# Protection of Children Consultation: Response from Online Safety Act Network

Professor Lorna Woods OBE, University of Essex & Maeve Walsh, Director OSA Network
Contact at **hello@onlinesafety.net**

## Introduction

### About the OSA Network

The Online Safety Act Network brings together over 60 civil society organisations, campaigners and advocates with an interest in the implementation of the OSA. More details about our work are here.

### Structure of our Submission

This submission is divided into a summary analysis section (also available here) and 10 supporting sections, each of them covering a specific issue arising from Ofcom's consultation. Many of these supporting sections mirror analysis from our submission to Ofcom's illegal harms consultation; we have tried where possible not to duplicate material and instead provide cross-references, where required, to that submission and its supporting evidence.

Each section is structured in broadly the same way - mirroring our previous response - which we hope provides a consistent approach and will enable Ofcom to make best use of our analysis in its entirety as well as within the individual teams leading on different parts of the consultation:

- Issue
- What the Act says
- Parliamentary debate
- Ofcom's proposals
- Evidence
- Recommendation

This response should be read in conjunction with the table we provide at annex A (analysis of volume 3 functionality risks vs volume 5 mitigation measures).

Organisations within our network will be submitting their own individual responses to this consultation. We do not repeat the expert analysis and evidence on their particular areas of interest in our submission but would see much of this as supporting evidence for the broad, structural themes we have focused on here. To that end, we have supported the analysis put forward by the Children's Coalition.

**July 2024**

**Contact: Maeve Walsh**

**maeve@onlinesafetyact.net**

## Summary

1. Ofcom's protection of children consultation is the second major plank of its implementation of the regulatory regime that it will be enforcing under the Online Safety Act 2023. The first - the illegal harms consultation - closed in February 2024 and Ofcom's response has not yet been published.

2. Ofcom refers to its attempts to provide alignment and consistency between the two consultations at a number of points in the documentation. For example, on age assurance they have "aimed to ensure consistency with our Illegal Harms Consultation and Part 5 Guidance" (Summary; p7); they have "sought to align our draft Children's Risk Assessment Guidance with our draft Illegal Harms Risk Assessment Guidance where possible" (Summary; p10); and "our approach [to governance] is consistent with our Illegal Harms Consultation. This means service providers who must comply with both illegal content safety duties and children's safety duties can choose to adopt a single process that covers both areas" (Summary; p12). Many of the measures proposed in the children's codes mirror those in the illegal harms codes. (Proposed codes at a Glance)

3. We raised a number of concerns about the approach taken by Ofcom in its illegal harms proposals, not least as we felt that the strategic choices they had taken risked setting the regime off on a weak footing that would not be easily revised in subsequent iterations of the codes of practice. Our full response to the illegal harms consultation is here and a public statement, co-signed with a number of the organisations in our network, is here. Those concerns remain - not least, as the mirroring of the approaches and broadly similar measures from the illegal harms consultation bakes the same weaknesses into this one.

4. Ofcom - in volume 1 - sets out that the feedback which it received on the illegal harms consultation may result in a changed approach to some elements of the illegal harms proposals - and consequently the children's proposals which mirror them. This is necessary if they are to maintain the consistency between the two parts of the regime:

   > "To ensure a coherent online safety regime and to help services understand their responsibilities, this consultation follows, as far as possible, a consistent approach with the Illegal Harms Consultation and Part 5 Consultation. We are currently carefully considering and analysing the responses received to these consultations.
   >
   > Some of the feedback we have received on our previous proposals may also be relevant to the approach currently proposed in this consultation. Where that is the case, we will take into account the feedback on our regulatory approach in the round to ensure that

> our approach remains consistent across our consultations. For example, several respondents to the Illegal Harms Consultation expressed concern that under the Act services which follow our Codes of Practice will be deemed compliant with the relevant safety duties even if there are risks in their risk assessment which are not fully addressed by Ofcom's proposed measures. We are considering this issue carefully and will provide a detailed response covering both the Illegal Harms and Protection of Children proposals following this consultation." (Volume 1; p20)

5. This therefore makes responses to this consultation both more straightforward and more challenging at the same time. Straightforward in the sense that much of our analysis and feedback is the same; we provide cross-references to our previous submission and supporting evidence where appropriate but, in many cases, the substantive commentary and analysis is restated here. It is more challenging, however, in that we do not know how extensive Ofcom's revisions will be as a result of the illegal harms consultation nor whether they will be (relatively speaking) superficial (eg, additional measures added to the codes of practice, for instance) or fundamental and transformative to the regime as a whole (eg a more comprehensive approach to safety by design, or a different approach to governance and risk assessment).

6. We have chosen therefore to emphasise, where applicable, the same points we made in response to the earlier consultation, linking them to material from the current consultation to show that using the same (consistent) approach will lead to - in our view - similar (limited) regulatory outcomes. We also question whether this truly does deliver the "strongest protections for children" promised by the Government and enshrined in the Act at section 1 3 (b) (i). We hope that Ofcom will therefore address our feedback in the round when it responds to both consultations later in the year. Our work and advocacy through the legislative process, and now during the implementation phase, has only ever been with the intention of ensuring robust, outcomes-focused regulatory interventions that make the UK the safest place to be online.

## Our proposed recommendation

7. In our previous response, we made a recommendation for an amendment to the illegal harms codes of practice that - we felt - would resolve a number of the structural issues within Ofcom's approach, including the shortcomings of the evidential threshold it had set itself before measures could be included in the codes, its approach to proportionality, the lack of a true focus on safety by design biting at the level of systems and the limitations of its risk assessment guidance. We do not know whether this suggestion has been taken on board by

Ofcom nor whether a measure like this will appear in subsequent iterations of the codes. But we remain of the view that it is the most efficient and effective way to resolve the similar issues we have identified in this consultation and to ensure that there is a step change in the safety of users on all regulated services as soon as practicably possible. This approach is also, in our view, very much aligned with the intentions behind the Government's policy goals and parliamentary support for, and amendments to, the Bill during its passage.

