Your response

Please refer to the sub-questions or prompts in the annex to our call for evidence.

Question	Your response
Question 1: Please provide a description introducing your organisation, service or interest in Online Safety.	*Is this response confidential? – N* Carnegie UK's objective is better wellbeing for people in the UK and Ireland. Over the past three years, we have shaped the debate in the UK on reduction of online harm through the development of, and advocacy for, a proposal to introduce a statutory duty of care to reduce Online Harms. Our proposal is for social media companies to design and run safer systems – not for government to regulate individual pieces of content. Companies should take reasonable steps to prevent reasonably foreseeable harms that occur through the operation of their services (for example, the impact of recommender systems), enforced by a regulator. The proposal has been developed by Professor Lorna Woods (Professor of Internet Law, University of Essex), William Perrin (Carnegie UK Trustee) and (Carnegie UK Associate). It draws on well-established legal concepts to set out a statutory duty of care backed by an independent regulator, with measuring, reporting and transparency obligations on the companies. Our way of working is to develop and publish detailed public policy proposals, drawing on our extensive legal, regulatory and policymaking expertise, for debate and adoption by others. For example, we published a draft Online

	Harms Bill to demonstrate that a systems-based regime is easy to legislate for. Over the past 18 months, we have carried out work – in conjunction with a number of civil society organisations, academics and other expert groups – to develop principle-based model codes of practice that act at a systemic level to help tech companies assess and reduce the prevalence of online harm on their services. The work started with a code of practice on hate speech, which then informed ad hoc advice for the UN Special Rapporteur on Minority Issues. We then adapted this approach to produce – through the same collaborative process – a code of practice on online violence against women and girls and are currently working with another civil society partner on a code on mis- and disinformation. We will refer to these examples throughout the rest of our submission, setting out both the principles that we feel online services should follow in addressing the particular functionality or design choice and, where relevant, extracting the subject-specific application of that principle from one or other or the published codes.
Question 2: Can you provide any evidence relating to the presence or quantity of illegal content on user-to-user and search services? **IMPORTANT: Under this question, we are not seeking links to or copies/screenshots of content that is illegal to hold, such as child sexual abuse. Deliberately viewing such images may be a criminal offence and will be reported to the police.**	*Is this response confidential? – N* We are not an expert in this regard and would refer Ofcom to the work of, for example, Internet Watch Foundation and NSPCC with regard to child sexual abuse and exploitation material, or to the Institute of Strategic Dialogue, Centre for Countering Digital Hate or Hope Not Hate on extremism.
Question 3: How do you currently assess the risk of harm to individuals in the UK from illegal content presented by your service?	*N/A*

Question 4: What are your governance, accountability and decision-making structures for user and platform safety?	*N/A*
Question 5: What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements?	*Is this response confidential? - N* There are a few more general points to make upfront before we turn to the detail in our proposed codes of practice. Firstly, consideration needs to be given for what terms of service more generally should cover: our work on codes of practice has not touched on this because they have been specifically focused on a type of content which, by definition, will already be "in" the terms of service. There may be some issues that need to be considered across all services – and this will be implicit form the illegal content safety duty and the children's safety duty) and others that may be flagged up by the risk assessment. The service provider should look to its risk assessment to identify the minimum that should be in the terms of service. There is also a consideration as to whether it is appropriate to bundle the community guidelines in terms of service and/or privacy policies. We can see arguments both ways: for a user, looking at three separate documents is not effective if you are trying to understand what's ok and not ok online. But conversely, terms of service could be written in legal language whereas community standards might more appropriately been written in less formal language - and made prominent more easily. Our proposed codes dealt with the issue of clarity and accessibility by suggesting that a service provider should make its terms of service (including any privacy policy) and/or community standards visible to would-be

	users and advertisers before they sign up to the service. The terms of service and/or community standards must be expressed in clear and easy to understand language. This includes providing different language versions of the terms of service and/or community standards appropriate to the territories in which the service is made available. It must ensure that training and awareness tools are readily available to users on the Terms of Service and Community Guidelines to ensure users are aware of permitted content and behaviours on the platforms. These policies should be kept under review, updated when appropriate and users informed of any such changes. In addition, we recommend that a service provider must prompt its users to consider their safety and privacy settings and that these features should be designed appropriately in the light of the risks present on the service. To show how this works in practice, with regard to a particular type of harm, we set out below the relevant section from our code of practice on violence against women and girls (p15) *Terms of Service constitute the contract between the service provider and the user. They are important in communicating the service provider's values. As such, they may include community standards (though sometimes Terms of Service and Community Standards are used interchangeably) or acceptable use policies, understood as the content and behaviour rules the provider will enforce. The Community Standards should make clear the service provider's position on VAWG. This is not the same as saying, however, that platforms must actively seek*

