pint

# Consultation response form

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

https://www.ofcom.org.uk/online-safety/protecting-children/protecting-children-from-harms-online/

| | |
|---|---|
| **Consultation title** | Consultation: Protecting children from harms online |
| **Organisation name** | Samurai Labs |

# Your response

| Question | Your response |
|---|---|
| **Volume 2: Identifying the services children are using**<br>**Children's Access Assessments (Section 4).**<br>https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/284469-consultation-protecting-children-from-harms-online/associated-documents/vol2-identifying-services-children-are-using.pdf?v=336051 | |
| **Do you agree with our proposals in relation to children's access assessments, in particular the aspects below. Please provide evidence to support your view.**<br>1. Our proposal that service providers should only conclude that children are not normally able to access a service where they are using highly effective age assurance?<br>2. Our proposed approach to the child user condition, including our proposed interpretation of "significant number of users who are children" and the factors that service providers consider in assessing whether the child user condition is met?<br>3. Our proposed approach to the process for children's access assessments? | Confidential? – Y / N |
| **Volume 3: The causes and impacts of online harm to children**<br>**Draft Children's Register of Risk (Section 7)**<br>https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/284469-consultation-protecting-children-from-harms-online/associated-documents/vol3-causes-impacts-of-harms-to-children.pdf?v=336052 | |
| **Proposed approach:**<br>4. Do you have any views on Ofcom's assessment of the causes and impacts | Confidential? – Y / N |

| Question | Your response |
|---|---|
| of online harms? Please provide evidence to support your answer.<br>a. Do you think we have missed anything important in our analysis?<br>5. Do you have any views about our interpretation of the links between risk factors and different kinds of content harmful to children? Please provide evidence to support your answer.<br>6. Do you have any views on the age groups we recommended for assessing risk by age? Please provide evidence to support your answer.<br>7. Do you have any views on our interpretation of non-designated content or our approach to identifying non-designated content? Please provide evidence to support your answer.<br>**Evidence gathering for future work:**<br>8. Do you have any evidence relating to kinds of content that increase the risk of harm from Primary Priority, Priority or Non-designated Content, when viewed in combination (to be considered as part of cumulative harm)?<br>9. Have you identified risks to children from GenAI content or applications on U2U or Search services?<br>a) Please Provide any information about any risks identified<br>10. Do you have any specific evidence relevant to our assessment of body image content and depressive content as kinds of non-designated content? Specifically, we are interested in: | When it comes to suicide and self-harm content, there is some evidence suggesting that generative AI can encourage or promote it:<br>● A man died by suicide after conversations with generative AI, which provided arguments for sacrificing oneself for the planet: Euronews.<br>● GenAI Replika has been documented to mitigate suicides among students (Nature), although there are reports of users stating that it actually encouraged them to do so: Reddit 1 and Reddit 2.<br>● Another evidence of potentially dangerous outputs related to suicide and self-harm is associated with Google AI: Rolling Stone.<br>Additionally, regarding violent content and other types of risk, Humane Intelligence conducted a large-scale red teaming study, resulting in a report on the percentages of successful reports for various categories of harmful outputs: Humane Intelligence Report. |

| Question | Your response |
|---|---|
| a) (i) specific examples of body image or depressive content linked to significant harms to children,<br>b. (ii) evidence distinguishing body image or depressive content from existing categories of priority or primary priority content.<br>11. Do you propose any other category of content that could meet the definition of NDC under the Act at this stage? Please provide evidence to support your answer. | |
| **Draft Guidance on Content Harmful to Children (Section 8)** | |
| 12. Do you agree with our proposed approach, including the level of specificity of examples given and the proposal to include contextual information for services to consider?<br>13. Do you have further evidence that can support the guidance provided on different kinds of content harmful to children?<br>14. For each of the harms discussed, are there additional categories of content that Ofcom<br>a) should consider to be harmful or<br>b) consider not to be harmful or<br>c) where our current proposals should be reconsidered? | Confidential? – Y / N<br>Yes, I agree with the proposed approach and would like to provide further evidence regarding the prevalence of various types of suicide and self-harm encouragement content on the social platform Pinterest, and how such content can proliferate and reach the most vulnerable users: Google Document.<br>Additionally, it is not uncommon for both youth and adults to express their suicidal intentions, including sharing their definitive plans, on user-to-user online platforms. In our work, we have encountered various forms of such cries for help, ranging from situations where individuals leave a suicide letter within the community just before an attempt, to instances where people search for a suicide buddy (suicide pacts). Such comments and posts can potentially negatively affect vulnerable users or lead to potential suicide pacts being arranged by users in crisis.<br>I believe that companies should not only implement measures to detect such content but also provide users in distress with relevant help while making efforts to prevent tragedies.<br>On Reddit, we have encountered content where users shared their stories of suicidal crises, including descriptions of lethal substances and their dosages, |