8. The amendment is provided upfront here as context for the material that follows.

We suggest the following wording is inserted in the draft codes for both illegal harms and protecting children, between the section on governance and accountability and the section on content moderation, which 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).

"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 measures taken in reducing the risk of harm identified in the risk assessment. These measures should feed back into the risk assessment."

9. 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 pre-determined 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.

10. 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.

11. Before we set out the detail of this response, we felt it was important to acknowledge a few particular things that set the children's consultation apart from the illegal harms consultation that preceded it.

### Some positives

12. There is a greater sense of **consistency and coherence** between the constituent parts of this consultation. The illegal harms consultation felt like it had been rushed in some places - understandable given how quickly it was published, following Royal Assent for

the Online Safety Act - leading to gaps, differences in tone and approach and internal inconsistencies between different parts of the documentation. The children's consultation - while still overly long and repetitive in places - is more coherent and, as a result, easier to navigate.

13. Ofcom has worked hard **to respond to feedback from civil society** on its handling of the first consultation. The summary document is a welcome "way-in" for small organisations looking to engage with the detail and there is more (but not a huge amount of) acknowledgement of the evidence from civil society organisations that acts as a counterbalance to the evidence from industry and tech platforms. That said, the consultation is still very long (1300+ pages vs 1900+) and the terms in which feedback is requested are fixed by the questions that Ofcom chooses to ask relating to the specific proposals, rather than open in the sense of seeking views as to the overall framework (within which those specific proposals sit) and its potential effectiveness.

14. As mentioned above, many of the issues that were raised in the illegal harms consultation have been acknowledged - though they have not been worked through to the new proposals.

15. There is evidence in some parts of the consultation (notably the children-specific aspects) of **a shift away from prescriptive "tick-box" approaches** to compliance to one where the responsibility is put on service providers to exercise a duty of care to the children who are using their platforms. For example, in the child access assessment volume, there is emphasis on companies deciding what they have to do. E.g.

> "In the draft guidance we reflect that it is for services to understand the effectiveness of their age assurance methods and processes, in addition to the access control methods and processes. This could be through the service provider's own testing, or by making the relevant enquiries of third-party providers. In practice, where evidence materialises which suggests that there is a reduction in effectiveness in a relevant principle or a combination of principles, services should repeat their children's access assessment". (Vol 2 4.55)

16. There is also a welcome warning to services - contained in volume 4 on risk assessment - that if they are "already implementing measures such that they assess their risk level to be low or negligible, they should continue doing so. Stopping implementing such measures or changing them may constitute a significant change (see Step 4 below) and may increase their risk level." (volume 4 pp56-57). This (to an extent) addresses concerns raised in response to the first consultation that the tick-box, prescriptive

approach to measures in the codes - aligned with the safe harbour promise - could mean services making a decision to stop using existing protective or mitigating measures as they were no longer required to be compliant with the regulation. This is a very welcome shift. However, in terms of upholding age terms and conditions, the proposal is to measure this on a tick-box consistency metric rather than outcomes.

### Some caveats

17. There is no doubt that the combination of the age assurance measures and the new measures relating to recommender systems are significant steps forward in increasing the protections for children, particularly in relation to reducing their exposure to - and the impact of - Primary Priority Content and Priority Content that is harmful and, in some cases, life-threatening. But the limitations of the measures in addressing wider **safety by design** factors remain. compounded by the safe-harbour compliance threshold which does not prioritise overall improvements in the protection of children. For example:
    a. The **age gating requirement** sits on top of all the other obligations and is the only substantive new measure to protect children (and, as such, a single point of failure). The risk assessment obligations in this consultation are no more stringent than those proposed in the illegal harms consultation nor do they have to undertake any significant redesigns of their services as a result of the risks that may be identified. This means that for services, by keeping children off their platform, their obligation - as set out in section 1 of the Act, to "design and operate" safer services to ensure that a "higher standard of protection is provided for children than for adults" - is diluted.
    b. Measures that address **the recommender system** are quite far down the product development and design process. A more robust "safety by design" approach, allied with rigorous risk assessment and product safety testing, would be looking at many more aspects of the overall service before then. (We would refer here to the four-stage model, developed by Prof Lorna Woods in work for Carnegie UK; see p9 here.)
    c. There is a significant gap in the lack of any measure in the codes relating to **livestreaming**, not least as the risk register picks this up as a functionality that causes harm in a number of areas covered by the children's safety duty and the fact that DCMS, back in 2021, specifically included practical guidance for companies on livestreaming in its "Principles of Safer Online Platform Design". Similar gaps, which we cover further in section 10, are evident with location information, large group messaging and ephemeral messaging which Ofcom

identifies have specific risks of facilitating harm to children but which are not covered by any measures.

d. While there is more evidence and commentary presented here by Ofcom than previously on the **influence on the business model** on harms to children, particularly the financial incentives for influencers propagating harmful content or views, there are no new measures proposed to address this.

## Some ongoing concerns

18. We noted that the illegal harms consultation frequently mentioned that the draft codes of practice were **first iterations**; the same is true here. One of the reasons given for this previously was that Ofcom's information-gathering powers only came into effect via a commencement order from 10 January - too late for the first consultation - but it was clear in statements from the Ofcom senior management during the previous consultation that they saw these powers as a route to amassing much more of the evidence they needed to fill in the gaps and/or provide more evidence-based measures for further versions of the codes.

