confidential? –N We provide here the relevant sections from our full response on safety by design and on the gap between the risk profiles and the draft children's safety codes. We wish to draw Ofcom's attention upfront to the recommendation we make for an additional measure to be inserted into the codes which we include above and also, in its entirety, here. **Recommendation** We suggest the following wording is inserted in the draft codes <u>for both illegal harms and protecting children</u>, between the section on governance and accountability and the section on content moderation, which follows the order of areas in which

Question	Your response
24. Are there other areas in which we should consider potential future measures for the Children's Safety Codes? a) If so, please explain why and provide supporting evidence.	measures should be taken identified in section 10 (4) and section 27 (4) (on illegal harms) and 12(8) and 29 (4) (child safety duties). --------- **"Design of functionalities, algorithms and other features** **Product testing** For all services, suitable and sufficient product testing should be carried out during the design and development of functionalities, algorithms and other features to identify whether those features are likely to contribute to the risk of harm arising from illegal content on the service. The results of this product testing should be a core input to all services risk assessments. **Mitigating measures** For all services, measures to respond to the risks identified in the risk assessment should be taken, including but not limited to, providing extra tools and functionalities, including additional layers of moderation or prescreening, by redesigning the features associated with the risks, by limiting access to them where appropriate or where the risk of harm is sufficiently severe by withdrawing the function, algorithm or other feature. Decisions taken on mitigating measures, as part of the product design process or as a response to issues arising from the risk assessment, should be recorded. (Note: this would be included in the record keeping duties under section 23 (u2U) and section 34 (search).) **Monitoring and measurement** All services should develop appropriate metrics to measure the effectiveness of the mitigating

Question	Your response
	measures taken in reducing the risk of harm identified in the risk assessment. These measures should feed back into the risk assessment." -------- The obligation here is to have a mechanism to consider how to mitigate, rather than requiring the use of particular technologies or the introduction of predetermined safeguards in relation to technologies. Significantly, and given the proposal is based on the duty of care, the measure of success is not wholly about output measures (though they may indicate whether an effective process is in place) but about the level of care found in outcome-oriented processes and choices. Assessment is about the features taken together and not just an individual item in isolation. Given that, the outcome may not be wholly successful; what is important, however, is the recognition of any such shortfall and the adaptation of measures in response to this. It may be that the language of the obligation should recognise that the measures proposed should be appropriate bearing in mind the objective sought to be achieved (in the sense that an arguable claim can be made about appropriateness rather than there being pre-existing specific evidence on the point). We note that Ofcom has proposed criteria for assessing the effectiveness of age verification criteria (technical accuracy, robustness, reliability and fairness) that are more about outcomes than specific outputs; it may be that analogous criteria could be introduced to assess the processes adopted to identify harms to select appropriate mitigation measures. Significantly, the extent of the testing and assessing obligation should be proportionate, bearing in mind the provider's resources, reach and severity of likely impact on groups of users. The lack of reach and the less complex internal environment should of course mean that in any event the process will be less sizable for smaller providers than larger. **Safety by design**

Question	Your response
	We provide here the general commentary and analysis on this issue as context for our response on the codes of practice framework in the next section. We have also included the specific examples (eg on the recommender system, content moderation and age assurance) from this section in other relevant sections of this consultation proforma. We noted in our previous submission the relatively late insertion of a new "section 1" in the Online Safety Act, setting out the overall objectives of the legislation, including a duty on providers to ensure that services are "safe by design". As with our previous submission, we provide evidence - often interlinked - throughout this document that provides evidence of the choices that Ofcom has made which – taken together – we believe will not deliver this stated outcome. **What the Act says** Section 12 (8) describes the children's safety duties and mirrors section 10 (4) in the illegal content duties. Schedule 4 (the Online Safety Objectives) is relevant here, as are the new duties on Ofcom, set out in section 91, which amend Section 3 of the Communications Act 2003, including: (2) In subsection (2), after paragraph (f) insert— "(g) the adequate protection of citizens from harm presented by content on regulated services, through t*he appropriate use by providers of such services of systems and processes* designed to reduce the risk of such harm" (our emphasis) **Parliamentary debate** In our previous submission, we provided relevant extracts from Hansard where the integral nature of a "safety by design" approach was emphasised by Peers, including Lord Parkinson - the Government Minister - who introduced the new "clause 1" by saying that it was the Government's intent that "a

Question	Your response
	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." (Hansard 6 July column 1320) The "by design" approach raises the question of whether, where there is evidence of harm connected to particular features, the obligation should be on the companies to be the subject to the burden of rectification – even to the point of rolling back specific features (e.g. push notifications which have given rise to concerns about addiction in the US) until the evidence is there to make them safe enough: product withdrawals are known in other industries and indeed TikTok recently suspended a feature on its new Lite App in response to an investigation into its child safety impacts under the European Digital Services Act. **Ofcom's proposals** Ofcom's Approach document, published alongside the illegal harms consultation last November, says "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." (p5). This is heartening – and reflects the Government's intention, as set out in Parkinson's above statement. But - as the approach and measures in the children's consultation mirror those set out in the illegal harms consultation - it is worth repeating here that this objective does not flow through the subsequent proposals (including the approach to governance and risk assessment, proportionality decisions and the differentiated approach to size) nor to the codes themselves. Our analysis of their proposals starts with the two new buckets of measures that are included in the children's consultation - on age gating and the rec-

Question	Your response
	ommender system - and then moves on to the features and functionalities that are identified as causing harm in the risk profile volume (volume 3) but which are not covered in the measures. (Our table at annex A provides an at-a-glance comparison.) Age gating In the children's Summary document (p13 onwards), Ofcom sets out the "safer platform design choices" that it is consulting on: "We are also proposing a range of safety measures that focus on service providers ensuring they make foundational design choices, so children have safer online experiences. These cover three broad topics: • understanding which users are children so that those children can be kept safe; • ensuring recommender systems do not operate to harm children; and • making sure content moderation systems operate effectively. With the exception of the proposals around the recommender systems (which is welcome), these topics - and the measures related to them which we discuss below - do not go much further than the ex-post measures Ofcom set out in the illegal harms consultation. In fact, two-thirds of the 36 measures recommended for U2U platforms, and all but one of the 24 measures for search services, are the same or equivalent versions. Age assurance - e.g. keeping children off platforms - is a tool to prevent harm but not a "safety by design" choice that fundamentally changes the platform itself for all users, whether they are children or not. We refer Ofcom here to the analysis by 5 Rights/Children's Coalition of the age assurance proposals. Content moderation is about dealing with

Question	Your response
	content that is already posted rather than addressing the system which it flows over. In the Proposed codes at a glance, the description of measures highlights how they are limited to cutting off access to the service to children (by age assurance) for PPC content and some PPC, then to cut off access at more granular content level using age assurance, then to use age verification to assess recommender system usage, plus content moderation. This is not safety-by-design but the application of safety tech on top of a system that is deemed to be harmful to the users that the regulatory framework is designed to protect (and at a higher level than adult users, too). We discuss the age assurance measures in more detail in response to questions 31-35 below. <u>Recommender systems</u> The measures relating to the recommender system - while welcome and integral to a platform or service's design - still relate largely to the content that flows over the system and that is promoted by its algorithm rather than the deployment of a recommender system itself. The recommender system may not be a problem, per se: it's how it's designed, the values it incorporates and the way it is used by the service provider. The consultation also does not consider how recommender systems form part of the suite of incentives for content creation (see also our commentary on business models, below) and how being picked up by the algorithm is important for advertising revenue and other promotions. Moreover, it is relatively far down the design stack in terms of its impact. We have concerns here that this narrow approach will ultimately be a missed opportunity, resulting in piecemeal impacts on children with little shift in the culture of safety within companies and the overall safety of products used by children, particularly those in vulnerable groups with shared characteristics.