| Question | Your response |
|---|---|
|  | implicitly encouraging others to follow a similar path. Often, such comments were written in response to a post from a user struggling with suicidal thoughts. These comments could pose a risk to the most vulnerable users and should potentially be taken down. At the same time, the users posting these comments should be provided with relevant help. |
| **Volume 4: How should services assess the risk of online harms?**<br>**Governance and Accountability (Section 11)** |  |
| 15. Do you agree with the proposed governance measures to be included in the Children’s Safety Codes?<br>a) Please confirm which proposed measure your views relate to and explain your views and provide any arguments and supporting evidence.<br>b) If you responded to our Illegal Harms Consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response.<br>16. Do you agree with our assumption that the proposed governance measures for Children's Safety Codes could be implemented through the same process as the equivalent draft Illegal Content Codes? | Confidential? – Y / N |
| **Children’s Risk Assessment Guidance and Children’s Risk Profiles’ (Section 12)** |  |

| Question | Your response |
|---|---|
| 17. What do you think about our proposals in relation to the Children’s Risk Assessment Guidance?<br>a) Please provide underlying arguments and evidence of efficacy or risks that support your view.<br>18. What do you think about our proposals in relation to the Children’s Risk Profiles for Content Harmful to Children?<br>a) Please provide underlying arguments and evidence of efficacy or risks that support your view.<br>Specifically, we welcome evidence from regulated services on the following:<br>19. Do you think the four-step risk assessment process and the Children’s Risk Profiles are useful models to help services understand the risks that their services pose to children and comply with their child risk assessment obligations under the Act?<br>20. Are there any specific aspects of the children’s risk assessment duties that you consider need additional guidance beyond what we have proposed in our draft?<br>21. Are the Children’s Risk Profiles sufficiently clear and do you think the information provided on risk factors will help you understand the risks on your service?<br>a) If you have comments or input related to the links between different kinds of content harmful to children and risk factors, please refer to Volume 3: Causes and Impacts of Harms to Children Online which | Confidential? – Y / N |

| Question | Your response |
|---|---|
| includes the draft Children’s Register of Risks. | |
| **Volume 5 – What should services do to mitigate the risk of online harms**<br>**Our proposals for the Children’s Safety Codes (Section 13)** | |
| **Proposed measures**<br>22. Do you agree with our proposed package of measures for the first Children’s Safety Codes?<br>a) If not, please explain why.<br>**Evidence gathering for future work.**<br>23. Do you currently employ measures or have additional evidence in the areas we have set out for future consideration?<br>a) If so, please provide evidence of the impact, effectiveness and cost of such measures, including any results from trialling or testing of measures.<br>24. Are there other areas in which we should consider potential future measures for the Children’s Safety Codes?<br>a) If so, please explain why and provide supporting evidence. | Confidential? – Y / N<br>With One Life Project, we are employing Artificial Intelligence to identify signs of suicidal crises within user comments, including descriptions of suicidal thoughts, declarations, plans, and methods, as well as suicide and self-harm encouragement. Last year, we supported over 25,000 people on Reddit, where AI detected individuals in suicidal crises (more than 250 people daily). Human experts then reached out to those in distress with supportive interventions, providing them with user support materials such as Find a Helpline (a global network of verified helplines) and self-help resources. Additionally, a team of suicidology experts addressed barriers to mental health help-seeking.<br>This proactive approach enabled us to reach individuals in distress, including children, on a much larger scale. Traditionally, platforms rely on user reporting, where a person encountering a distressed post can report it, prompting an automated message with a helpline number. However, various studies indicate that the majority of relevant content may not be reported. Moreover, in some posts, individuals describe being in the midst of an active suicide attempt. In such cases, emergency services should be contacted immediately.<br>Out of the 25,000 people to whom we provided interventions, 4 required active rescues. Emergency services were called for 3 teenagers under 18, all resulting in lives being saved. The fact that suicide notes or cries for help are posted online creates a unique |