19. In the short timescales between the commencement of the information-gathering powers and the publication of the children's consultation, we would not expect material evidence to have been gathered to influence the proposals. However, we were surprised that these **information-gathering powers** had not even been used by the time the consultation was issued, especially given the number of areas that Ofcom flags as lacking evidence.[1] Given that lack of evidence is frequently cited as a reason for not recommending specific measures (and that lack of evidence does not mean lack of harm), this further delays the production of more robust iterations of the codes.

20. Moreover, as we note below, there is much evidence that has already been amassed by Ofcom in relation to harm that does not lead to a requirement on companies to mitigate that harm. We refer Ofcom here to the advisory from the US Surgeon-General on the need for urgent action to minimise harms to children and adolescents:

> "The current body of evidence indicates that while social media may have benefits for some children and adolescents, there are ample indicators that social media can also have a profound risk of harm to the mental health and well-being of children and adolescents. At this time, we do not yet have enough evidence to

[1] "We have not yet formally requested information from service providers as our information-gathering powers only came into effect in Jan 2024" - para 14.27

> determine if social media is sufficiently safe for children and adolescents. We must acknowledge the growing body of research about potential harms, increase our collective understanding of the risks associated with social media use, and urgently take action to create safe and healthy digital environments that minimise harm and safeguard children's and adolescents' mental health and well-being during critical stages of development." (Social Media and Youth Mental Health: May 2023, p4)

21. We remain concerned in that regard that **Ofcom has not been bold enough**. Arturo Bejar, the Meta whistleblower who has recently testified to the US Congress, observed: "Social media companies are not going to start addressing the harm they enable for teenagers on their own. They need to be compelled by regulators and policy makers to be transparent about these harms and what they are doing to address them." See also Bejar's interview at the recent FOSI conference in Paris.

22. Also as previously, we remain concerned that Ofcom has made a number of choices in how it is approaching the legislative framework that it has not fully justified and which we argue are not required by the language of the Act; there are inconsistencies between its analysis of the harms it has evidenced and the mitigation measures it proposes; and there are some significant judgements (such as the **primacy of costs** in its proportionality approach) on which it is not consulting but which fundamentally affect the shape of the proposals that flow from them. With regard to Ofcom's perspective on costs, these are largely based on companies having to change things as a response to the need for regulatory compliance (eg existing market participants); they do not take into account the impact on new entrants, who would be in a position to design in better safety at (presumably) a lower cost but, under these proposal, would currently have no incentive to do so.

23. Moreover, until we see evidence to the contrary in Ofcom's response to the illegal harms consultation, we are concerned that the framework as proposed at this stage will not be "iterated" in subsequent versions of the codes: the combination of the focus on content-moderation and the rules-based, tick-box approach to governance and compliance is likely to become the baseline for the regime for years to come.

24. The **piecemeal basis** in which Ofcom has approached the selection of measures contained in the codes – only adding those where (in their opinion) there is enough evidence – rather than stepping back to consider the risk-based outcome the legislation

compels companies to strive to achieve continues to concern us. Unless the combined response to the illegal harms consultation and this consultation suggests a significant shift in approach, the chance to introduce (as Parliament intended) a systemic regulatory approach, rooted in risk assessment and "safety by design" principles, will be lost.

25. It is worth noting here the definition of "safety by design", put forward by DCMS in its guidance for companies on the "principles of safer online platform design":

> "Safety by design is the process of designing an online platform to reduce the risk of harm to those who use it. Safety by design is preventative. It considers user safety throughout the development of a service, rather than in response to harms that have occurred."

26. Finally, we do not see how - given that two-thirds of the 36 U2U measures are direct "lift and shift" copies of those in the illegal harms consultation, it is debatable whether (without the age-gating to prevent children accessing services and the, admittedly welcome, measures on the recommender system) the codes here deliver the **"higher protection" to children** promised by the Government. Nor are they sufficiently future-proofed to provide preventative protection as technology evolves.

27. For all the reasons above, we are urging Ofcom to adopt a measure - that is both intended to return the obligation to the providers to try to mitigate risk while evidence on the effectiveness of measures for inclusion in future iterations of the codes is assessed by Ofcom, and a means of future-proofing the codes as new evidence of harms continue to emerge - as we set out at para 8 above.

## Our analysis

28. Following the structure of our previous response, we set out below our analysis of the building blocks of the regime proposed by Ofcom, provide evidence (or refer back to previously cited evidence) for alternative approaches and recommend specific revisions to the codes of practice that we believe can be made in their first iteration, rather than waiting for a second round of consultations.

29. We start with analysis of five fundamental issues which run through the whole regime - and have been replicated in the children's proposals as per the illegal harms proposals - and therefore provide the basis on which many of the specific recommendations are made. These issues are not out for consultation but we hope that in light of our previous feedback, Ofcom has been reviewing the choices they have made here along with the impact they are having on the consequential measures recommended and their likely impact. These issues are:

    1: Weak "safety by design" foundations;
    2: Decisions on the burden of proof/evidence threshold;
    3: The approach to "proportionality";
    4: The approach to human rights.

30. We then look at a series of specific implementation issues that are a concern and cover some gaps in the final section.

31. As previously, we are grateful to the 60+ organisations, experts and academics in our network for their comments and inputs to the series of discussions which have informed our analysis and to the Ofcom representatives who have met with us bilaterally or as part of larger group discussions. We do not speak on the network's behalf and - with regard to this particular consultation - we are not as well placed to speak on some of the recommendations as the leading children's charities, so we defer to their judgement and provide cross-references - where appropriate. In particular, we support the position of the Children's Coalition and their concerns about the approach being taken to age assurance (which might result in under-age children remaining on platforms) and the lack of requirement for a differential experience for children of different ages.

32. We are also submitting parts of this written response via the proforma, where they are relevant to the specific questions contained there.