Question	Your response
	In the introductory sections to volume 3 (risk register), Ofcom's description of recommender systems highlights the problems: "The functionalities and characteristics we describe as risky are not inherently harmful and can have important benefits. For example, recommender systems benefit internet users by helping them find content which is interesting and relevant to them. The role of the new online safety regime is not to restrict or prohibit the use of such functionalities or characteristics, *but rather to get services to put in place safeguards which allow users to enjoy the benefits they bring, while managing the risks appropriately*." (our emphasis) (vol 3, page 4) It is not clear what "safeguards" mean here. Is this post-hoc, after content has been created? If so, this is not "safety by design" - it implies that the recommender system will run as previously but overlaid with interventions to meet the measures required in the codes. In that regard, Ofcom's approach does not fit with what is in the Act or in the risk register. In the next section, we also look at how the business model affects the creation and promotion of harmful content - intersecting with the recommender system in a way that is about system design choices as much as the motivation of the individual content creators. Ofcom describes this interplay in para 7.12.5: "The choice architecture of a service (i.e. the design of the choice environment in which a user is making decisions) can be *designed to influence or manipulate users into acting in ways that serve commercial interests but may be detrimental to individual or societal interests (e.g. spending time engaging with the service, in the case of advertising revenue models*)" (our emphasis) <u>Business model</u> The business model is referred to in the risk assessment and risk profiles - and more emphasis is given to it than in the previous consultation - but no consideration is given in the codes of practice to

Question	Your response
	measures to mitigate or curtail the commercial incentives for content creation (eg clickbait farms or harmful influencers (such as Andrew Tate) where content is used as a means to make money for the creators and often constitutes their sole purpose for being on the platform. In the risk register, Ofcom specifically mentions the recent rise in influence of Andrew Tate in its discussion of the financial incentives to create and share harmful content and, notably, how the monetisation incentive combines with the recommender system to result in harmful content being pushed to younger users without their prior engagement: > "Such content can be created by ordinary users or by content creators. Content creators typically earn money on social media from advertising, in proportion to their number of followers. This means they face similar financial incentives to services, whose revenue depends on number of users and/or user engagement, and so they can be incentivised to create harmful or extreme content, if such content drives their followers and hence their earnings. Services are then incentivised to recommend such engaging content to users (including children) to sustain their revenue. For instance, the evidence shows that hateful and misogynistic videos posted by content creators can be popular on social media and are recommended to young users without them having proactively 'liked' or searched for such content." (7.12.7) In addition, Ofcom acknowledges that: "Due to the nature of risk, we also distinguish two ways in which goods or services may be promoted on a service. This distinction was made because in some cases services are paid to promote content as 'advertisements' which represent a source of revenue. In contrast, while users can promote goods and services by posting them for sale, in many cases the service

Question	Your response
	is not paid to advertise them. The risks associated with how a service generates revenue differ according to which functionalities are offered to users and how they might be used." (para 7.30) But there is a "third way" here - that of content creators being incentivised by financial reward (the monetisation of content) to create ever more controversial, provocative or potentially viral content with a view to increasing their revenue. This is not addressed in the measures. Finally, the advertising-based model is specifically mentioned in relation to eating disorders (potentially one of the categories of non-designated content): "Advertising-based business models may increase the risk of children encountering eating disorder content. Services which optimise revenue based on user base and engagement have incentives to develop service designs and features that maximise engagement and drive revenue, even if this is at the expense of exposing child users to harmful content. As set out earlier in this section, eating disorder content can generate high engagement, especially within eating disorder communities." (Also Vol 3, para 7.3.101) <u>Metrics</u> Linked to the business model - and particularly the incentives for content creators to maximise engagement - design choices relating to metrics and their impact on children's content exposure and creation are identified as a function that is potentially harmful but are not covered by the mitigating measures. For example: "Ofcom research also reported that many children, and particularly those seeking social validation or looking to build their online following, said they shared violent content to gain popularity, due to the high levels of engagement that violent content would typically gain. Others reported that some of their friends shared violent content as they thought it was "funny" to surprise them with it."

Question	Your response
	(Volume 3, para 7.6.11) Volume 3 also notes the influence of "likes" in the incentivisation of children to take part in dangerous stunts (see 7.8.10 and 7.8.14). <u>Addictive design</u> There is some interesting evidence presented in volume 3 (section 7.13) in relation to the impact of design choices - including infinite scroll and autoplay, and alerts and notifications - on the time spent by children online. This is linked to the issues above relating to the business model (incentivisation for content creators) and also to the use and influence of metrics on user engagement. But there are no corresponding measures to mitigate it in the codes of practice despite the fact that Ofcom clearly states that: "Evidence suggests that the greater the time spent on services by a child, the higher the risk of encountering any harmful content that may be present on that service. Some service features and functionalities are designed to influence certain behavioural outcomes, such as high usage or specific kinds of engagement. Children may be particularly vulnerable to being influenced in this way." (p245) Ofcom goes on to say: "We understand that these features and functionalities can be fundamental to how services operate, and a significant source of revenue for services in proportion to their number of users and/or user engagement. This might include encouraging users to spend money on a particular service, or in the case of advertising-based business models, simply spend time engaging with a particular service while being exposed to ads." (para 7.13.3) This comment suggests that the explanations given to Ofcom by service providers about the nature of their service are (as with other evidence) being taken at face value: that addictive design is an integral part of social media services and, in order to

Question	Your response
	comply with the children's safety duties, some kind of "safety tech" fix must be retrospectively applied to mitigate the harm, rather than imposing a requirement on the services to address the design at source. (We refer back to the recent DSA example mentioned above, where action by the Commission temporarily stopped a new feature on TikTok that had addictive design elements.) Both metrification and addictive design are linked directly to the way in which recommender systems work - part of a wider suite of features and functionalities that drive engagement and keep users on platforms. Ofcom refers again to this aspect in its risk assessment guidance: "Further, in our research into features and functionalities we understand that affirmation based features play an outsized role in children seeking social validation through online services because they facilitate children receiving affirmation from others, and can lead to children spending more time online. It follows that services introducing changes which impact the prevalence of these functionalities could lead to more children spending more time on the service which could amount to a significant change in risks posed to children." (Volume 4, 12.100) Yet there are no measures, or even an open requirement to act upon the identification of harm arising from these features or functionalities (or combination thereof), to address it. As with much of the work across both risk profile volumes, Ofcom has identified quite specifically how these features and functionalities are part of the problem the OSA is trying to solve but then has done nothing on this via the codes. In the absence of evidence that Ofcom deems suitable to inform the recommendation of measures to

Question	Your response
	address these features and functionalities, an alternative approach would be to turn them off by default for children - using the age gating measures as the means by which to apply this default. There is evidence that children don't like the addictive design elements of their social media experience. Such a measure would not make services unviable, just less profitable. We refer Ofcom to the court filings in the US relating to the Californian case on adolescent social media addiction and to the advisory from the US Surgeon General in May 2023 (Social Media and Youth Mental Health). Size of company With specific reference to measures that could be seen as touching on "safety by design" (including written statements of responsibilities or expectations of product testing), Ofcom makes an upfront judgement that these can only be reasonably expected of large or multi-risk companies – thereby undercutting at the outset the overarching legislative objective in the Act. Significantly, in the proposals set out on governance in volume 4, Ofcom - in a proposal that it acknowledges "mirrors an equivalent one in the illegal harms consultation" (para 11.89) - sets out that a written statement of responsibilities for senior members of staff would: "include ownership of decision-making and business activities that are likely to have a material impact on children's online 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 or behavioural biases have been taken into account), how data related to children's online safety is collected and processed, and how humans and

Question	Your response
	machines implement trust and safety policies. Depending on a service's structure, key responsibilities in children's 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." (Vol 4, 11.87) However, as with the illegal harms consultation, this statement of responsibilities is only recommended for large or multi-risk services despite the acknowledgment that "decision-making and business activities are likely to have a material impact on user safety outcomes", which goes to the heart of safety by design. Indeed, as we set out below, the Government's Impact Assessment makes reference to the fact that building in safety by design is a way for smaller platforms to reduce regulatory compliance costs. Ofcom itself has recognised that smaller providers are likely to have less complex systems which would suggest safety by design would be - in process terms - less complex than for larger operators. Ofcom also only makes a few brief references to product safety testing, which we would include as a component of an overall "safety by design" approach. In Volume 3, Ofcom says: "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). Then, in a table suggesting a number of "enhanced inputs" to help companies build up their "risk assessment evidence base", "results of product testing" are included: "We use 'product' as an all-encompassing term that includes any functionality, feature, tool, or policy that you provide to users for

Question	Your response
	them to interact with through your service. This includes but is not limited to whole services, individual features, terms and conditions (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." (Table 9.5) (Our emphasis) This is an "enhanced input": an expectation for larger services only. Ofcom's rationale for this distinction between "core" and "enhanced" inputs is: "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) This therefore means that not only is product testing to ensure user safety not expected of smaller companies, it is not something that Ofcom feels should be carried out as part of a risk assessment to inform the measures that smaller services might feel they need to take in order to make their products safe. (We set out more on the implications of the differentiated approach to size in Ofcom's proposals below.) Implicitly in this, Ofcom is seeing severity of harm as being about the number of people affected, not the severity of harm caused, an approach which is not necessarily mandated by the

Question	Your response
	Act but which occurs repeatedly throughout the consultation. This seems to run counter to a "safety by design" approach. It is in marked contrast to the approach of the CMA and the ICO who suggest in a joint paper that testing is key to prevent harmful design in choice architecture; the paper notes that there are different ways of testing. The resources available to a service provider could thus inform the sort of testing rather than the question of whether service providers should test. Content-focused measures We make a final point here about the content-focused nature of the assessment of risk and harm. In our response to the illegal harms consultation, we included analysis which we had published as a standalone blog on Ofcom's approach to the illegal content judgements guidance. We don't intend to rehearse or repeat the arguments again here but make a couple of observations about how far this may have influenced - in a way that is not required by the Act - Ofcom's approach to PPC, PC and NDC in the children's duties and the decisions it has made in relation to design-based measures in the codes. While the Act itself is problematic, in its designation of content in those three categories, it refers in slightly different ways across PPC, PC and NDC or "content of a particular kind" (eg Section 41). Ofcom, conversely, refers to "*examples of kinds of conten*t" (eg para 8.20) which is a much more specific description bringing service's attention to individual pieces of content rather than "kinds" of content. This inevitably leads to an ex-post perspective on harm - eg, does this individual piece of content fit one of the categories in the Act and how was it dealt with by the service provider? Rather than, how does the service design lead to the creation, promotion and engagement with "content of a particular kind" in a way that is harmful to children?