| Question | Your response |
|---|---|
| | opportunity to proactively detect and assist these individuals using AI before it's too late.<br>Posts from users in suicidal crises should receive special protection, and potential encouragement must be detected within these contexts, as sometimes a single comment can prompt an attempt. However, we cannot simply delete the comment of someone in crisis who may unintentionally encourage others. Such users should receive an intervention and be encouraged to seek help. |

| | |
|---|---|
| **Developing the Children’s Safety Codes: Our framework (Section 14)** | |
| 25. Do you agree with our approach to developing the proposed measures for the<br>Children’s Safety Codes?<br>a) If not, please explain why.<br>26. Do you agree with our approach and proposed changes to the draft Illegal Content Codes to further protect children and accommodate for potential synergies in how systems and processes manage both content harmful to children and illegal content?<br>a) Please explain your views.<br>27. Do you agree that most measures should apply to services that are either large services or smaller services that present a medium or high level of risk to children?<br>28. Do you agree with our definition of ‘large’ and with how we apply this in our recommendations?<br>29. Do you agree with our definition of ‘multi-risk’ and with how we apply this in our recommendations?<br>30. Do you agree with the proposed measures that we recommend for all services, even those that are small and low-risk? | Confidential? – Y / N |
| **Age assurance measures (Section 15)** | |

| | |
|---|---|
| 31. Do you agree with our proposal to recommend the use of highly effective age assurance to support Measures AA1-6? Please provide any information or evidence to support your views.<br>a) Are there any cases in which HEAA may not be appropriate and proportionate?<br>b) In this case, are there alternative approaches to age assurance which would be better suited?<br>32. Do you agree with the scope of the services captured by AA1-6?<br>33. Do you have any information or evidence on different ways that services could use highly effective age assurance to meet the outcome that children are prevented from encountering identified PPC, or protected from encountering identified PC under Measures AA3 and AA4, respectively?<br>34. Do you have any comments on our assessment of the implications of the proposed Measures AA1-6 on children, adults or services?<br>a) Please provide any supporting information or evidence in support of your views.<br>35. Do you have any information or evidence on other ways that services could consider different age groups when using age assurance to protect children in age groups judged to be at risk of harm from encountering PC? | Confidential? – Y / N |
| **Content moderation U2U (Section 16)** | |

| | |
|---|---|
| 36. Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.<br>37. Do you agree with the proposed addition of Measure 4G to the Illegal Content Codes?<br>a) Please provide any arguments and supporting evidence. | Confidential? – Y / N |
| **Search moderation (Section 17)** | |
| 38. Do you agree with our proposals? Please provide the underlying arguments and evidence that support your views.<br>39. Are there additional steps that services take to protect children from the harms set out in the Act?<br>a) If so, how effective are they?<br>40. Regarding Measure SM2, do you agree that it is proportionate to preclude users believed to be a child from turning the safe search settings off?<br>The use of Generative AI (GenAI), see Introduction to Volume 5, to facilitate search is an emerging development, which may include where search services have integrated GenAI into their functionalities, as well as where standalone GenAI services perform search functions. There is currently limited evidence on how the use of GenAI in search services may affect the implementation of the safety measures as set out in this code. We welcome further evidence from stakeholders on the following questions and please provider | Confidential? – Y / N |

| | |
|---|---|
| arguments and evidence to support your views:<br>41. Do you consider that it is technically feasible to apply the proposed code measures in respect of GenAI functionalities which are likely to perform or be integrated into search functions?<br>42. What additional search moderation measures might be applicable where GenAI performs or is integrated into search functions? | |

| | |
|---|---|
| **User reporting and complaints (Section 18)** | |
| 43. Do you agree with the proposed user reporting measures to be included in the draft Children's Safety Codes?<br>a) Please confirm which proposed measure your views relate to and explain your views and provide any arguments and supporting evidence.<br>b) If you responded to our Illegal Harms Consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response.<br>44. Do you agree with our proposals to apply each of Measures UR2 (e) and UR3 (b) to all services likely to be accessed by children for all types of complaints?<br>a) Please confirm which proposed measure your views relate to and explain your views and provide any arguments and supporting evidence.<br>b) If you responded to our Illegal Harms Consultation and this is relevant to your response here, please | Confidential? – Y / N |

