Your response

Question	Your response
Question 1: Do you have any comments on our proposed approach to 'content and activity' which 'disproportionately affects women and girls'?	Confidential? – N One critical gap we observe in the current guidance is the lack of explicit attention to children's unique vulnerabilities in the context of gender-based online harms. The guidance should explicitly recognise how cognitive, emotional, and social development stages intersect with gender norms to shape risks. Additionally, while overt harms such as CSAM are more readily identified and moderated; subtle and cumulative harms across childhood, such as the reinforcement of oppressive gender stereotype and bias, often go undetected and unchallenged. We need a more fully developed *proactive* underpinning framework that integrates, rather than relying primarily on, *reactive* safety measures. This would provide *both* more robust regulatory outcomes for all children and young people *and* greater agility in responding effectively and quickly to new technological developments, and how children use them. There are four key ways in which the proposed approach could be strengthened to more effectively reflect the developmental and gendered nature of online harm for under-18s. **A. Integrate a child development and gender vulnerability framework** Do's review research on GenAI chatbots and children shows that girls under 18 face compounded risks in online spaces due to their developmental stage and gender-specific targeting (e.g., body shaming, sexual grooming). Left unaddressed, this early exposure can shape their perceptions of acceptable behaviour, relationships, and self-worth, entrenching patterns of inequality and harm into adulthood. A framework that recognises these compounding vulnerabilities would ensure the guidance addresses not just illegal harms but the cumulative shaping of harmful behavioural norms. **B. Adress cross-gender social norm formation** While content often targets girls, the effects, although different, can also be important for *boys and non-binary youth* in the longer term. Harmful gender norms shape behaviours, language, and expectations between peers at school, home, and play. Once these patterns take root, they become difficult to challenge. Interventions should therefore not only protect girls but build broad-based interpretive and resilience skills in all children to disrupt this social 'leakage'. **C. Embed preventive, interpretive design interventions for children** The draft guidance rightly encourages reactive tools (e.g., reporting, moderation), but stronger emphasis is needed on preventive, design-

Question	Your response
	level interventions tailored for children at different ages. Platforms should integrate tools and prompts that help children interpret and question harmful norms, especially those that are subtle or socially reinforced. Normative harms, such as exclusionary messaging, stereotype reinforcement, or biased recommendation patterns, require proactive, systemic responses. The guidance should go further in setting clear expectations for how services detect, surface, and mitigate these less visible risks. Examples include narrative prompts, visual flags, and "pause and reflect" moments tailored to child users. These design interventions advance both Action 5 and Action 7 by supporting child agency and resilience. **D. Adopt a child-centred design approach to underpin these measures**. This meansbuilding environments that are developmentally and age-appropriate, and empowering. For example, interfaces, prompts, language, and support are matched to the child's age, cognitive ability, and emotional maturity. This is more than simply blocking or flagging content. It addresses how content is recommended systemically; how interactions are shaped (e.g., through gamification or popularity cues); and what coping resources are offered when harm occurs. More importantly, children themselves, across gender, socio-economic background, ability, and culture, should be actively involved in the design, development, and ongoing feedback processes. Initiatives such as the Children's AI Summit by The Alan Turing Institute illustrate the importance and feasibility of child participation in shaping technology systems that affect them
Question 2: Do you have any comments on the nine proposed actions? Please provide evidence to support your answer.	Confidential? – N *Action 6: Reduce the circulation of online gender-based harms* **A. Limitations of hash matching alone** While hash matching is a valuable starting point for reducing the circulation of known non-consensual intimate images (NCII), my research and engagement with technology governance show that this approach has critical limitations in today's evolving digital landscape. • hash matching only works for content that has already been identified and reported. It cannot prevent the spread of newly created abusive materials**,** including those generated using AI tools. • perpetrators can easily bypass basic hashing systems by making minor edits (e.g., cropping, resizing, adding filters) which often renders the hash unrecognisable. • hash-matching is focused on static images. It is not sufficient for detecting harmful videos or increasingly prevalent deepfakes, which present a rapidly growing threat to women and girls online. **B. Support for proactive, AI-based detection tools**