Question	Your response
	This perspective is the one which the ICJG proposes in relation to criminal offences. This may have been understandable in the context of the ICJG and the concerns about the mental element (though we still have concerns about the precise approach adopted) but there is no similar requirement for mental element here. Instead the emphasis is on the likely impact on users, which is looking at the prediction of harm arising from classes of material. Furthermore, in our view it is an approach that is not appropriate given that the taking down individual pieces of harmful content is not a requirement for compliance. ## Evidence ### Safety by design The evidence we would like to draw Ofcom's attention to here is the same as that submitted in our response to the illegal harms consultation. It includes: • The Government's 2021 guides on "safety by design" for online platforms, unreferenced in Ofcom's material, which set out that this was a "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.. By considering your users' safety throughout design and development, you will be more able to embed a culture of safety into your service." Ofcom makes no reference to this work in its risk profile evidence (volume 3), though it does quote extensively from DCMS-commissioned research from Ecorys on the impact of online harms to children. • The Government's own Impact Assessment, which says "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

Question	Your response
	from the start thereby minimising compliance costs." • The Australian e-Safety Commissioner's Safety By Design principles • The OECD's recent report on safety by design for children. • Children's coalition, 5 Rights and NSPCC consultation responses **Harmful Design** The evidence we would like to draw Ofcom's attention to here is the same as that submitted in our response to the illegal harms consultation, including recent US court 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. These include: US court filings • State of NY, Erie County vs Meta et al re radicalisation - March 2024 • 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 Media Addiction - May 2023 Whistleblower material • Arturo Bejar in conversation with Stephen Balkam - FOSI conference - June 2024 • Arturo Bejar testimony to Congress - November 2023

Question	Your response
	• 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: Daniel Tucker - February 2024 • 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: Callie Lewis - December 2019 Transparency reports • Digital Services Act Transparency database

Developing the Children's Safety Codes: Our framework (Section 14)	
25. Do you agree with our approach to developing the proposed measures for the Children's Safety Codes? a) If not, please explain why. 26. Do you agree with our approach and proposed changes to the draft Illegal Content Codes to further protect children and accommodate for potential synergies in how systems and processes manage both content harmful to children and illegal content? a) Please explain your views. 27. Do you agree that most measures should apply to services that are either large services or smaller services that present a medium or high level of risk to children? 28. Do you agree with our definition of 'large' and with how we apply this in our recommendations? 29. Do you agree with our definition of 'multi-risk' and with how we apply this in our recommendations? 30. Do you agree with the proposed measures that we recommend for all services, even those that are small and low-risk?	Confidential? – Y / N As we described in detail in our response to the illegal harms consultation, we have concerns that 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. We refer to our large evidence table at annex A which compares the functionalities identified in volume 3 with the measures (or lack thereof) to address them in volume 5. The extracts below provide further context to this. Just as with the risk profile work in the illegal harms consultation, volume 3 of the suite of children's documents is a commendable standalone document and is analytical and thorough in identifying the functionalities that contribute to this prevalence and/or risk of harm to individuals from the categories of content designated in the OSA. Many of these functionalities are vectors for multiple harms. However, there is the same structural problem with the illegal harms proposal in that this assessment does not flow through to the mitigation measures set out in the Codes of Practice (Annex 7) (for user-to-user services) and Annex 8 for search, which focus primarily on ex-post measures (content moderation) - with the exception of the new age assurance measures and measures relating to the recommender system. Again, the rules-based nature of the Codes - specifying specific recommended measures rather than obligations aimed towards the achievement of desired outcomes - and the fact that these are designed as a "safe harbour" (eg if companies follow the measures they will be judged to have complied with

	their duties under the Act*), means that there is no incentive for companies to implement mitigating measures beyond those described in the codes. This is the case even if their risk assessment has flagged that their service poses particular risks from other functionalities (arising from design choices) and despite the fact that the risk assessment notes the need for voluntary actions over and above what is set out in the codes. The Atlantic Council makes this point: "if compliance replaces problem-solving, it establishes a ceiling for harm reduction, rather than a floor founded in user and societal protection." (p 36) (*The "safe harbour" provision is described here: > "Services that choose to implement the measures we recommend in Ofcom's Children's Safety Codes will be treated as complying with the relevant children's safety as well as their reporting and complaints duties. This means that Ofcom will not take enforcement action against them for breach of that duty if those measures have been implemented. This is sometimes described as a "safe harbour. However, the Act does not require that service providers adopt the measures set out in the Children's Safety Codes, and service providers may choose to comply with their duties in an alternative way that is proportionate to their circumstances ." (Para 13.4)) Furthermore, smaller companies are in many instances exempt from implementing particular mitigating measures due to Ofcom's proportionality analysis. (See our response to question 58 below.) We have produced a supporting document (annex A) to illustrate where the gaps between the analysis of harm and the recommended mitigations of it lie, along with a summary "at a glance" table. We have previously published a blog discussing the choices made in relation to the illegal harms codes of practice and compliance, which we also draw from below.

What the Act says

Section 12(4) includes at (b) design of functionalities, algorithms and other features, all of which – as we set out below – are lacking measures in this first iteration of the codes. The significance of the Codes is seen in section 49, which envisages two ways in which in-scope providers can comply with their relevant statutory duties: (a) compliance through recommended measures; and (b) compliance through alternative measures, but with caveats. Section 49 states that a service provider:

> "is to be treated as complying with a relevant duty if the provider takes or uses the measures described in a code of practice which are recommended for the purpose of compliance with the duty in question."

This means that service providers which choose to implement measures recommended to them for the kinds of content and their size or level of risk indicated in the regulator's Codes will be deemed as compliant with the relevant duty and Ofcom will not take enforcement action for breach of that relevant duty against those services. The level and nature of Ofcom's recommendations are therefore significant for the level of safety provided to users and the extent to which the Act's objectives are achieved.

In the event of identifying potential risks in services that are not adequately addressed by the existing Codes, and where transparency measures prove ineffective, Ofcom has the authority to update and enhance the Codes (see sections 47(1) and 48 of the Act) - a point which Ofcom recognises when it notes that the development of the Codes will be an iterative process. This, of course, has the disadvantage of introducing further delays to the effective implementation of the regime.

Schedule 4 provides further requirements about the measures to be included in any codes, as we discuss below.

Parliamentary debate

	In Lords Committee stage day 1, the Government Minister Lord Parkinson said: "Through their duties of care, all platforms will be required proactively to identify and manage risk factors associated with their services in order to ensure both that users do not encounter illegal content and that children are protected from harmful content. To achieve this, they will need to design their services to reduce the risk of harmful content or activity occurring and take swift action if it does". (Column 725) At Lord Committee stage day 3, in response to a debate on the nature of cumulative harm, Lord Parkinson said: > "The Bill will address cumulative risk where it is the result of a combination of high-risk functionality, such as live streaming, or rewards in service by way of payment or non-financial reward. This will initially be identified through Ofcom's sector risk assessments, and Ofcom's risk profiles and risk assessment guidance will reflect where a combination of risk in functionalities such as these can drive up the risk of harm to children. Service providers will have to take Ofcom's risk profiles into account in their own risk assessments for content which is illegal or harmful to children. The actions that companies will be required to take under their risk assessment duties in the Bill and the safety measures they will be required to put in place to manage the services risk will consider this bigger-picture risk profile." (Lords Committee stage 27 April 2023 Column 1385) Later in Lords Committee stage, when challenged by Baroness Morgan as to why the Government would not concede on a code of practice for women and girls, Lord Parkinson set out a number of reasons why the existing codes would be sufficient in this regard. He also replied directly to Morgan's claim that the Bill "misses out the specific course of conduct that offences in this area can have" and referred to (then) clause 9 re services needing to mitigate and

	manage the risk of being used for the commission or facilitation of an offence. Parkinson said: "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." **Ofcom's proposals** As in the illegal harms consultation (largely because the bulk of the measures are the same), Ofcom has in the main interpreted "measures described" as requiring very specific recommendations to which proportionality and costs criteria have to be applied on an individual basis before they can be "recommended for the purpose of compliance". Ofcom is pre-assessing proportionality here to limit the scope of the measures recommended, rather than allowing services to make their own assessments. This section repeats the analysis we provided in our previous consultation. It is fundamental to what we perceive as the problem in Ofcom's approach and one which we feel is still not fully understood. We submit that Ofcom's chosen approach is not required by the Act and does not reflect Parliamentary intention. One implication of section 236(1) in this context is that the obligations to take or use measures – notably those set out in non-exhaustive lists under sections 12(8) for user-to-user as well as 29(4) for search services - are not limited to specific types of technology but extend to processes as well. A requirement for an obligation to be clear and precise (Schedule 4, para 2b) does not mean that a service provider should have no choice or discretion in responding to the obligation; rather what it means is that the service provider should be able to understand the nature of the requirement. Ofcom is not precluded from imposing process requirements and offering illustrative examples of good or best practices when making recommendations of a procedural nature. Indeed, it is arguable that Ofcom could