## The legislative benchmark

33. As with our previous submission, we refer Ofcom to Schedule 4 of the Online Safety Act which sets out the Online Safety Objective. Here it is specifically relevant to note the expectations that "a service should be designed and operated in such a way that" —

    (i) the systems and processes for regulatory compliance and risk management are effective and proportionate to the kind and size of service,

    (vi) the service provides a higher standard of protection for children than for adults,

    (vii) the different needs of children at different ages are taken into account,

    (ix) there are adequate controls over access to, and use of, the service by children, taking into account use of the service by, and impact on, children in different age groups;

    And that it should be "designed and operated so as to protect individuals in the United Kingdom who are users of the service from harm, including with regard to—

    (i) algorithms used by the service,

    (ii) functionalities of the service, and

    (iii) other features relating to the operation of the service."

34. As we set out above and in our specific areas of focus, the choices that Ofcom have made in developing their proposals do not align with the overall objectives of the Act, especially the central element of safety by design. There is a focus on individual measures rather than returning the obligation to service providers to ensure that their services taken in the round are safe. Moreover, in considering the issue of mitigating measures, there is little consideration of how those measures intersect with each other. This does not provide the paradigm shift in safety for children that was envisaged by legislators in passing the Act. There is not enough focus on - or indeed urgency to understand - what the impact is of the very many gaps where Ofcom has determined that evidence is insufficient to make recommendations for measures in the codes of practice.

35. Fundamentally, the lack of an approach within the risk assessment process to what would be termed in other industries "product safety" is as marked here as it was in the

illegal harms consultation. Introducing age gating - as mandated by the Act - will go a long way to keeping children safe by preventing them from using products or by serving them a different content-feed to those experienced by adult users. This is a zero sum approach that suggests children may explore online or they may be safe, but not both - and this runs contrary to the fundamental approach of privacy by design (as outlined by Couvakian[2], specifically principle 4) But this does not equate to upstream product safety testing nor - when regulated services carry out their risk assessment - is there a requirement to mitigate the harms caused by all the functionalities integral to their product, even if their risk assessment identifies that they may be in play.

36. As we said in our previous submission, "this atomistic approach to the codes creates a structural problem: Ofcom is thinking about adding bits on as and when evidence is available rather than stepping back and thinking about how to approach safe design based on a risk assessment: e.g. how do you make a product or a service so that it orientates itself towards safety? We do not think that it is acceptable to address this by promising further iterations to fill the gaps or to add on more individual pieces to the codes."

37. Hence our recommendation at paragraph 8 above: to put in place a system to identify appropriate measures to address the risks arising from the design and functionality of their service, as identified in their risk assessments, that are proportionate to the size and type of service and crucially the severity of the harms (whether understood as the number of people harmed or as the degree of harm particular individuals suffer), bearing in mind best practice and the state of the art.

[2] Ann Cavoukian, Privacy by Design - The 7 Foundational Principles: Implementation and Mapping of Fair Information Practices, available: https://privacy.ucsc.edu/resources/privacy-by-design---foundational-principles.pdf

## Issue 1: Weak "Safety by Design" Foundations

### Issue

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. It says that "The duties set out in subsections (2) and (3) apply across all areas of a service, including the way it is designed, operated and used as well as content present on the service, and (among other things) require the provider of a service to take or use measures in the following areas, if it is proportionate to do so—

> (a) regulatory compliance and risk management arrangements,
>
> (b) design of functionalities, algorithms and other features,
>
> (c) policies on terms of use,
>
> (d) policies on user access to the service or to particular content present on the service, including blocking users from accessing the service or particular content,
>
> (e) content moderation, including taking down content,
>
> (f) functionalities allowing users to control the content they encounter,
>
> (g) user support measures, and
>
> (h) staff policies and practices.

We set out the detail of Schedule 4 (the Online Safety Objectives) above.

Also relevant here is part of 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 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 recommender 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 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 section nine.

Recommender systems

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.

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)

Business model

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 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 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)

Metrics

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." (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).

Addictive design

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 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 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 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. (We set out more on the implications of the differentiated approach to size in Ofcom's proposals in section six, 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 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?

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 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
- 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

## Recommendation

Supported by the evidence and analysis we provided previously, we repeat our recommendation that Ofcom makes a small but significant change to its draft codes of practice for both illegal harms and children's protection. This would put a requirement on all regulated companies specifically to take measures to address harms that have been flagged in their risk assessment that arise from the features and functionalities of their service, drawing on current good practice, and to regularly monitor the measures' effectiveness. (Current good practice could include interventions that Ofcom has discussed but for which the evidence base is missing at the moment.) This provides an interim step, in the absence of the evidence Ofcom feels it requires to recommend specific measures, that would go a long way to ensuring that the regulatory regime begins on the right footing and starts, from the outset, delivering the "safety by design" intent of the Act and the general mitigation duty at section 12 2(c) for user-to-user services and 28 (2) for search.

We also recommend that product testing should be included in the codes of practice, appropriate to the size of the company and the risks its products pose, and that the results of this testing should be a core input to the risk assessment.

We suggest the following wording is inserted in the draft codes for both illegal harms and protecting children, between the section on governance and accountability and the section on content moderation, which 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).

"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 measures taken in reducing the risk of harm identified in the risk assessment. These measures should feed back into the risk assessment."

As we said in paras 9-10 above, 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 pre-determined 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); 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.

## Issue 2: Decisions on the Burden of Proof/Evidence Threshold

### Issue

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 measuresand for Ofcom to use this practical evidence to change the recommendation or add additional measures. (See section 5 on measures and the codes below.)

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 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 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 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 (see our references provided in section 1, above) 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 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 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 also our discussion on the measures in the codes of practice in section 5, below.

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 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 dopamine 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. (See section 1 above.)

**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 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 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 in section 1 above, with suggested wording, 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.

## Issue 3: the approach to proportionality

### Issue

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 - which we discuss further below. We look below in section 7 at how the principle of proportionality plays into Ofcom's differentiated approach to small and large companies.

