

Consultation response form

Introduction:

The below response is on behalf of Meta only.

In relation to WhatsApp LLC: WhatsApp does not have specific comments relating to the questions below but would like to note the following:

For private messaging services that do not host publicly visible content, the safety challenges are distinct. Protecting users, including children, is shaped primarily by how the product itself is designed, whether that is limiting unwanted contact or making it simple to block and report. In this context, product design plays as important a role as content policies in ensuring a safe experience.

Question	Your response
The incidence of content that is harmful to children on regulated services We invite stakeholders to provide evidence relating to the incidence of PPC, PC or NDC on regulated user-to-user, search or combined services. Please provide as much detail as you can. This could include but is not limited to evidence about:	
The quantity, prevalence or presence of specific kind(s) of PPC, PC or NDC on regulated services and the frequency with which children are encountering this content.	Non-confidential <u>Combined response for all of Section A:</u> Meta has globally applicable content policies in place which outline what is and is not allowed on our platforms. These are called the Community Standards. They cover a wide range of issues, for example : • Child Sexual Exploitation, Abuse and Nudity; • Suicide, Self-injury and Eating Disorders;

Question	Your response
	● Adult Sexual Solicitation and Sexually Explicit Language; ● Bullying and Harassment; ● Hateful Conduct; ● Violent and Graphic Content; and ● Violence and Incitement. We have nearly 3 billion users on each of Facebook and Instagram alone, and ~47-48 million on each in the UK. We remove ~265 million pieces of violating content per month, most of it proactively. We receive ~640 million user reports per month, and ~20 million in the UK. We measure the prevalence of violating content to gauge how we're performing against that goal. Prevalence considers all of the views of content on Facebook or Instagram and measures the estimated percentage of those views that were of violating content. <u>Transparency Reports</u>: Over time, we have developed measurements, metrics and data for the purposes of enforcing our policies. A milestone was the publication of our first <i>Community Standards Enforcement Report</i> in 2018 and we continue to publish this every quarter in order to track our progress and demonstrate our continued commitment to making Facebook and Instagram safe and inclusive, see here: https://transparency.meta.com/reports/community-standards-enforcement/ For more information on how Meta calculates prevalence, please see here: https://transparency.meta.com/policies/improving-prevalence-metric/
How demographic characteristics such as age, gender, race, sexual orientation and disability influence the likelihood that children will encounter specific kind(s) of PPC, PC or NDC.	Please see our combined response above, which is non-confidential.
The services or types of service where children are most likely to encounter specific kind(s) of PPC, PC or NDC.	Please see our combined response above, which is non-confidential.

Question	Your response
The severity of harm that UK children suffer, or may suffer, as a result of content that is harmful to children We invite stakeholders to provide any evidence regarding the severity of harm that UK children suffer, or may suffer, as a result of encountering PPC, PC or NDC. Please provide as much detail as you can. This could include but is not limited to evidence about:	
The impact that specific kind(s) of PPC, PC or NDC can have on a child's wellbeing, behaviours, or other physical or psychological outcomes. This might include evidence concerning cumulative exposure to specific kind(s) or combinations of PPC, PC or NDC.	Confidential
Any causal links between a child encountering specific kind(s) of PPC, PC or NDC and subsequently suffering physical or psychological harm.	
How demographic characteristics such as age, gender, race, sexual orientation and disability can influence the severity of harm that UK children suffer, or may suffer, as a result of encountering specific kind(s) of PPC, PC or NDC.	
Whether it is appropriate to make changes to the kinds of PPC and PC in the Act	

Question	Your response
We invite stakeholders to provide evidence that suggests any kind(s) of PPC and/or PC in the Act are inappropriate or incorrectly categorised. This could include but is not limited to evidence indicating that:	
Any kind of PPC or PC in the Act should cease to be categorised as content harmful to children.	Non- Confidential <u>Combined response for all of Section C:</u> <u>Overview of Meta’s Content Policy Development Process</u> Our policies, including the Community Standards, define what is and isn’t allowed on Meta’s platforms. There are a number of reasons why Meta may draft a new policy or revise an existing one. Sometimes we identify a policy gap, discourse evolves in a new direction, or an external stakeholder helps us understand that a policy does not fully address an important issue. In other cases, our policies must evolve to meet new risks, such as a new type of scam or pattern of behaviour. No matter the impetus, at Meta we develop policies to create a better experience for our users. The paramount goal of our Community Standards is to give people a voice. We want to ensure that our policies are not too restrictive, allow people to express themselves freely, and can be enforced fairly and consistently. We also look to international human rights standards. Over the years, our teams have gathered information through both internal and external engagement involving difficult policy issues, which by their very nature tend to evolve over time. Sometimes we are able to update our policies based on our existing body of knowledge gleaned from speech trends on our platforms. However, speech is constantly evolving, and when we explore a significant change to an existing policy or the creation of a new one, we may kick off an extensive policy development process culminating in a Policy Forum (see below). Throughout the process, we consider new inputs from external experts on the topic, relevant internal and external research, and considerations from internal teams, who can help inform how a change might impact our ability to review content accurately. Dialogue is at the heart of how we develop our policies at Meta. While it is our responsibility to write the policies that govern our