	make more use of objective-focussed process obligations to cover gaps in mitigations that are currently found in the recommended measures. There are many instances where a functionality has been found to be problematic in Vol 3 and for the purposes of the risk register, but where Vol 5 finds the evidence of those solutions not to be specific enough to justify making a specific technical recommendation. An approach based on broader process-based obligations orientated towards the Act's objectives could also be within the scope of Section 49(1) which would allow a much more flexible orientation towards user safety while still satisfying the requirements for clarity and precision and allowing for proportionality of response. As we set out in our response to questions 4-7 above, throughout the consultation document, Ofcom makes its own judgements – without qualification – about a) what evidence it deems to be acceptable to support the inclusion of measures in the codes of practice (we talk further about evidence thresholds in response to questions 4-7, above); and b) what measures it deems proportionate for services to implement to mitigate the harms they may have already identified in their risk assessment. While there is some methodology set out in Volume 3 about what evidence they have accepted for the purpose of the risk register, for Volume 5 (the codes) there is no equivalent. This is a different issue from when the threshold has been reached - and why. The wording of the Act, however, does not imply that this is for Ofcom to judge – rather that it is for providers to "take or use measures … if it is proportionate to do so" (s 12 (8)). Despite this, Ofcom is taking a rules-based, prescriptive, de minimis approach to safety, which does not take into account the fact that the Act itself says the duties apply across all areas of the service "including the way it is designed, operated as well as used" and that the duties "require the provider to take or

	use measures" in areas, including "regulatory compliance and risk management arrangements", "design of functionalities, algorithms and other features". On the impact of proportionality, we refer to our response to question 58 below. We understand that Ofcom is taking a cautious approach with regard to the obligations imposed on companies - if not as regards the harms continued to be experienced by children - that it is reliant on evidence and that its proportionality assessment is stringent. However, there is a fundamental choice that has been made - integral to the illegal harms approach and therefore repeated here - about the approach to the codes that does not fit with the legislative intent: the regime was supposed to be principles-based or risk-based. While Schedule 4, para 1(a) does require Ofcom to "consider the appropriateness of provisions of the code of practice to different kinds and sizes of Part 3 services and to providers of differing sizes and services", it does not have to pre-judge all the measures it recommends on that basis nor is it required to set down specific rules. While there are expectations that obligations should be clear (and not impose unnecessary obligations on service providers) this does not mean more general obligations cannot be imposed. Indeed, as Lord Parkinson remarked; > "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." (Lords Committee stage 25 April 2023) It is reasonable as the regulator to place an expectation on the companies to respond to outcome-defined obligations. Ofcom's Economic Director, Tania Van Den Brande set out the problems with a rules based approach in 2021:

	"..rules are at a greater risk of leading to undesirable effects if a given conduct can be harmful, neutral or beneficial depending on the circumstances of the market or the characteristics of the firm they apply to.....Rules can also become outdated in highly dynamic markets." Despite the amount of evidence Ofcom has collected on the nature of harm, its decision to follow a rules-based model of recommendations has significantly limited the likelihood that companies will take a risk-based approach to mitigation. Furthermore, the rigid rules-based approach then requires Ofcom to decide, based on its proportionality assessment, that it should exempt smaller services from following those rules – rather than specifying an outcome or a principle and judging whether the regulated service has acted proportionately in its response. We discuss the issue relating to small companies further in response to questions 27-30; but deciding whether or not to apply code of practice measures to all companies, based on Ofcom's own assessment of the "onerous" (a word used, thankfully, fewer times in this consultation than previously) impact they might have on their profitability, is entirely inconsistent with Ministerial expectations that the Act's safety duties would apply to all regulated services, regardless of size – with the proportionality test being for companies to judge and account for to Ofcom, rather than Ofcom making that decision for them upfront. **Evidence** We set out our evidence on this disconnect between the harms identified and the measures proposed to address them in the updated table at annex A, which is attached to this submission as a PDF and which can be found on our website here. With the exception of recommender systems and age assurance, the measures recommended in the children's codes of practice mirror those in the illegal harms codes. There are a few additional points

	we would like to make in this regard, to supplement the comparative work provided in the annex. This is largely to highlight the gaps in measures, where we feel these are not justified, particularly when the codes are intended to deliver a "higher protection" for children. • There is no justification for **measures on livestreaming** to be omitted in relation to children given the number of types of harm it is linked to. Rather weakly, Ofcom argues (in volume 3 para 7.17) that "while livestreaming can be a risk factor for several kinds of harm to children, as it can allow the real-time sharing of content such as suicide and self-harm, it also allows for real-time updates in news, and can provide children with up-to-date tutorial videos and advice or encourage creativity in streaming content. These considerations are a key part of the analysis underpinning our Code measure." A small amount of benefit is used to make the case against a measure to mitigate a large amount of harm. Ofcom might understandably not want to "ban livestreaming" for children, but there would be interventions (aligned with the precautionary approach we advocated at Carnegie UK, see questions 4-7) that could introduce friction into its use. Friction would not prevent the positive use cases continuing (eg, educational broadcasts - though there is no evidence that educational content has to be live-streamed or that there is inherent value to be gained from doing that by contrast to other forms of audiovisual dissemination) while the negatives (children livestreaming themselves doing dangerous stunts, self-harming, or engaged in violent activities) could be minimised. Notably, a number of such practical measures were set out by DCMS, back in 2021, when it included guidance for companies on livestreaming in its "Principles of Safer Online Platform Design". Ofcom makes no reference to this in its proposals, nor does it consider the distinction between the issues around children having the ability to

	livestream versus the ability to receive content that is livestreamed; arguably these raise different issues in relation to harm. • Two other new functionalities have been identified in the risk register as posing specific harms to children but which were not included in the illegal harms analysis: **stranger pairing** and **ephemeral messaging**, neither of which have corresponding measures. Other functionalities that crop up multiple times in relation to multiple PPC or PC risks but with no mitigating measures recommended include: **hashtags, group messaging**, **direct messaging** and **anonymous profiles**. • There are no measures to address some of the risks relating to the **business model**s (as per our analysis in section 1), despite these being identified as something that the services' risk assessments must cover (eg "Assess the level of risk of harm to children and how that is affected by characteristics of a service and how it is used, including: user base, functionalities, algorithmic systems, and the business model"; para 2.30) • The incentives for children to chase likes or other visible metrics and incentives - another non-financial engagement aspect - is not addressed. • There is no requirement on platforms to do anything or make any modifications to the way their service is operating based on **feedback from children**, despite the fact that Ofcom recognises that "certain service characteristics play an important role in children's experiences of harm online" and that children themselves are aware that "any engagement, including reporting and signalling negative engagement could lead to similar content being recommended". (Vol 3, para 6.10) • **Large group messaging:** While the measures in the codes allow children to refuse invitations to groups, there are no considerations of systemic actions that regulated services might take when aware of the presence of large groups containing children on their

	platforms. For example, should they consider what content is being posted, what the connection is between the children, how many adults are also involved, etc? Also, regarding the observation at vol 5, 21.62 that "evidence suggests that the main risks of being unwillingly added to group chats by others are related to pornographic content, eating disorder content, bullying content, abuse and hate content and violent content", there is a wider consideration as to whether adding or inviting children to groups should be allowed as a functionality per se, regardless of whether there is enough evidence about which types of harmful content they might be exposed to. At the very least, the measure relating to their ability to refuse invitations should be applied to all services. • **Emerging technologies - metaverse, genAI etc:** We noted in our illegal harms consultation that the Government, during the passage of the Bill, said it was "technology neutral" and that harms arising from new technologies (such as the metaverse, immersive technologies or GenAI) would be covered if they were 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) As we noted in the illegal harms response, there is plenty of evidence already of harm from both technologies in the here and now - including child sexual abuse within VR environments and a virtual gang-rape of an under-16 in the metaverse. Deepfake porn has risen up the agenda and fraud is also a significant area of concern. In the illegal harms consultation, there was no indication from Ofcom of the 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. The same concerns arise here. The metaverse is mentioned in volume 3 in relation to exposure to porn (7.1.13). Ofcom identifies the **risks arising from Gen AI** (particularly the links between immersive environments and bullying, vol 3, para 7.5.60) and its links to both eating disorder content (7.3.57) and bullying (7.5.87). Ofcom notes that children are early adopters of new technologies and "gen AI models can present a risk of harm to children", para 7.14.22).

It is also noted as a risk factor in relation to search, "as these tools can both return indexed results, as described above, and generate novel content in response to prompts, which could be considered harmful to children." (7.10.5) See also para 7.14.27 for a full summary of the evidence available of the risks GenAI pose to children. We look specifically at GenAI in our case study provided in response to question 4-7 above.

Despite this, Ofcom concludes that "the evidence base for children's interaction with genAI will be limited" and does not suggest a corresponding measure.

Size of platforms

Despite the children's code duties applying to all services (if they are likely to be accessed by children), regardless of size, Ofcom's recommended measures in the codes of practice do not apply equally to all of them. Instead, as in the illegal harms consultation, they are differentiated according to size and then differentiated further based on the services' own risk assessments.