### 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. The safety duties for children are set out at Section 12; Section 13 then sets out the interpretation of the safety duties, including this on "proportionate"—

> (a) all the findings of the most recent children's risk assessment (including as to levels of risk and as to nature, and severity, of potential harm to children), and
>
> (b) the size and capacity of the provider of a service.

A comparable approach to proportionality is found in the analogous provisions for search services in section 30.

In Schedule 4, which sets out details on how Ofcom should approach the codes of practice, it says:

> 2 (c) the measures described in the code of practice must be proportionate and technically feasible: measures that are proportionate or technically feasible for providers of a certain size or capacity, or for services of a certain kind or size, may not be proportionate or technically feasible for providers of a different size or capacity or for services of a different kind or size; (NB this does not mention cost in relation to proportionality)

2 (d) then makes a specific reference to proportionality in relation to the risk of harm:

> "the measures described in the code of practice that apply in relation to Part 3 services of various kinds and sizes must be proportionate to OFCOM's assessment (under section 98) of the risk of harm presented by services of that kind or size."

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 in section 1 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.

## Issue 4 The approach to human rights

We published a detailed analysis on this issue by Prof Lorna Woods in relation to the illegal harms consultation to which we refer Ofcom as evidence in this section. In relation to the children's consultation, we would like to draw attention to the following points and provide a bit of extra commentary below:

- The reference in the summary document to "content that is legal but is nevertheless harmful" does not take into account content that might infringe rights (eg privacy) within a regulatory rather than criminal regime. (Summary doc p10 para 2.17)
- The discussion of human rights in volume 1 notes abuse etc but doesn't recognise in its analysis from other perspectives than the speaker. The issues that are picked up in this volume are not followed through in the rest of the consultation.
- Similarly, the reference to the UNCRC (para 2.49) is not pulled through elsewhere.
- Reference to children being "discouraged" from expressing themselves online (vol 3 para 6.3).
- The discussion at vol 3, page283 onwards doesn't pick up that restrictions here are about reach and not about prohibition so are less intrusive.

Ofcom's commentary notes that the rights analysis is complex given the need to balance the rights of multiple users and has to take into account the adverse impact of the exercise by one person's freedom of expression rights on others' ability to exercise their rights, as well as a state's positive obligations in this context.

In Volume 3, Ofcom notes the silencing impact on children being discouraged from expressing themselves, and the fact that this affects those in minoritized groups particularly. Rights here are not being equally protected, yet the rights enumerated in the Convention are to be enjoyed without discrimination. Nonetheless despite this initial analysis, the rights assessment was not fully pulled through into the discussion of measures, and specifically the impact of the UNCRC, mentioned in Vol2, was not pulled through and analysed in the context of the risk register or the code of practice.

While, on the whole, the rights analysis (based on the Convention) did not prevent the adoption of measures, it is unclear what role rights played in relation to issues which were just not discussed. There was no explanation of measures that had been considered but not adopted. This gap means it is also unclear the extent to which Ofcom had regards to the need to protect fundamental rights. We would however like to emphasise that, in terms of a proportionality analysis in the context of human rights, measures which relate to limiting reach (rather than taking down content) have been considered to be less rights intrusive - see for example, the report of Irene Khan on Gendered Disinformation. There, the Special Rapporteur remarked:

> "Systemic regulation, which emphasises "architecture over takedown", allows for more proportionate responses and is likely to be better aligned with freedom of expression standards."

## Issue 5: disconnect between risk analysis and the recommended mitigation measures

### Issue

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, which we cover in the safety by design section above.

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 section 3 above))

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

We included the relevant text from Section 12 (4) on the children's safety duties above.

Section 236(1) of the Online Safety Act then describes "measures" as follows

> "any reference to a measure includes a reference to any system or process relevant to the operation of an internet service or any step or action which may be taken by a provider of an internet service to comply with duties or requirements under this Act."

In addition, Schedule 4 of the OSA sets out the approaches that Ofcom must take to drawing up the codes of practice. Under the General Principles, it says:

> (d) the measures described in the code of practice that apply in relation to Part 3 services of various kinds and sizes must be proportionate to OFCOM's assessment (under section 98) of the risk of harm presented by services of that kind or size.

Schedule 4 also includes the following at section 3: OFCOM must ensure that measures described in codes of practice are compatible with pursuit of the online safety objectives, which we have extracted at (page/para) above. As well as setting out a number of objectives relating to systems and processes in section 3(a), the objectives specify at 3(b):

> (b) a service should be designed and operated so as to protect individuals in the United Kingdom who are users of the service from harm, including with regard to—

(i) algorithms used by the service,

(ii) functionalities of the service, and

(iii) other features relating to the operation of the service.

Schedule 4 requires that the recommendations be clear and precise, but this does not mean that the service providers should have no freedom of choice.

Finally, Schedule 4 also requires that Ofcom ensure that (9(1)) Codes of practice that describe measures recommended for the purpose of compliance with a duty set out in section 10(2) or (3) (illegal content) must include measures in each of the areas of a service listed in section 10(4) (our emphasis).

As we can see above, 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 section 2, 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 section 2, 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 Section 3.

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 section 7; 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 section 2) 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** (see section 10 below), **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)
- Ofcom identifies the **risks arising from Gen AI** (particularly the links between immersive environments and bullying, vol 3, para 7.5.60) and the fact children are early adopters of new technologies and "gen AI models can present a risk of harm to children", para 7.14.22). Despite this, it concludes that "the evidence base for children's interaction with genAI will be limited" and does not suggest a corresponding measure (see analysis in section 2 above and section 10, below)

## Recommendation

We refer Ofcom back to our recommendation in section one which sets out additional measures to be added to the draft codes of practice which require companies to take mitigating measures

based on the risks arising from their services' "functionality, algorithms and features" that they have identified in their risk assessment.