Question	Your response
	platforms, we elicit a range of voices in and out of Meta to bring in the perspective of the people who use, or may be potentially impacted by, the policies that govern our products. This includes eliciting input from external stakeholders with a range of perspectives and expertise. We may also consider empirical evidence to guide our decisions. This typically includes original public opinion surveys conducted in countries around the world, the analysis of internal data about what people are posting and how different policies could affect the conversation on our platform, and reviewing the best available published research —from both our internal user research teams and external experts. After we begin a policy development process, conduct engagement, research, and analysis, we present our most significant policy decisions at a meeting called the Policy Forum. The Policy Forum brings together subject matter experts from across the company to discuss potential policy updates. We also at times invite external partners such as academics, experts, practitioners, or Oversight Board members and staff to observe. This engagement allows external partners a chance to better understand our policy development process, holding us accountable to making our policies and processes better. Stakeholder engagement helps to make our policies more nuanced and better attuned to local context. It introduces us to new perspectives, allows us to share our thinking on policy options and roots our policies in sources of knowledge and experience beyond Meta. <u>Perspectives on categories of PPC and PC</u> We broadly agree with the PPC and PC categories, however we have a concern that the level of detail set out in the categories as provided for in the Guidance may be too prescriptive. If categories are defined in too much detail, it may limit our ability to respond agilely to emerging harms and mitigations as they develop. This is explained in more detail below. This area is fast moving and ever evolving. Content moderation and harmful content areas are not stagnant, we are constantly iterating on policies internally (e.g. Policy Forum)

Question	Your response
	and we think the guidance should offer more flexibility and be futureproof. For example, AI generated content or manipulated content is a highly fluid space. We develop internal policies to enforce but it would be difficult to maintain detailed Code measure guidance which may not keep pace with how these particular trends evolve. While we are mostly in agreement with the PPC/PC categories, we have some feedback on certain aspects of Ofcom's Guidance on Content Harmful to Children. Overly detailed definitions: <u>Self-Injury & Recovery Content</u>: Ofcom's guidance provides granular examples of what constitutes harmful content, including narratives describing self-injury and imagery of healed scars in a self-injury context. The guidance suggests that even recovery content can be harmful to vulnerable children. This means Ofcom's threshold could require removal of content that is intended to provide hope or reduce stigma, which is not inherently harmful and may be beneficial for some users. ● <i>Example</i>: Narratives about self-injury, even without graphic imagery, are considered violating by Ofcom. This could lead to removal of personal stories or support content that is unobjectionable and valuable for recovery communities. ● In our external engagements on Disordered Eating and Self Injury, stakeholders, including health and wellness creators and people with lived experience, emphasised the importance of distinguishing between content that promotes self-injury/disordered eating and content that discusses recovery. There was broad agreement that recovery content such as sharing personal journeys of healing or support resources should be allowed, provided it is not graphic or promotional of harmful behaviours. This is why Meta marks as sensitive or age gates this type of content rather than removing it. The stakeholders also broadly noted that survivors and those in recovery have the right to freedom of expression. Prohibiting self-expression about one's body or recovery