Ofcom's tear sheet sets out "at a glance" its proposals and who they apply to. The explainer Ofcom published towards the end of the illegal harms consultation stressed (again) the iterative nature of the codes. As with their chosen approach to mitigating measures, we are concerned that this means a "lowest-common denominator" baseline for the codes when they come into force – and one which in many areas may even risk weakening existing protections.

We also do not think that Ofcom's approach to proportionality and size is justified by the legislative framework nor reflects the intention of Parliament.

At the risk of repeating ourselves, we set out our concerns again with reference to material from the children's consultation proposals.

	Parliamentary debate Throughout the development of the Bill, Government Ministers were at pains to stress that all platforms would be covered by the duties relating to protection of children. Here, for example, is former DCMS Minister Chris Philp at the Second Reading of the Bill in the Commons in April 2022: "all platforms, regardless of size, are in scope with regard to content that is illegal and to content that is harmful to children. (Hansard link here) As we can see from the duties in the Act above, there is much stress on "proportionate" measures – which Government Ministers, in Parliament, were also at pains to emphasise when challenged on the number of businesses that were potentially within scope of the legislation. For example, Lord Parkinson – in response to an amendment proposed by Baroness Fox, to exempt small services – said the following at Lords Committee stage: "My Lords, I am sympathetic to arguments that we must avoid imposing disproportionate burdens on regulated services, but I cannot accept the amendments tabled by the noble Baroness, Lady Fox, and others ..... 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, as the noble Baroness, Lady Kidron, rightly echoed.... The Bill has been designed to avoid disproportionate or unnecessary burdens on smaller services … 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."

	Despite that recognition, it is also clear that proportionality was not intended as a vehicle to undercut protection; rather it acknowledged the need to recognise the risk of harm posed by the service. We discussed in our previous response the intersection with the Parliamentary debates on categorisation of services, in particular where the threshold would be set for "category 1" services with respect to their extra duties. This is not relevant to the children's consultation - the child access assessment is the prerequisite for compliance with the children's safety duties - but the arguments put forth there still apply to the decisions being made about differential duties for services within the children's codes of practice: > "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… 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." (Lord Parkinson at Lords Report Stage 19 July 2023) We see below that – by mirroring the proposals from the illegal harms consultation in the children's consultation – Ofcom is indeed, from the outset of the regulatory regime, giving small companies many excuses for not dealing with illegal content as well as content harmful to children. **Ofcom's proposals** Ofcom says in its summary document: "We recognise that the size, capacity, and risks of services differ widely, and we therefore do not take a one-size-fits-all approach. Instead, we have set out what types of service we think should use specific safety measures to comply with their duties, with the most extensive expectations on the riskiest services."

	Yet, despite the very strong commitments from the Government, Ofcom is exempting small and/or single risk services from many of the measures in the codes on the grounds of proportionality and cost. This compounds the fact that these services are also in effect let off carrying out a robust risk assessment: if they don't assess their own risk adequately (meaning risks might be under- assessed resulting in a lower risk classification for Ofcom's framework), and they also don't have to comply with all the measures in the codes, the small-but-risky services will not be required to address the children's safety duties appropriately. Ofcom do acknowledge however that "Our framework for defining the kinds of services in scope of each measure, including with reference to size and risk thresholds, is broadly similar to that adopted for our Illegal Harms Consultation. We have not yet processed all responses to our 2023 Illegal Harms Consultation and it is possible that in light of these responses we may make adjustments to this framework in future." (14.51) The definition of large companies is the same in both the illegal harms and children's proposals; equivalent to the DSA definition VLOPs – 7 million monthly users in the UK (vol 4, 14.57). Ofcom goes on to say that "Our proposed definition of a large service captures services with the widest reach among UK children. Nevertheless, we recognise that the size of the total UK user base is not a precise proxy for the number of children using a service, which services are generally less able to measure accurately and robustly". Reliance on a numerical perspective is problematic. Using either profitability or the size of the user base to define risk of harm excludes from mitigating action the types of harm that minority or intersectional groups might experience from smaller sites that are designed to target them and overlooks the potential severity of that harm to individuals. "But the Act is equally clear that we must take account of the size and capabilities of the wide range of services in scope of the protection of children duties. These vary enormously and therefore we have not taken a one-size-fits-all approach. Measures

	that are appropriate and proportionate for the biggest and riskiest services may not be achievable for smaller and less risky firms, and when applied broadly they could lead smaller services to withdraw from the UK or reduce investment. Where this hampers competition and innovation, this can reduce the benefits of online life for all users, including children. For this reason, we have proposed different measures according to the level of risk posed by services, their size and resources. We propose that all services accessed by children – regardless of their size or risk – implement a core set of measures to protect children online. We propose additional measures for services that pose a greater risk of harm to children, recommending costly measures for smaller services only where there is clear risk of harm and where we have evidence that the measures proposed will make a material difference in dealing with this risk. Larger and better-resourced services that pose the most material risks to many children will be expected to go even further (3.18 & 3.19) Elsewhere, Ofcom's justification for a differential obligation between small and large companies seems based on what they do already (e.g. large companies do more already) and the impact of the harmful consequence. This is a quantitative assessment of harm - how many people are harmed, not how badly they are hurt, and therefore is not well framed to assess the impact of small, single issue services. (We note above how the severity of harm is not taken into consideration in the proportionality assessment.) Placing low governance obligations on smaller companies does not make sense when many of these obligations are affecting basic principles for company or service operation (e.g. guidance on how to apply community guidelines, or on training moderators). The response from smaller companies may be

simpler, to take account of the size and lack of complexity of their operation, but the basic principles still remain.

The only measures in the children's codes of practice that apply to all U2U services (annex 7) or all search services (annex 8), regardless of risk or size, are the same as those that applied to all services in the illegal harms codes (both references given below)

Children's code Ref	Illegal harms ref	Measure
User-to-user code		
GA2	3B	Named person accountable to the most senior governance body.
CM1	4A	Content moderation systems or processes designed "swiftly take action" against content harmful to children
UR1-UR4	5A-H)	Measures relating to reporting and complaints
TS1 & TS2	6A&B	Terms of service measures
	NA	The age assurance measures apply to "all user-to-user services" based on whether they host or do not prohibit Primary Priority Content or Priority Content
Search code		

GA2	3B	Named person accountable to most senior governance body
SM1	4A	Systems and processes designed to take appropriate action" on PPC, PC or NDC
UR1-3 & 5	5 A-H	7 of the 9 measures relating to reporting and complaints
TS1 & TS2	6A&B	Publicly available statements

We refer Ofcom back to our previous submission for our analysis of how the differentiation of size and risk plays out in relation to the measures.

Evidence

What is marked in this consultation compared to the previous one, is that Ofcom provides its own commentary on the evidence of the risks posed by small and niche sites - though it does not work this through to specific measures and/or the extension of other measures intended only for larger sites.

For example:

> "Smaller services can pose a particular risk of harm because they may be more focused on niche interests or topics and can therefore present a higher risk of encountering harmful content, if these topics are likely to contain content harmful to children. Smaller services may also have fewer resources available to moderate content, and therefore present a higher risk of hosting harmful content. For example, evidence suggests that content promoting suicide and self-harm can be shared within online communities, some of

	which exist on smaller, more niche services. Refer to Section 7.2 and 7.3 on Suicide and self harm content and Eating disorder content for more detail." (7.14.13) "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. The research suggests that some communities, and even entire services, are 'deeply hateful'; that the Terms of Use for these services are 'more lax' than mainstream services, and do not explicitly prohibit hate speech. Comparison of hate content within these services, and more mainstream ones, found that while even in the more extreme parts of the internet not all posts are hateful, the level of hate is significantly higher than in mainstream services." (7.4.31) "Although there is a lack of evidence on children's use of these smaller niche services, there is a risk that children might encounter hate content on large social media services, and then be led to smaller, niche services with higher volumes of hate content and therefore higher risk of harm. Our Illegal Harms Register (Section 6F.32) notes that 'perpetrators of hate offences' tend to use services with large and small user bases in different ways. Research has found that some potential perpetrators are incentivised to maintain a presence on larger mainstream social media services, where they build their network further with new users, attracting them with 'borderline' hate content (such as by sharing incendiary news stories and provocative memes). These networks of users are then directed towards less-moderated services. In these spaces, users discuss and share hate content more openly." (7.4.32, also 7.4.26 and 7.4.27)