## Issue 6: Small vs Large Companies Makes Size Rather than Risk the Primary Aspect

### Issue

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, which we set out in section 5, 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. .

### What the Act says

Section 7(2) says: "All providers of regulated user-to-user services that are likely to be accessed by children must comply with the following duties in relation to each such service which they provide—

(a) the duties about children's risk assessments set out in section 11, and

(b) the duties to protect children's online safety set out in section 12(2) to (13).

Section 12 (8) for user-to-user services, which we discuss above, sets out the types of measures providers may be "required" to take "if it is proportionate to do so". Section 13 for user-to-user and section 30 for search define "what is proportionate for the purposes of this section", stating that "the following factors, in particular, are relevant—

(a) all the findings of the most recent children's risk assessment (including as to levels of risk and as to nature, and severity, of potential harm to children), and

(b) the size and capacity of the provider of a service.

Section 91 also inserts into Section 3 of the Communications Act 2003 new duties on Ofcom, including:

"(4A) …. OFCOM must have regard to such of the following as appear to them to be relevant in the circumstances—

(a) the risk of harm to citizens presented by regulated services;
(b) the need for a higher level of protection for children than for adults;
(c) the need for it to be clear to providers of regulated services how they may comply with their duties set out in Chapter 2, 3, 4 or 5 of Part 3, Chapter 1, 3 or 4 of Part 4, or Part 5 of the Online Safety Act 2023;
(d) the need to exercise their functions so as to secure that providers of regulated services may comply with such duties by taking measures, or using measures, systems or processes, which are (where relevant) proportionate to—

(i) the size or capacity of the provider in question, and
(ii) the level of risk of harm presented by the service in question, and the severity of the potential harm"

### 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 equivalent** | **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 selfharm 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 reports covered in section 4, above.

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 in section one, above, for addition to the draft codes as we recommend that this applies to all services regardless of size.

## Issue 7: Governance and Risk Assessment

### Issue

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 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 section 2.

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.

### What the Act says

The risk assessment duties are at section 11 for User to User and section 28 for Search. Regulated services are required to carry out a "suitable and sufficient" risk assessment, keep it up to date and redo it "before making any significant change to any aspect of a service's design or operation." For User-to-User services, section 11 (6), requires that the risk assessment to take into account "the risk profile that relates to services of that kind" —

(a) the user base, including the number of users who are children in different age groups;
(b) the level of risk of children who are users of the service encountering the following by means of the service—

(i) each kind of primary priority content that is harmful to children (with each kind separately assessed),
(ii) each kind of priority content that is harmful to children (with each kind separately assessed), and

(iii) non-designated content that is harmful to children,

giving separate consideration to children in different age groups, and taking into account (in particular) algorithms used by the service and how easily, quickly and widely content may be disseminated by means of the service;

(c) the level of risk of harm to children presented by different kinds of content that is harmful to children, giving separate consideration to children in different age groups;

(d) the level of risk of harm to children presented by content that is harmful to children which particularly affects individuals with a certain characteristic or members of a certain group;

(e) the extent to which the design of the service, in particular its functionalities, affects the level of risk of harm that might be suffered by children, identifying and assessing those functionalities that present higher levels of risk, including functionalities—

(i) enabling adults to search for other users of the service (including children), or
(ii) enabling adults to contact other users (including children) by means of the service;

(f) the different ways in which the service is used, including functionalities or other features of the service that affect how much children use the service (for example a feature that enables content to play automatically), and the impact of such use on the level of risk of harm that might be suffered by children;

(g) the nature, and severity, of the harm that might be suffered by children from the matters identified in accordance with paragraphs (b) to (f), giving separate consideration to children in different age groups;

(h) how the design and operation of the service (including the business model, governance, use of proactive technology, measures to promote users' media literacy and safe use of the service, and other systems and processes) may reduce or increase the risks identified.

A smaller set of factors are included at section 28 (5) for search.

### 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 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 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.
- 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 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 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 the recommendation in section 1 above - 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.

## Issue 8: Age assurance

### Issue

We have noted in section 1 (above) that the implementation of age assurance measures is not a fundamental "safety-by-design" measure. For services that are fundamentally harmful - eg their content is entirely inappropriate for children and under-18s - this is right. For others, the requirement builds in safety to the architecture of the service through age gating - either in its entirety or in part, based on the types of content it serves - but the service that sits behind the age-gating may not intrinsically be made any safer.

The alignment of the approach taken by Ofcom on governance, risk assessment and (most of) the measures in the codes of practice between the illegal harms consultation and the children's consultation bears this out. Therefore, there is potentially an incentive for services that *could* make themselves safer to decide not to bother with the extra costs that might be incurred and just bar children from accessing their sites.

The approach taken to age assurance draws from the proposals for the part 5 guidance for pornography providers, which Ofcom consulted on earlier in the year. Ofcom says: "The overarching aim of age assurance measures for services under the children's safety duties is to help ensure children are protected from harm and receive age-appropriate experiences. We have also aimed for alignment with Part 5 guidance to create a clear and consistent regulatory regime for services." Ofcom does not, however, set out any measures or guidance for platforms to provide a range of "age-appropriate" experiences: it is a one-size-fits-all requirement for those that might be accessed by children.

### What the Act says

At Section 11 (3), the Act says that service providers have "a duty to operate a service using proportionate systems and processes designed to

(a) prevent children of any age from encountering, by means of the service, primary priority content that is harmful to children;

11 (4) then says: The duty set out in subsection (3)(a) requires a provider to use age verification or age estimation (or both) to prevent children of any age from encountering primary priority content that is harmful to children which the provider identifies on the service.