Question	Your response
	journey, in the absence of promotion of dangerous behaviours, can be seen as disproportionate and potentially counterproductive. There was also concern about the risk of over-enforcement, for example, removing content that factually discusses self-injury or recovery, which could harm those seeking support or community. <u>Eating Disorder Content:</u> The guidance requires removal of content that "glamourises, glorifies, romanticises or normalises eating disorders," including images of emaciation with aspirational commentary. Meta distinguishes between content that promotes or facilitates eating disorder behaviours and content that discusses eating disorders in a recovery or non-promotive way. Where content includes indicators that it is encouraging, normalising, or instructing eating disorder behaviour, we remove it given the elevated risk of harm, particularly for vulnerable users to prevent exposure that promotes or depicts eating disorders. We also recognise that some content referencing eating disorders, such as recovery narratives, admissions, or educational discussion can be protected expression and may provide support or awareness benefits. However, it can still be sensitive or triggering, especially for minors. In these cases we apply strong viewer protections, including age gating and sensitivity screens, to limit under 18 exposure while preserving space for lawful and socially valuable expression. We are concerned that an approach which treats all eating disorder-adjacent content as requiring removal, without accounting for context and intent, could lead to over-removal of non-harmful material such as recovery stories or educational material. This would reduce access to supportive resources and undermine legitimate expression. Lack of future-proofing • <u>Static Policy Examples:</u> The guidance relies on specific examples and context, such as "body checking" or "side-by-side comparisons" for weight loss. Ofcom's static definitions may not adapt to changing social norms, technology, or new forms of content, leading to over-removal or under-protection as the online environment evolves. • <u>Recovery and Support Content:</u> The guidance's lack of nuance in distinguishing between harmful and supportive

Question	Your response
	content (e.g., recovery narratives, educational material) means it is not flexible enough to accommodate future understandings of mental health, body image, or social dialogue. Potential for removal of unobjectionable content Ofcom's guidance would require content to be removed that is not inherently harmful and may be unobjectionable to the majority of users, such as: ● Healed scars in a recovery context: Our research and stakeholder engagement shows that recovery communities can reduce isolation, which is one of the strongest drivers of relapse and crisis escalation. ● Content about body ideals in a non-eating disorder context: Not all content is disordered; blanket suppression can stigmatise normal health, exercise and aesthetic expression. People can benefit from encouragement around strength, mobility, or recovery from illness or injury. ● Meta's approach is informed by extensive external engagement with over 70 experts worldwide, including clinicians, academics, NGOs, creators, and people with lived experience. These stakeholders consistently emphasized the importance of distinguishing between content that promotes or glamorises eating disorders/self-injury (which should be removed) and content that discusses recovery, support, or education (which should be allowed with safeguards). ● There is broad agreement that recovery and support content can play a vital role in reducing stigma, providing hope, and supporting individuals in their recovery journey. Over-removal of such content risks silencing those seeking help and undermining access to supportive resources. Hate speech: Ofcom's hate speech guidance, while rooted in the intention to protect users, is excessively restrictive and risks eroding the principles of open discourse and democratic debate that underpin

Question	Your response
	both UK and international human rights frameworks. Meta’s approach is grounded in the belief that protecting people from harm must be balanced with safeguarding free expression. In revising our former Hate Speech policy - now the Hateful Conduct policy - we focused on covering the highest-risk content, simplifying our rules, and enabling meaningful discussion on issues of public importance. This means reaffirming our commitment to free expression, even when it results in content that some may find offensive. Such speech is often part of robust public debate whether in Parliament, the media, or our communities. Removing it from our platforms would diminish users’ ability to engage in conversations that matter to their daily lives. Our Hateful Conduct policy is carefully crafted to navigate the inherent complexity of hate speech. There is no universally accepted definition of ‘hate speech’ under international law. Our Hateful Conduct policy seeks to strike an appropriate balance between freedom of expression and the right to non-discrimination in line with international human rights standards. As noted in Ofcom’s guidance on Content Harmful to Children, providers should have particular regard to users’ rights of freedom of expression under Article 10 of the European Convention on Human Rights and the Human Rights Act 1998, which according to <i>Handyside v UK App no 5493/72 (ECHR, 7 December 1976)</i>, includes the right to “offend, shock, or disturb”. To ensure our policies are both effective and principled, Meta consults global experts with whom we collaborated extensively to refine our approach. Our policies are under constant review, shaped by feedback from internal and external stakeholders, and frequently debated in political and media circles worldwide. Recent reviews have shown that enforcement can sometimes inadvertently or disproportionately restrict legitimate political debate. Our Hateful Conduct policy will continue to protect a broad range of characteristics including race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease. We are also maintaining policies against dehumanising speech and other forms of hateful conduct such as harmful stereotypes, serious insults, and others, and will continue to assess the impact of these changes on content affecting young people. Other policies remain applicable to address similar types of potential harmful content, such as our Violence and Incitement,