Question	Your response
	We recommend combining foundational approaches like hash matching with AI-based detection tools that align better with the safety-by-design approach. We already have concrete examples of how this can work: • NSFW Classifiers can identify intimate and explicit content using convolutional neural networks, even when it hasn't been previously flagged • Microsoft's Deepfake Detection Tools analyse subtle inconsistencies in facial movements and visual cues to detect synthetic videos created for harassment. • Multimodal AI Models go further by combining video analysis, speech recognition, and NLP to detect unsafe or risky conversations in private conversations. • The iCOP 2.0 system, developed by the University of Bristol and deployed in Southeast Asia, shows how AI-based tools can strengthen law enforcement's response to online child sexual exploitation and abuse. This model offers a powerful example of how such technologies can be adapted to address gender-based harms more broadly. This is also strongly supported by academic research. Thiel (2023) has demonstrated that AI models can identify and eliminate harmful content, including CSAM, even when embedded in generative ML training data. The study identified hundreds of known and new CSAM images within the LAION-5B dataset, which was used to train models like Stable Diffusion, a common foundation for many downstream generative AI applications. These findings highlight the urgent need for stronger safeguards in dataset curation, model development, and the hosting of generative AI systems, especially as these models become more accessible, open-source, and vulnerable to misuse or accidental reproduction of harmful content. **C. Limitations and risks of automated content moderation (ACM)** "*2.70. Many of the methods in this section (persuasion, removal, and reduction) often rely on automated processes. Importantly, many of them may use 'proactive technology' within the meaning of the Act*." While Ofcom acknowledges the role of automated processes in persuasion, removal, and reduction strategies, we urge caution against assuming that these technologies are, or can become, sufficiently accurate, nuanced, and unbiased in addressing complex gender-based harms given the current state-of-the-art. Automated systems still struggle to understand coded misogyny, cultural nuances, and the power dynamics embedded in online harassment. This results in harmful content slipping through undetected (false negatives) and, equally concerning, the removal of survivor voices or important conversations about gender-based violence (false positives).

Question	Your response
	To improve the effectiveness and accountability of ACM, I recommend: • **Adopting a multi-layered approach**: ACM must be complemented by trained human moderators, particularly for high-impact and context-dependent cases. Automated systems should flag potential harms, but final decisions, especially concerning gender-based violence, should involve human oversight. • **Enhancing training data for intersectionality**: Platforms should be required to evidence the diversity of their training datasets to better capture the lived experiences of women and girls from varying backgrounds, including considerations of race, class, age, sexuality, and disability. • **Transparency and user control**: ACM decisions must be explainable, with clear appeals processes. Women and girls should have greater control over their online experience, including granular privacy controls and the ability to pre-approve interaction. These safeguards must also be developed with clear, age-appropriate, and easy-to-use tools that help children of different ages understand what they are signing up for. • **Independent auditing and public reporting**: Platforms should undergo regular, independent audits of their moderation systems and publish disaggregated data to ensure transparency and accountability in how gender-based harms are being addressed. This could involve setting up a published scorecard, overseen by Ofcom (see response to question 4). **Supporting research:** • Lee, H. E., Ermakova, T., Ververis, V., & Fabian, B. (2020). *Detecting Child Sexual Abuse Material: A Comprehensive Survey.* Forensic Science International: Digital Investigation, 34, 301022. • Thiel, D. (2023). Identifying and Eliminating CSAM in Generative ML Training Data and Models. Stanford Digital Repository. Available at https://purl.stanford.edu/kh752sm9123. https://doi.org/10.25740/kh752sm9123.
Question 3: Do you have any comments about the effectiveness, applicability or risks of the good practice steps or associated case studies we have highlighted in Chapter 3, 4 and 5?	Confidential? – N Current reporting mechanisms are often intimidating or confusing for younger users, requiring complex navigation or adult-like articulation of issues. I suggest: **A. Make child reporting mechanisms age-appropriate and intuitive** I recommend that Ofcom's guidance explicitly address the design of child-friendly reporting and support systems. Children must feel safe, empowered, and capable of seeking help when they encounter harmful content or experiences. Some of child-friendly and age-appropriate reporting

Question	Your response
Are there any additional examples of good practices we should consider? Please provide evidence to support your comment.	options could be explored such as: Visual icons and guided input rather than open-text boxes; Confidential routes to trusted services (e.g., helplines, NGOs); Simple and age-appropriate language. The guidance should encourage platforms to provide clear, accessible pathways for children to seek confidential support, including partnerships with child-focused helplines and NGOs. **B. Reduce reliance on reporting alone** Children may be unable or unwilling to report. At younger ages, they lack developmental capacities for rational reflection. Adolescents need to take risks and experiment with autonomy, but can lack adults' capacity to understand long term implications of actions, especially when these implications are not directly physical. We support embedding the following types of measures, adapted by age: • Algorithmic 'exposure breaks' (nudges, limiters) especially for younger children • Behavioural analytics for when children experience escalating and intensifying exposure to harmful material ("boiling frog" scenarios) • Built-in check-ins or pop-up interventions for adolescent users These enhancements support a more active safety-by-design approach (Action 4, Action 5). **C. Draw on child psychology expertise in design** The design of moderation and support should integrate child development knowledge, particularly around the need for agency, and the importance of self-expression, and shame, at different developmental stages.
Question 4: Do you have any feedback on our approach to encouraging providers to follow this guidance, including our proposal to publishing an assessment of how providers are addressing women and girls' safety? Do you have any examples or suggestions of other ways we could encourage providers to take up the 'good practice' recommendations?	Confidential? – N I strongly support Ofcom's plan to publish assessments of provider actions. This is an important accountability mechanism that can put real public pressure on platforms to improve their safety measures. To strengthen this further, I recommend the following: • **Develop a public rating or scorecard system** on safety by service, similar to environmental or data privacy scores. This should include transparent metrics on how platforms handle age-disaggregated, gender-based harms, systemic measures, moderation effectiveness, user controls, and response times to reports of abuse. • **Offer technical support grants or pilot collaborations:** Ofcom could act as a convenor, connecting committed online platforms with technological solutions and expert partners through initiatives such as safety sandboxes in collaboration with universities or regulators, allowing platforms to test moderation tools in controlled environments. Or targeted technical support grants to help platforms assess, implement, and refine content moderation technologies with expert input.