	As we flagged in our illegal harms consultation, there is increasing evidence of the direct offline harm caused by dedicated, single-risk sites. For example: • 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. In relation to the concern about small suicide sites and message forums that sit behind URLs, the ICO has had to cope with some of this in the UK with cold calling companies going into insolvency the moment the ICO goes after them with regulatory measures (in the ICO's case mainly fines) but then the person behind the company pops up again with another company and carries on doing the same thing. You could have a forum that then changes its name slightly but has the same people behind it. Who is the provider (see s 226(3) on this) and more specifically can Ofcom keep a track of them? The enforcement plan does not seem to consider this issue (and that of 'refusenik' sites) in general. We have recently published a blog post on this issue specifically. The differential requirements relating to even core expectations such as content moderation is surprising given how central this function is to the duties in the Act – and how its under-resourcing in even the largest platforms has been evidenced to cause

	harm. We refer Ofcom here to the evidence we previously provided from US court filings and from Revealing Reality. We also refer to the extracts from the X/Twitter Australian transparency elsewhere in this response.. The way Ofcom applies its risk assessment approach focuses on size and number of risks but not on the severity of risks, which allows the small, niche sites to slip through the net. The risk assessment process, as we have described above, is too focused on corporate risks and managing external reputational issues, with governance requirements related to the type of information they should be assessing, in what form. There is no requirement to look at testing or risk assessment of the actual impact of the products or services that they are responsible for. Furthermore, many of the governance requirements are only applied to larger platforms. **Recommendation** We recommend that Ofcom review its definition of proportionality to ensure that all services, regardless of size, are required to take measures that will address the risks they have identified in their risk assessment if they correspond to one or more of the risks set out in the risk register. We also recommend that Ofcom remove the differentiation based on size that it has applied to the specific measures recommended in the codes of practice and require services instead to decide on – and justify to Ofcom – whether their adoption of these measures is proportionate to the risks posed by their services. We refer back to the recommendation we propose for addition to the draft codes as we recommend that this applies to all services regardless of size. **Safety by Design** With specific reference to measures that could be seen as touching on "safety by design" (including written statements of responsibilities or expectations of product testing), Ofcom makes an upfront

	judgement that these can only be reasonably expected of large or multi-risk companies – thereby undercutting at the outset the overarching legislative objective in the Act. Significantly, in the proposals set out on governance in volume 4, Ofcom - in a proposal that it acknowledges "mirrors an equivalent one in the illegal harms consultation" (para 11.89) - sets out that a written statement of responsibilities for senior members of staff would: > "include ownership of decision-making and business activities that are likely to have a material impact on children's online 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 or behavioural biases have been taken into account), how data related to children's online safety is collected and processed, and how humans and machines implement trust and safety policies. Depending on a service's structure, key responsibilities in children's 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." (Vol 4, 11.87) However, as with the illegal harms consultation, this statement of responsibilities is only recommended for large or multi-risk services despite the acknowledgment that "decision-making and business activities are likely to have a material impact on user safety outcomes", which goes to the heart of safety by design. Indeed, as we set out below, the Government's Impact Assessment makes reference to the fact that building in safety by design is a way for smaller platforms to reduce regulatory compliance costs. Ofcom

	itself has recognised that smaller providers are likely to have less complex systems which would suggest safety by design would be - in process terms - less complex than for larger operators. Ofcom also only makes a few brief references to product safety testing, which we would include as a component of an overall "safety by design" approach. In Volume 3, Ofcom says: "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). Then, in a table suggesting a number of "enhanced inputs" to help companies build up their "risk assessment evidence base", "results of product testing" are included: > "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 (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." (Table 9.5) (Our emphasis) This is an "enhanced input": an expectation for larger services only. Ofcom's rationale for this distinction between "core" and "enhanced" inputs is: "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) This therefore means that not only is product testing to ensure user safety not expected of smaller companies, it is not something that Ofcom feels should be carried out as part of a risk assessment to inform the measures that smaller services might feel they need to take in order to make their products safe. Implicitly in this, Ofcom is seeing severity of harm as being about the number of people affected, not the severity of harm caused, an approach which is not necessarily mandated by the Act but which occurs repeatedly throughout the consultation. This seems to run counter to a "safety by design" approach. It is in marked contrast to the approach of the CMA and the ICO who suggest in a joint paper that testing is key to prevent harmful design in choice architecture; the paper notes that there are different ways of testing. The resources available to a service provider could thus inform the sort of testing rather than the question of whether service providers should test.
Age assurance measures (Section 15)	
31. Do you agree with our proposal to recommend the use of highly effective age assurance to support Measures AA1-6? Please provide any information or evidence to support your views. a) Are there any cases in which HEAA may not be appropriate and proportionate?	Confidential? –N We note that in the children's Summary document (p13 onwards), Ofcom sets out the "safer platform design choices" that it is consulting on: "We are also proposing a range of safety measures that focus on service providers ensuring they make foundational design choices, so children have safer online experiences.

b) In this case, are there alternative approaches to age assurance which would be better suited? 32. Do you agree with the scope of the services captured by AA1-6? 33. Do you have any information or evidence on different ways that services could use highly effective age assurance to meet the outcome that children are prevented from encountering identified PPC, or protected from encountering identified PC under Measures AA3 and AA4, respectively? 34. Do you have any comments on our assessment of the implications of the proposed Measures AA1-6 on children, adults or services? a) Please provide any supporting information or evidence in support of your views. 35. Do you have any information or evidence on other ways that services could consider different age groups when using age assurance to protect children in age groups judged to be at risk of harm from encountering PC?	These cover three broad topics: • understanding which users are children so that those children can be kept safe; • ensuring recommender systems do not operate to harm children; and • making sure content moderation systems operate effectively. With the exception of the proposals around the recommender systems (which is welcome), these topics - and the measures related to them which we discuss below - do not go much further than the ex-post measures Ofcom set out in the illegal harms consultation. In fact, two-thirds of the 36 measures recommended for U2U platforms, and all but one of the 24 measures for search services, are the same or equivalent versions. Age assurance - e.g. keeping children off platforms - is a tool to prevent harm but not a "safety by design" choice that fundamentally changes the platform itself for all users, whether they are children or not. We refer Ofcom here to the analysis by 5 Rights/Children's Coalition of the age assurance proposals. Content moderation is about dealing with content that is already posted rather than addressing the system which it flows over. In the Proposed codes at a glance, the description of measures highlights how they are limited to cutting off access to the service to children (by age assurance) for PPC content and some PPC, then to cut off access at more granular content level using age assurance, then to use age verification to assess recommender system usage, plus content moderation. This is not safety-by-design but the application of safety tech on top of a system that is deemed to be harmful to the users that the regulatory framework is designed to protect (and at a higher level than adult users, too). On the age assurance proposals specifically, we note that Ofcom's proposals here are the same as

	those set out in their consultation on the part 5 duties for pornography providers. This is good in terms of consistency of approach and in ease of regulatory enforcement. As such, the analysis we provided to Ofcom's consultation on those duties applies and we provide the relevant sections in full below in the evidence section. We make here a few observations of some of the - perhaps unintended - consequences of Ofcom's decision to place so much weight by the age assurance measures to provide protection of children and not (as we have argued above) to ensure that all the other aspects of regulatory compliance are as robust as possible. • There is no requirement to do this for illegal content, just for content that is designated as Primary Priority Content (PPC) or Priority Content (PC) or non-designated content (NDC). This means that sites that might be primarily set up for disseminating illegal content don't need to keep children off (though it is arguable whether they would comply with any of the regulatory requirements anyway) unless illegal content is seen as also falling within the categories of content harmful to children. However, this does beg the question as to whether it would be better for small, high-harm platforms to be subject to age-gating rather than for Ofcom to be attempting to manage the content via risk registers and related measures. • Ofcom has not attempted to introduce measures that would take into consideration the different age groups of children who might be on platforms and how harm manifests itself according to age, although some of this is described in the risk register. Ofcom says that this is difficult, though it would seem that the bigger platforms are already very well aware of the ages of children on their platforms to a fairly precise degree of accuracy. See Arturo Bejar from 36 mins

	here where he mentions "talking to regulators in the UK" and being aware that: "*Social media companies .. particularly Meta .. misrepresent what they are able to do. For example, they talked about their inability to detect under-13 accounts … It's not that hard to find an account that an 8 year old makes. These are all problems that are solvable*." If platforms know the age of their users, it should be possible for them to introduce different measures for those different users. It appears here - as Bejar suggests - that Ofcom is taking at face value platforms describing what they are doing now, without looking at what the capacity of age-verification might be - if properly applied, as required under the Act. • There is a flaw too in using age gating as the means to prevent harm in otherwise anodyne or relatively risk-free environments. If, for example, the service is a small gaming platform that might have instances of severe harm but not in large quantity or on a large scale, then its requirements under the age assurance duties will mean that those instances of severe harm will not get addressed. Eg Volume 4, 12.50: "However, for the avoidance of doubt, we expect that any service with more than 1 million (or between 100,000 and 1 million) monthly UK child users would need a range of robust evidence to demonstrate that it does not in fact pose high (or medium) risk of harm to children in respect of a given kind of content." • Related to this, an obligation/dependency on age verification potentially means that the quality of the service providers' risk assessments are secondary - eg if children aren't on the platform, then they don't need to keep monitoring risks. • There is also the question as to what happens if the percentage of content that is "principal purpose" is just below the threshold designated for age assurance measures to prevent children's access.