And at 11 (6): If a provider is required by subsection (4) to use age verification or age estimation for the purpose of compliance with the duty set out in subsection (3)(a), the age verification or

age estimation must be of such a kind, and used in such a way, that it is highly effective at correctly determining whether or not a particular user is a child.

There are no comparable requirements within the Act for search services.

### Ofcom's proposals

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 - e.g. 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.

## Issue 9: Violence Against Women and Girls (VAWG)

### Issue

Forty-four organisations and individuals signed an open letter to Ofcom's CEO, Melanie Dawes, about their concerns with the approach taken in the illegal harms consultation, which remain valid in respect of many aspects of the children's consultation. While Ofcom officials - at all levels of the organisation - are keen to stress in public and in private that the protection of women and girls is a key priority for them in their implementation of the Online Safety Act regime, the foundations on which the guidance on VAWG (due next spring) will sit are - we fear - not strong enough to provide the level of protection promised by the previous Government during the passage of the Bill through Parliament. We set out more detail on our concerns below. A fuller version of this analysis will be submitted separately to Ofcom.

### Analysis

Weak levels of protection

The intersection of the measures proposed here and the forthcoming guidance that Ofcom needs to produce by spring next year with regard to protecting women and girls is important.

The VAWG sector campaigned strongly for a mandatory code of practice to be included in the Act and an amendment to that effect from Baroness Morgan had cross-party support during the passage of the Bill. In the Lords' debate on that amendment, the then Government Minister, Lord Parkinson, suggested that the existing codes of practice on illegal harms and children's safety would be enough:

> "all service providers must understand the systemic risks facing women and girls through their illegal content and child safety risk assessments. They must then put in place measures that manage and mitigate these risks. Ofcom's codes of practice will set out how companies can comply with their duties in the Bill. *I assure noble Lords that the codes will cover protections against violence against women and girls. In accordance with the safety duties, the codes will set out how companies should tackle illegal content and activity confronting women and girls online. This includes the several crimes that we have listed as priority offences, which we know are predominantly perpetrated against women and girls. The codes will also cover how companies should tackle harmful online behaviour and content towards girls."* (Our emphasis) (Hansard: 16 May 2023)

Eventually, the Government conceded and brought forward its own amendment to require Ofcom to produce guidance on VAWG. When he spoke to this amendment, Lord Parkinson again

stressed how the codes of practice were a fundamental part of delivering improved protections for women and girls:

> "Ofcom's codes of practice will set out how companies can comply with the duties and will cover how companies should tackle the systemic risks facing women and girls online. Stipulating that Ofcom must produce specific codes for multiple different issues could, as we discussed in Committee, create duplication between the codes, causing confusion for companies and for Ofcom … G*overnment Amendment 152 will consolidate all the relevant measures across codes of practice, such as on illegal content, child safety and user empowerment, in one place, assisting platforms to reduce the risk of harm that women and girls disproportionately face*." (Our emphasis) (Hansard: 12 July 2023)

There are two points here: what is in the code(s) matters as to how effective the VAWG guidance can be; the guidance is not an alternative route to providing mandatory protections as it is not enforceable. The code(s) will set the foundation for how effective the more wide-ranging guidance can be in changing the culture of online VAWG, both in terms of services' prioritisation of measures or product redesigns to reduce it and the experience of users as a result. Secondly, VAWG guidance that relates to risks to women and girls that have been included in either of the risk registers relating to the (mandatory) codes will not be enforceable; without an overarching obligation to put in place mitigating measures to address design or functionality risks identified in the risk assessment (as in section 1 above), companies do not have to act on them. It is important then that Ofcom gets the balance right between what is in the code(s) and what is in guidance. Without a sufficient suite of measures (see analysis here) to address the identified risks of harm to women and girls in the codes - the "relevant measures" which Parkinson envisaged would be consolidated in the guidance - then the guidance itself risks being insufficient.

Gendered harms

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 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.")

Age-appropriate experiences

Ofcom's decision not to require services to deliver age-differentiated experiences for users under-18 - which the Children's Coalition have flagged in their response - is also problematic. For example, para 8.2.9 refers to BBFC and telecoms operators standards in relation to porn but there is no consideration given to the fact that this is an under-18 blanket age restriction and there should be a watershed comparison for younger age groups. The definition of porn as PPC means it's narrowly focused but there isn't any additional consideration for sexually suggestive material which might be harmful to young children (as identified by their assessment of harms).

We note that - as in many other areas - Ofcom cites “limited evidence” as the reason for not recommending differential measures for different age groups, despite the fact that (at 15.98): “We also note that the severity of impacts faced by children within particular age groups when exposed to PC may vary quite significantly and some children will be more vulnerable than others, even in older age groups such as neurodivergent children and children whose gender, race and sexuality may impact the harm they experience from content outlined in Sections 7.4-7.8 in Volume 3 the causes and impacts of harms to children.”

Features and functionalities

We welcome the controls around recommender systems, which would be likely to have a cross-harm effect including for issues more likely to impact girls. But other issues and specifically functionalities are not thoroughly dealt with. These include issues where anonymous or fake accounts are a specific factor - for example, material containing self-harm which girls have an increased likelihood of encountering. There are VAWG aspects to services which allow the creation of multiple/disposable accounts - this might have links to sub-criminal stalking, for example, or bullying. Here the response is not about stopping the problem (through perhaps considering checks on users with multiple accounts) but by putting the onus on users to block/mute accounts (21.76). While the proposed measure is welcome, it does not go to the route of the problem.