Question	Your response
	• **Facilitate co-creation panel(s) between platforms, survivors, and civil society**, ensuring that the voices of those with lived experience of online abuse are central to the design and evaluation of safety measures (see reponse to question 5, below). These panels could also be used to develop offline tools, resources, and training materials for parents, educators, and other responsible adults, helping to prevent the 'leakage' of harmful behaviours and norms from online spaces into offline contexts. • **Support the development of shared safety tools and technologies** that can be adopted across platforms, reducing duplication of effort and raising the baseline of safety industry-wide. This includes investment in open-source safety technology, such as advanced content moderation tools and reporting systems.
Question 5: Do you have any comments on our impact assessment, rights assessment, or equality impact assessment? Please provide any information or evidence in support of your views.	Confidential? – N To meet its regulatory obligations under Section 54 and the Human Rights Act, Ofcom must ensure rights-based frameworks translate into meaningful interventions. These comments support Actions 1, 2, and 3. I suggest: • **Make lived experience and child-centred design central to understanding platform dynamics** by formally integrate the voices of those with lived experience, especially children and young people, into the design, implementation, and ongoing evaluation of safety measures. This is not simply about inclusion and participation; it is the *only* way to understand how different age groups actually use, experience, and are affected by online platforms in real time. Some features may pose unexpected risks for particular age groups, while others may offer protective effects that only emerge through direct engagement with users. Ofcom should therefore support the creation of **standing participation panels**, including survivor panels, youth advisory boards, and expert working groups. Such panels are common in the health field that has similarly consequential and complex developments (called 'public and patient involvement and engagement', PPIE). Panel(s) must not be one-off consultations but embedded mechanisms for continuous insight, enabling more agile and responsive regulatory decision-making. Participation panels offer a vital early warning system—identifying emerging harms, surfacing age-specific challenges, and tracking how responsible adults, such as parents and educators, are actually navigating and mitigating these risks. • **Provide clearer guidance on how to avoid reinforcing surveillance harms, particularly for women and girls in minoritised communities.** While content moderation and platform safety interventions are nec-

Question	Your response
	essary, they must not inadvertently enable over-policing or surveillance of already marginalised groups. The guidance should explicitly state that data collection and monitoring practices must uphold privacy rights and be subjected to rigorous human rights impact assessments. This is especially important given that racially minoritised women, LGBTQ+ communities, and those from low-income backgrounds, are more likely to experience both digital harms and disproportionate surveillance. • **Recognise and address the tensions between safety and privacy rights**. While automated monitoring tools may help flag harmful content, their deployment should always be accompanied by transparency, accountability, and strict data minimisation practices. Users – and their responsible adults - must have access to clear information about how their data is used in safety interventions and retain control over their personal information. This needs to be understandable to young people independently, and to adults responsible for caring for children.
Question 6: Do you agree that our draft Guidance is likely to have positive effects on opportunities to use Welsh and treating Welsh no less favourably than English? If you disagree, please explain why, including how you consider the draft Guidance could be revised to have positive effects or more positive effects, or no adverse effects or fewer adverse effects on opportunities to use Welsh and treating Welsh no less favourably than English.	Confidential? – N We agree the guidance supports Welsh language rights. However, we suggest Ofcom strengthen its commitment to linguistic and cultural inclusion through the following: • **Consider funding outreach and translation partnerships** with Welsh-speaking women's groups to ensure that safety guidance, reporting tools, and online resources are fully accessible in Welsh. This will help reach communities who may otherwise be underserved by digital safety initiatives. • **Include case studies from Welsh digital contexts** to ensure cultural and linguistic relevance. Highlighting specific examples of online harms and safety interventions within Wales would provide practical, relatable insights for service providers and demonstrate Ofcom's commitment to addressing regional diversity. • **Ensure online safety resources and reporting mechanisms are available in Welsh by default**, particularly for platforms operating in Wales. This should include not only translations of policies but also culturally relevant messaging and the option for users to report harms and seek support services in their preferred language. • **Engage directly with Welsh language advocacy organisations and women's groups** to co-design solutions that reflect the lived experiences of Welsh-speaking communities. This collaboration would help identify any unique barriers faced by these groups and ensure that their needs are meaningfully integrated into platform safety strategies.

Please complete this form in full and return to OS-Section54@ofcom.org.uk.