	Evidence We include here the main points we made with regard to Ofcom's similar approach in the part 5 guidance for pornography service providers. We also refer to the submissions from children's charities, particularly 5 Rights and NSPCC on this topic. With regard to the principles-based approach, we noted that Ofcom does not provide sufficient criteria by which it will measure those outcomes and/or the providers' compliance with their duties. Ofcom put forward arguments about the "nascent" age verification industry (see above, though we also note age verification in some form or other has been required under the Communications Act for more than a decade) which they said justify not having an output level score (especially in relation to technical accuracy). There is a difference between recommending a particular tool (which Ofcom in our opinion rightly is not doing, both in the part 5 guidance and these proposals) and measuring effectiveness of any tool. If the concern is that any one tool could not be effective enough, techniques could be used in combination with other tools. Ofcom's narrow approach means that it is precluding the potential effectiveness of combinations of techniques that might lead to the same outcome. We note that Ofcom provides criteria describing different aspects of effectiveness. While we agree with these aspects, they do not in themselves provide a definition for highly effective. While we appreciate that there may be challenges in specifying a metric by which to judge "highly effective" age assurance technologies, there would be no reason why Ofcom could not specify a metric for each of their criteria that would indicate that the method adopted – and/or the implementation and enforcement of that method – by the regulated provider is "highly effective". If, in practice, the application of that age assurance method falls below the metric specified, the written record could then be used by Ofcom to determine whether providers had used their best efforts and/or acted in good faith to ensure its effective implementation and identify those providers

	who had done neither. Ofcom however say that they are not doing "setting a base level for score" so because of the "nascent" age assurance industry and because they want to "allow space for important innovation in the safety tech sector". In our view, metrics related to Ofcom's criteria (rather than types of technology) would not preclude innovation in this field. **Recommendation** We would suggest that Ofcom looks again at the definition of "highly effective" and also, in light of Arturo Bejar's comments, uses their information-gathering powers as a priority to understand what is already technically feasible for the companies with regard to age assurance and updates the measures in their next iteration of the codes accordingly
Content moderation U2U (Section 16)	
36. Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views. 37. Do you agree with the proposed addition of Measure 4G to the Illegal Content Codes? a) Please provide any arguments and supporting evidence.	Confidential? – N Response: We cover this in relation to our general responses to the codes of practice (and the measures recommended) and differential approach to large and small sites, based on cost and size. We also refer to the analysis at annex A. We would also make this point from the perspective of VAWG-related harms: It is a significant concern that there are no measures requiring services use some form of automated content moderation, particularly for large or multi-risk services. Whilst the Codes set out what companies must do in response to harmful content, they are much less clear about how this content should be identified in the first place. There is a significant risk that this will enable services, particularly those who are looking to take a 'hands-off' approach to moderation, to avoid putting proactive systems in place.

	Human moderation alone will not be able to effectively assess whether content is PPC or PC at the scale and speed required. This means that there is a real risk that misogynistic material, as well as other harmful content which disproportionately impacts girls, will not be meaningfully identified and removed / hidden / downranked.
Search moderation (Section 17)	
38. Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views. 39. Are there additional steps that services take to protect children from the harms set out in the Act? a) If so, how effective are they? 40. Regarding Measure SM2, do you agree that it is proportionate to preclude users believed to be a child from turning the safe search settings off? The use of Generative AI (GenAI), see Introduction to Volume 5, to facilitate search is an emerging development, which may include where search services have integrated GenAI into their functionalities, as well as where standalone GenAI services perform search functions. There is currently limited evidence on how the use of GenAI in search services may affect the implementation of the safety measures as set out in this code. We welcome further evidence from stakeholders on the following questions and please provider arguments and evidence to support your views: 41. Do you consider that it is technically feasible to apply the proposed code measures in respect of GenAI	Confidential? – N Response: We cover this in relation to our general responses to the codes of practice (and the measures recommended). We also refer to the analysis at annex A.

functionalities which are likely to perform or be integrated into search functions? 42. What additional search moderation measures might be applicable where GenAI performs or is integrated into search functions?	

User reporting and complaints (Section 18)	
43. Do you agree with the proposed user reporting measures to be included in the draft Children's Safety Codes? a) Please confirm which proposed measure your views relate to and explain your views and provide any arguments and supporting evidence. b) If you responded to our Illegal Harms Consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response. 44. Do you agree with our proposals to apply each of Measures UR2 (e) and UR3 (b) to all services likely to be accessed by children for all types of complaints? a) Please confirm which proposed measure your views relate to and explain your views and provide any arguments and supporting evidence. b) If you responded to our Illegal Harms Consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response. 45. Do you agree with the inclusion of the proposed changes to Measures UR2 and UR3 in the Illegal Content Codes (Measures 5B and 5C)?	Confidential? – N Response: We cover this in relation to our general responses to the codes of practice (and the measures recommended) and the differential approach to large and small sites, based on cost and size. We also refer to the analysis at annex A. We would also make this point from the perspective of VAWG-related harms: The proposals on user reporting and complaints put much burden on children to provide the evidence for platforms to take action on harmful content. We note that Ofcom is seeking additional evidence in relation to user reporting: we would urge them in this regard to include a measure or recommendation in the codes of practice to use Trusted Flaggers. Trusted Flaggers with expertise in this online VAWG could strengthen reporting systems and ensure the onus is not on children to report harm. We also note that much of the burden is passed to children in terms of managing their own safety. Ofcom notes the evidence that "Children in particular are often dissuaded from reporting content or complaining, as they do not think anything will come of their complaint. Our research into children's attitudes to reporting echoes this finding, and suggests that if children receive no update on the outcome of their complaints, they do not believe they have been taken seriously." (7.11.43)

a) Please provide any arguments and supporting evidence.	There is lots of evidence further cited on this issue, including how delays in removing reported accounts can exacerbate harms to children. Later. at 7.11.53, Ofcom notes: “Some children use the available tools to protect themselves online, such as blocking content or blocking accounts, although use remains low, possibly due to the reasons set out in the ‘User reporting and complaints’ subsection.” While measures relating to simplifying reporting and complaints are welcome - particularly given the evidence as to the inadequacy of the processes currently used - there is no requirement on, or means by which to incentivise, services’ improvements in this area nor are any metrics required to be collected on the types and volumes of reports. Moreover, in relation to networks of accounts that are generating the most complaints from children, there is no obligation on companies to track this and take action (such as disrupting or blocking them) in response to the levels of complaints received from children. Ofcom would not have had to come up with a specific measure but instead put an obligation on companies to devise appropriate metrics that were context- and business-specific, use the information this provided as part of the suite of inputs to their risk assessment and devise a mitigation measure accordingly. Transparency reporting and researcher access to data are other complementary routes to this and should be considered by Ofcom in building its evidence base.

Terms of service and publicly available statements (Section 19)	
46. Do you agree with the proposed Terms of Service / Publicly Available Statements measures to be included in the Children's Safety Codes? a) Please confirm which proposed measures your views relate to and provide any arguments and supporting evidence. b) If you responded to our illegal harms consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response. 47. Can you identify any further characteristics that may improve the clarity and accessibility of terms and statements for children? 48. Do you agree with the proposed addition of Measure 6AA to the Illegal Content Codes? a) Please provide any arguments and supporting evidence.	Confidential? N Response: We cover this in relation to our general responses to the codes of practice (and the measures recommended) and the differential approach to large and small sites, based on cost and size. We also refer to the analysis at annex A.
Recommender systems (Section 20)	
49. Do you agree with the proposed recommender systems measures to be included in the Children's Safety Codes? a) Please confirm which proposed measure your views relate to and provide any arguments and supporting evidence. b) If you responded to our illegal harms consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response.	Confidential? – N Response: We cover this in relation to our general responses to the codes of practice (and the measures recommended) and the differential approach to large and small sites, based on cost and size. We also refer to the analysis at annex A. From a safety by design perspective, we also note that the measures relating to the recommender system - while welcome and integral to a platform or service's design - still relate largely to the content that flows over the system and that is promoted by

50. Are there any intervention points in the design of recommender systems that we have not considered here that could effectively prevent children from being recommended primary priority content and protect children from encountering priority and non-designated content? 51. Is there any evidence that suggests recommender systems are a risk factor associated with bullying? If so, please provide this in response to Measures RS2 and RS3 proposed in this chapter. 52. We plan to include in our RS2 and RS3, that services limit the prominence of content that we are proposing to be classified as non-designated content (NDC), namely depressive content and body image content. This is subject to our consultation on the classification of these content categories as NDC. Do you agree with this proposal? Please provide the underlying arguments and evidence of the relevance of this content to Measures RS2 and RS3. • Please provide the underlying arguments and evidence of the relevance of this content to Measures RS2 and RS3.	its algorithm rather than the deployment of a recommender system itself. The recommender system may not be a problem, per se: it's how it's designed, the values it incorporates and the way it is used by the service provider. The consultation also does not consider how recommender systems form part of the suite of incentives for content creation (see also our commentary on business models, below) and how being picked up by the algorithm is important for advertising revenue and other promotions. Moreover, it is relatively far down the design stack in terms of its impact. We have concerns here that this narrow approach will ultimately be a missed opportunity, resulting in piecemeal impacts on children with little shift in the culture of safety within companies and the overall safety of products used by children, particularly those in vulnerable groups with shared characteristics. In the introductory sections to volume 3 (risk register), Ofcom's description of recommender systems highlights the problems: "The functionalities and characteristics we describe as risky are not inherently harmful and can have important benefits. For example, recommender systems benefit internet users by helping them find content which is interesting and relevant to them. The role of the new online safety regime is not to restrict or prohibit the use of such functionalities or characteristics, *but rather to get services to put in place safeguards which allow users to enjoy the benefits they bring, while managing the risks appropriately*." (our emphasis) (vol 3, page 4) It is not clear what "safeguards" mean here. Is this post-hoc, after content has been created? If so, this is not "safety by design" - it implies that the recommender system will run as previously but overlaid with interventions to meet the measures required in the codes. In that regard, Ofcom's approach does not fit with what is in the Act or in the risk register.