In the context of self-harm material and also in relation to eating disorder material, for example, Ofcom also notes the impact of likes as validation (which arguably has impacts elsewhere too), but these are not considered in the Codes. While Ofcom suggests some limitations on being added to groups (but not for all services), it does not address stranger pairing which was highlighted in relation to abuse (which can have a gender-based component). In a number of instances, the business model is relevant but again not dealt with in the codes. We suggest that while the proposals on age-gating and recommender systems are important steps, that more should be done to tackle other functionalities - including those higher up the communication chain - and that obligations in relation to them (even a programmatic obligation such as we set out above) should be included - but that in that instance, understanding harm and solutions should be seen through a lens of gender. While we note that Ofcom has chosen to prioritise certain measures which it believes will materially improve the position for children (14.34), it is not clear on what basis this selection was made.

Content moderation

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.

Small platforms

We have noted that some services are subject to more limited obligations because of their size. Some of those obligations are, however, central to safety and a key example of this is guidance and training for moderators - Ofcom notes the difficulties in identifying harms in some context (eg self-harm; eating disorder) and these areas are ones in which the differential impact of harm has been noted. The obligation to train in relation to a topic should relate to the risk in relation to that subject on the particular service - not to the service's size, or how many risks the service faces. (Ofcom notes the evidence previously provided by Glitch on moderator training in gender-based violence at para 16.226.) This should be a base level obligation for all services - and as Ofcom notes, the scale of the job will vary so single risk platforms will have less to do.

Burdens on children

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.

### Recommendations

We refer Ofcom to the full response from the VAWG sector coalition which sets out a list of evidence-based recommendations to improve the measures in the codes ahead of the publication of the VAWG guidance early next year.

## Issue 10: Gaps and other consultation issues

### Issue

In this section we cover a number of issues emerging from the consultation, including gaps in the proposals that either have not been acknowledged by Ofcom or have been acknowledged but could be (partially) filled and some other points.

#### Gaps in protections

Ofcom has identified a number of gaps where it is looking to improve its evidence base before taking action:

> "We've pinpointed several critical areas that demand urgent attention and possibly further action. These include using automated content moderation to detect illegal and harmful content on a large scale, addressing the risks children face from emerging generative AI technologies, and tackling features that entice children to increase their screen time. Furthermore, we're exploring more tailored protection strategies for different age groups and examining how parental controls can not only empower parents but also enhance their children's safety online." (volume 5, 13.60)

We have concerns about the time it will take to amass this evidence and then to formulate a measure for inclusion in the codes to deal with the risks of harm which (in many cases) are already evidenced to a greater or lesser degree. We set out a few of these concerns below.

<u>Emerging technologies - metaverse, genAI etc</u>

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), and GenAI is linked to both eating disorder content (7.3.57) and bullying (7.5.87). 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.

Ofcom note, also in volume 3, that children are early adopters of new technologies: “Children are often early adopters of new technologies, and generative artificial intelligence (GenAI) models can present risk of harm to children. There is emerging evidence indicating that GenAI can facilitate the creation of content harmful to children, including pornography, content promoting eating disorders, and bullying content, which can be shared online and potentially encountered by children” (section 7.14). This early adoption tendency is earlier flagged as a key driver in services’ growth strategies too “given that user growth may directly reflect an increase in children using the service or an increased likelihood of a service appealing to children. A comparative example may be taken from GenAI. CHILDWISE research found that 59% of online 7–17-year-olds had used any of the following GenAI tools: ChatGPT, Midjourney, DALL-E, Snapchat MyAI – all of which were made available to the public in the last 2–3 years. Data from Ipsos iris suggests that the reach of the OpenAI website / ChatGPT among 15–17-year-olds rose in line with its growing popularity among adults between Nov 2022 – May 2023 (grew from

<50k to over 500k). This reflects that a rapid user base expansion can encompass a growth in children's engagement as well."

Yet, in a footnote on page 13 Ofcom says: "We are aware of the debate around the potential risks that GenAI may pose. Given the pace of developments in GenAI, and because the evidence base in this area is still developing, we have considered this technology in a limited way in this version of the Register. Our draft Register considers some of these risks." (Vol 3, p13)

And in the longer discussion in section 7.14, Ofcom says: "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." Ofcom then details their call for evidence and the programme of work they are undertaking to "understand more about the risks GenAI poses to children" and "explore" how regulated services are approaching safety for AI-generated content.

As we set out in our summary section above, this is absolutely a case for a precautionary approach - using the measure we suggest in the introductory section - to allow for protections to children while the evidence base develops. Not waiting for a number of years, as Ofcom seems to be prepared to do, to suggest measures for mitigating the risks.

VPNs

Risks arising from the use of VPNs are mentioned in a number of areas throughout the consultation docs, including in relation to pornography (7.1.4)., but there are no recommendations as to what to do about this and the workarounds that VPNs offer for services that wish to avoid regulatory compliance are not addressed.

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, not just those where "there is a medium or high risk for at least one of these kinds of content".

Reporting and complaints

We have noted above 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) 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' sub-section."

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.

## Iterative approach

We noted in our previous response that the iterative nature of the illegal harms codes was disappointing; their publication within a month of Royal Assent was cited as one of the reasons for a trade off between speed vs comprehensiveness. Six months on, there are still gaps in the children's codes and a reliance on the iterations of codes to fill them.

> "The proposals in this consultation mark a vital first step toward safeguarding children

> online. We're committed to continuously refining our strategies based on a dynamic understanding of both the digital landscape and children's experiences on the internet. Through an active programme of research and ongoing dialogues with services —including targeted information requests—we aim to keep our approach fresh and effective."

As we note above, Ofcom has not yet used its information-gathering powers even though they now have them, unlike when they published the illegal harms codes. As previously, there are no timescales for these subsequent iterations nor a sense of what evidence will be needed? (The calls for evidence within this document have a fairly vague timeframe.)

We remain of the view that there is a significant risk as a result that the "regime gets embedded in this "lowest common denominator" form and watered down, via company lobbying, judicial review actions etc, from there, rather than being built on stronger foundations and continuously improved."

**July 2024**

**Contact: maeve@onlinesafetyact.net**