Question	Your response
	Bullying & Harassment and our Coordinated Harm policies. Concerns with Ofcom's Approach Ofcom's examples of "abuse and hate content" are overly broad, encompassing language and opinions that, while offensive to some, are integral to ongoing societal debates about gender, identity, and social inclusion. For example, the guidance includes: • Arguments for gender-based limitations in certain professions or spaces: this prohibits content that is a common debate among sports clubs, competitions, at the dinner table and in parliament. For example "Trans people should not be allowed to play on women's rugby teams - it's not safe", "trans men should not be allowed in women-only rape and domestic violence shelters", "Trans men should not be allowed to share bathrooms with little girls at the swimming pool". Meta allows this content because we want to respect rights to freedom of expression, information and opinion for speech that does not amount to advocacy of incitement to hatred or discrimination. However, we remove content that states for example, [REDACTED] • Expressions of disgust in personal relationship contexts: This type of content is often part of legitimate social and cultural conversations. Meta believes that not all negative speech is inherently harmful or abusive. Meta's policies distinguish between content that is simply offensive or unpleasant and content that is hateful. Expressions of disgust about a relationship (e.g., "I can't stand my ex") are not the same as hate speech targeting a protected characteristic, [REDACTED] • Denial of the existence of certain protected characteristics: Some claims are embedded in political debates about law, categories, eligibility, definitions, etc. A blanket ban risks becoming overbroad and disproportionate, stifling important political and social debates Meta allows certain content framed as an opinion that does not directly target a person, for example, "I don't believe gender identity exists; I believe there are only two sexes". But we don't allow content advocating exclusion or violence such as [REDACTED]

Question	Your response
	• Statements of superiority between groups: This is overly broad and includes speech that can be non-hateful, non-protected, or non targeted. For example “Mexicans are better than Asians”. Some of this content may be offensive but not a direct attack on a protected group. However, content that aims to dehumanise certain groups or incite hatred is removed under our policies such as “Look at this German guy! They’re all trash”. • Through Meta's external engagement on hate speech and online harms, stakeholders have highlighted that: - Overly prescriptive definitions can lead to the removal of content that is important for key societal debates, awareness-raising, advocacy, counter-speech, or support—such as discussions about discrimination or sharing of personal experiences. - Freedom of expression is a core value, and policies should not conflate offense with harm. Several civil society partners and subject matter experts cautioned that removing content simply because it is controversial or challenges prevailing norms can stifle necessary societal debate. - International human rights standards should guide policy, with a focus on necessity and proportionality, as reflected in the ICCPR and ECHR frameworks. While controversial, such content is debated in Parliament, the media and society at large. The guidance conflates causing offence with causing harm, mandating removal of content that merely expresses attitudes rather than focusing on the prevention of real-world harm. Suppressing these forms of speech would not only restrict users’ ability to share their views on topics relevant to their daily lives, but also stifle opinions that, while controversial, are actively debated in Parliament, the media, and society at large. Meta’s approach ensures that our platforms remain spaces for open dialogue, while robustly protecting against hateful conduct targeting groups based on their protected characteristics. Meta’s Hateful Conduct policy is designed to strike a careful balance between preventing harm and preserving free expression, in line with international human rights standards. Ofcom’s

Question	Your response
	guidance, by contrast, risks silencing legitimate debate and undermining democratic values. Our ongoing commitment is to maintain policies that protect users from harm without sacrificing the open exchange of ideas that is essential to a democratic society. Because Ofcom has strong enforcement powers (including major fines), platforms may take a cautious approach and remove or restrict borderline lawful speech to reduce regulatory risk particularly due to difficulties in enforcement of policy at scale. This could create a chilling effect on speech, especially for controversial political or societal debates.
Any kind of PPC would be more appropriately categorised as PC.	Please see our combined response above, which is non-confidential.
Any kind of PC would be more appropriately categorised as PPC.	Please see our combined response above, which is non-confidential.
We invite stakeholders to provide evidence in support of additional kind(s) of content being added to the PPC or PC categories in the Act. This should include, but is not limited to:	
Evidence that the kind(s) of content pose a material risk to an appreciable number of children.	Non-confidential <u>Combined response for all of Section D:</u> Meta collects evidence of content posing a risk to children as part of our OSA Children's Risk Assessments.
Evidence that the kind(s) of content pose a risk of significant harm to children.	Please see our combined response above, which is non-confidential.
Evidence to support each kind of content being categorised as either PPC or PC.	Please see our combined response above, which is non-confidential.

Please tell us how you came across about this call for evidence.

- ☐ Email from Ofcom
- ☐ Saw it on social media

- ☐ Found it on Ofcom's website
- ☐ Found it on another website
- ☐ Heard about it on TV or radio
- ☐ Read about it in a newspaper or magazine
- ☐ Heard about it at an event
- ☐ Somebody told me or shared it with me
- ☐ Other (please specify)

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