	In the next section, we also look at how the business model affects the creation and promotion of harmful content - intersecting with the recommender system in a way that is about system design choices as much as the motivation of the individual content creators. Ofcom describes this interplay in para 7.12.5: "The choice architecture of a service (i.e. the design of the choice environment in which a user is making decisions) can be *designed to influence or manipulate users into acting in ways that serve commercial interests but may be detrimental to individual or societal interests (e.g. spending time engaging with the service, in the case of advertising revenue models*)" (our emphasis)
User support (Section 21)	
53. Do you agree with the proposed user support measures to be included in the Children's Safety Codes? a) Please confirm which proposed measure your views relate to and provide any arguments and supporting evidence. b) If you responded to our Illegal harms consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response.	Confidential? N Response: We cover this in relation to our general responses to the codes of practice (and the measures recommended) and the differential approach to large and small sites, based on cost and size. We also refer to the analysis at annex A.
Search features, functionalities and user support (Section 22)	
54. Do you agree with our proposals? Please provide underlying arguments and evidence to support your views. 55. Do you have additional evidence relating to children's use of search services and the impact of search functionalities on children's behaviour? 56. Are there additional steps that you take to protect children from harms as set out in the Act? a) If so, how effective are they?	Confidential? – N Response: We cover this in relation to our general responses to the codes of practice (and the measures recommended) and the differential approach to large and small sites, based on cost and size. We also refer to the analysis at annex A.

As referenced in the Overview of Codes, Section 13 and Section 17, the use of GenAI to facilitate search is an emerging development and there is currently limited evidence on how the use of GenAI in search services may affect the implementation of the safety measures as set out in this section. We welcome further evidence from stakeholders on the following questions and please provide arguments and evidence to support your views: 57. Do you consider that it is technically feasible to apply the proposed codes measures in respect of GenAI functionalities which are likely to perform or be integrated into search functions? Please provide arguments and evidence to support your views.	

Combined Impact Assessment (Section 23)	
58. Do you agree that our package of proposed measures is proportionate, taking into account the impact on children's safety online as well as the implications on different kinds of services?	Confidential? – Y / N As with the illegal harms consultation - and unsurprising given that the children's proposals so closely mirror them - 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. Severity of harm is not just about how many people are affected either; it concerns the intensity of impact too. Yet, despite the express recognition of the harms for the risk register, when discussing the measures for the code neither aspect is expressly considered. 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 is also an issue, particularly given that it downplays the severe harm that can occur to minoritised groups on targeted, small sites. **What the Act says** There are 53 references to "proportionate" within the Act. While the Act defines proportionality (in relation to safety duties), Ofcom has not expressly stated how it is approaching the required balancing act; this may be in part because of the structure of the document whereby an analysis of harms sits in the risk register volume. It would be helpful for these issues to have been pulled through so it is

clear how Ofcom is weighting the harm and balancing it against costs.

It is our assessment that the Act, as drafted, does not direct Ofcom to take costs into account as the main driver of whether measures are proportionate or not but to make a judgement as to whether the recommendation of the measures itself is proportionate based on the kind or size of a service and the likely level of risk that those services pose, according to the functionalities that are identified in the risk assessment and also to weigh that against the severity of the harms also identified in the risk assessment (including the recognition that some of those harms might constitute an interference with individuals' human rights).

Parliamentary debate

In the Lords Committee stage debate on 2 May, Lord Parkinson – the Government Minister – gave the following reassurances in relation to the child safety duties:

> "The provisions in the Bill on proportionality are important to ensure that the requirements in the child-safety duties are tailored to the size and capacity of providers. It is also essential that measures in codes of practice are technically feasible. This will ensure that the regulatory framework as a whole is workable for service providers and enforceable by Ofcom. I reassure your Lordships that the smaller providers or providers with less capacity are still required to meet the child safety duties where their services pose a risk to children. They will need to put in place sufficiently stringent systems and processes that reflect the level of risk on their services, and will need to make sure that these systems and processes achieve the required outcomes of the child safety duty. …
>
> The passage of the Bill should be taken as a clear message to providers that they need to begin preparing for regulation now—indeed,

	many are. Responsible providers should already be factoring in regulatory compliance as part of their business costs. Ofcom will continue to work with providers to ensure that the transition to the new regulatory framework will be as smooth as possible." (Hansard 2 May col 1485) **Ofcom's proposals** We have set out a lot of material in section 7, below, in relation to the judgements on "proportionality" that lead to differential obligations being placed on small and large services and do not propose to repeat them here. The following extracts are relevant here to demonstrate where costs are used as a means by which to judge proportionality though, on the basis of our reading of the two consultations, this seems to be less marked in the children's consultation than in the illegal harms consultation. That said, given that the bulk of the recommended measures and their application based on size of company is rolled over from the illegal harms consultation, we have to assume the same economic criteria applies to those equivalent measures without any modification, even if it is not explicitly described as such in this second consultation. For example, "Impacts on services are an important consideration to ensure that more costly requirements are justified, even where they could negatively affect users. For example, if a high-cost burden on services reduces investment in areas other than user safety or (in the most extreme cases) drives some services to stop operating in the UK, this means that both children and adults can no longer benefit from such services or new innovations. This does mean that services should not fulfil their duties to keep children safe because it is costly. Considering the cost impact on services aims to meet the child safety requirements under the Act without unduly undermining investment in high-

	quality online services that UK users can enjoy, including children." "At this stage we do not consider it proportionate to recommend this measure for services that are not multi-risk for content harmful to children. For the same reasons set out above, we expect that benefits would be limited for these services. While there are potentially some benefits for single-risk services and the costs of this measure in isolation could be manageable for some of them, we have considered the combined implications of this measure on top of others. As set out in our combined impact assessment Section 23, we consider that the overall cost burden on some single-risk services may negatively affect users and people in the UK, so we have prioritised other measures for them where the benefits are more material." We made a point in our illegal harms consultation, in relation to child sexual abuse, that the severity of the offence and the costs to society (quantified at c£2.bn in the "underestimate" provided in the Government's Impact Assessment) are significant. Yet Ofcom's consideration of the merits of CSAM measures were weighed up against the costs to business – without considering the extent of the harms to the individuals nor the costs to society to eradicate this sort of crime and to provide support to affected individuals: "The level of detail and complexity in the comparison of costs and benefits is greater for some measures than others. This sometimes reflects the availability of information. It can also reflect where a more detailed assessment is more likely to impact our recommendations, and when it can affect which services we recommend measures for. This is especially the case for some of the measures we recommend to reduce grooming and the hash matching measure we recommend to reduce CSAM, where we carefully consider whether to recommend the

	measures for smaller services". (Illegal Harms: Vol 4, 11.32) There is a further aspect of this in the children's consultation - the severity of harm does not feature in the approach to proportionality nor in the designation of measures for services. For example, "Services likely to be accessed by children are required by the Act to use proportionate safety measures to keep them safe. Our draft Children's Safety Codes provide a set of safety measures that online services can take to help them meet their duties under the Act. Services can decide to comply with their duties by taking different measures to those in the Codes. However, they will need to be able to demonstrate that they offer the appropriate level of safety for children." **Evidence** We refer Ofcom to the evidence we presented in our illegal harms consultation response, including; • The Government's 2022 Impact Assessment (IA) • The case of X/Twitter in Australia **Recommendation** Based on the Parliamentary debates, Government statements and the Government's own impact assessment, we would argue that Ofcom's interpretation of what is "proportionate" is not appropriate. We would refer back to the recommendation we make for additional measures relating to product safety testing and safety by design to be added to the draft codes, which would place the responsibility on services (of all sizes) to take measures that are proportionate to them to address the risk of harm that is identified in their risk assessment.
Statutory tests (Section 24)	

59. Do you agree that our proposals, in particular our proposed recommendations for the draft Children's Safety Codes, are appropriate in the light of the matters to which we must have regard? a) If not, please explain why.	Confidential? – Y / N
Annexes **Impact Assessments (Annex A14)**	
60. In relation to our equality impact assessment, do you agree that some of our proposals would have a positive impact on certain groups? 61. In relation to our Welsh language assessment, do you agree that our proposals are likely to have positive, or more positive impacts on opportunities to use Welsh and treating Welsh no less favourably than English? a) If you disagree, please explain why, including how you consider these proposals could be revised to have positive effects or more positive effects, or no adverse effects or fewer adverse effects on opportunities to use Welsh and treating Welsh no less favourably than English.	Confidential? – Y / N

Please complete this form in full and return to protectingchildren@ofcom.org.uk.