| | |
|---|---|
| signpost to the relevant parts of your prior response.<br>45. Do you agree with the inclusion of the proposed changes to Measures UR2 and UR3 in the Illegal Content Codes (Measures 5B and 5C)?<br>a) Please provide any arguments and supporting evidence. | |

| | |
|---|---|
| **Terms of service and publicly available statements (Section 19)** | |
| 46. Do you agree with the proposed Terms of Service / Publicly Available Statements measures to be included in the Children's Safety Codes?<br>a) Please confirm which proposed measures your views relate to and provide any arguments and supporting evidence.<br>b) If you responded to our illegal harms consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response.<br>47. Can you identify any further characteristics that may improve the clarity and accessibility of terms and statements for children?<br>48. Do you agree with the proposed addition of Measure 6AA to the Illegal Content Codes?<br>a) Please provide any arguments and supporting evidence. | Confidential? – Y / N |
| **Recommender systems (Section 20)** | |
| 49. Do you agree with the proposed recommender systems measures to be included in the Children's Safety Codes?<br>a) Please confirm which proposed measure your views relate to and provide any arguments and supporting evidence.<br>b) If you responded to our illegal harms consultation and this is relevant to your response here, please signpost | Confidential? – Y / N |

| | |
|---|---|
| to the relevant parts of your prior response.<br>50. Are there any intervention points in the design of recommender systems that we have not considered here that could effectively prevent children from being recommended primary priority content and protect children from encountering priority and non-designated content?<br>51. Is there any evidence that suggests recommender systems are a risk factor associated with bullying? If so, please provide this in response to Measures RS2 and RS3 proposed in this chapter.<br>52. We plan to include in our RS2 and RS3, that services limit the prominence of content that we are proposing to be classified as non-designated content (NDC), namely depressive content and body image content. This is subject to our consultation on the classification of these content categories as NDC. Do you agree with this proposal? Please provide the underlying arguments and evidence of the relevance of this content to Measures RS2 and RS3.<br>• Please provide the underlying arguments and evidence of the relevance of this content to Measures RS2 and RS3. | |
| **User support (Section 21)** | |
| 53. Do you agree with the proposed user support measures to be included in the Children's Safety Codes?<br>a) Please confirm which proposed measure your views relate to and | Confidential? – Y / N |

| | |
|---|---|
| provide any arguments and supporting evidence.<br>b) If you responded to our Illegal harms consultation and this is relevant to your response here, please signpost to the relevant parts of your prior response. | |
| **Search features, functionalities and user support (Section 22)** | |
| 54. Do you agree with our proposals? Please provide underlying arguments and evidence to support your views.<br>55. Do you have additional evidence relating to children's use of search services and the impact of search functionalities on children's behaviour?<br>56. Are there additional steps that you take to protect children from harms as set out in the Act?<br>a) If so, how effective are they?<br>As referenced in the Overview of Codes, Section 13 and Section 17, the use of GenAI to facilitate search is an emerging development and there is currently limited evidence on how the use of GenAI in search services may affect the implementation of the safety measures as set out in this section. We welcome further evidence from stakeholders on the following questions and please provide arguments and evidence to support your views:<br>57. Do you consider that it is technically feasible to apply the proposed codes measures in respect of GenAI functionalities which are likely to perform or be integrated into search functions? Please provide | Confidential? – Y / N |

| | |
|---|---|
| arguments and evidence to support your views. | |

| | |
|---|---|
| **Combined Impact Assessment (Section 23)** | |
| 58. Do you agree that our package of proposed measures is proportionate, taking into account the impact on children's safety online as well as the implications on different kinds of services? | Confidential? – Y / N |
| **Statutory tests (Section 24)** | |
| 59. Do you agree that our proposals, in particular our proposed recommendations for the draft Children's Safety Codes, are appropriate in the light of the matters to which we must have regard?<br>a) If not, please explain why. | Confidential? – Y / N |
| **Annexes**<br>**Impact Assessments (Annex A14)** | |
| 60. In relation to our equality impact assessment, do you agree that some of our proposals would have a positive impact on certain groups?<br>61. In relation to our Welsh language assessment, do you agree that our proposals are likely to have positive, or more positive impacts on opportunities to use Welsh and treating Welsh no less favourably than English?<br>a) If you disagree, please explain why, including how you consider these proposals could be revised to have positive effects or more positive effects, or no adverse effects or fewer adverse effects on opportunities to | Confidential? – Y / N |

| | |
|---|---|
| use Welsh and treating Welsh no less favourably than English. | |

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