	out criminal content, or monitor generally. Such general monitoring has adverse impacts for all users' freedom of expression and privacy and would be very difficult, if not impossible, to justify. There is a need to ensure that the Terms of Service are not rendered meaningless and that there is some mechanism that is proportionate and appropriate to ensure that they provide a realistic expectation for the user of the types of content and behaviour that they will and will not encounter on the service. *Terms of service should be easily visible before a user signs up to the service, be easy to understand (by the age groups using the service) and be available in languages used by the service's users. This is important as part of transparency, but also to hold service providers and users to account. Terms of Service and Community Guidelines should be kept under review, and revised where appropriate taking into account not just changes in external context but also learning from risk assessments, metrics on effectiveness of mitigation plans and complaints and moderation processes as well as any codes and or guidance from OFCOM. For regulated services to effectively address the risk of VAWG, their terms of service must explicitly state what activity and material they determine constitutes VAWG and how they will deal with it. Most importantly, services must then enforce these principles and ensure the Terms of Service are effective and operational.* *Terms of service must reflect the harms that occur to women and*

	girls, ensuring systems and processes are continually informed by victims' perspectives and safeguarding best practice. This information might, for example, come from the internal expertise within the company or third-sector partners who provide advisory input. The provider must also explain how terms are developed, enforced, and reviewed, and the role of victims' groups and civil society in developing them. The Terms of Service must explain the steps that regulated services will take if the terms of service are broken by users and be enforced by the online service. Evidence must be kept on individual cases, in line with GDPR requirements, regardless of the final decision. Within the service itself providers must ensure that training and awareness tools are readily available to users on the Terms of Service and Community Guidelines to ensure users are aware of permitted content and behaviours on the platforms, and that these tools are kept updated.
Question 6: How do your terms of service or public policy statements treat illegal content? How are these terms of service maintained and how much resource is dedicated to this?	*N/A*
Question 7: What can providers of online services do to enhance the transparency, accessibility, ease of use and users' awareness of their reporting and complaints mechanisms?	*Is this response confidential? – N* We refer back to the principles-based approach set out in answer to question 5. On the specific functionality and design choices here, we suggest the following: <u>User empowerment tool</u>s Our codes envisaged that a service provider must consider what tools, in addition to content and behaviour reporting tools, are necessary to allow users to improve their

	control of their online interactions and to improve their safety; this is now reflected in the terms of clause 14 of the OSB. Such tools could include: a) controls over recommendation tools, so a user can choose, for example, to reject personalisation; b) user-set filters (over words, images or topics); c) tools to limit who can get in touch/follow a user, or to see a user's posts; d) tools to allow users to block or mute users, or categories of user in advance (eg anonymous accounts); e) Controls for the user over who can and cannot redistribute their content or user name/identity in real time. In addition, a service provider must ensure that these tools are easy to use by all groups of users accessing the service (for example considering the age of users) and take reasonable steps to ensure their prominence such that users are aware they exist. Reporting and complaints: A service provider must have reporting processes that are fit for purpose, that are clear, visible and easy to use and age-appropriate in design and cover all content and behaviour (whether user-generated, service generated (eg autocompletes) or advertising-based). A service provider must consider whether some forms of complaint (eg harassment; image-based sexual abuse) need specially designed reporting processes. We also recommend that a service provider must provide the opportunity for non-users who are affected by content or behaviour on the service to report that content and/or behaviour; and that providers should record the complaints in a sufficiently granular manner to feed into risk assessment review processes. The typology of categorisations

should be developed with survivor representatives.
To show how this works in practice, with regard to a particular type of harm, we set out below the relevant section from our code of practice on violence against women and girls (p24)

> *Complaints processes provide vital early warning of VAWG problems on a service, as well as a mechanism to deal with a problem in an individual case. The adequacy of complaints processes should be part of the risk assessment. The provider should also ensure that the design of complaints mechanisms is user-centric: that is, visible, easy to use and age and language appropriate. Complaints processes should not just be limited to complaints about individual items of content. They should allow for complaints about a series or pattern of communications as well as to features of the services itself (for example, the way the recommender algorithm works, or other 'dark patterns' and nudges, or tools for creation). The regulator must regularly assess whether such processes are fit for purpose. Regulated services must work to identify trends and developments in user reporting and incorporate this in any transparency reporting obligations to the regulator.*
> *Good practice in responding to VAWG content that is flagged to an online service might include the following:*
>
> *• all platforms must acknowledge reports within 24 hours. Reports must be actioned within a specific time frame set and published by the provider in their Terms of Service*

	and in response to a report made (this may vary dependent on harm reported); *• data should be gathered on response times to ensure these commitments are met;* *• companies should track where multiple reports are made by an individual as this may indicate increased risk of harm;* *• victims must be able to provide the username of the perpetrator, rather than reporting individual pieces of content;* *• reporting avenues should be provided for non-users to flag harmful content;* *• users should have access to clear flagging processes that identify whether their issues are VAWG related as well intersecting with other types of abuse such as racist, homophobic abuse. This is in addition to more specific flagging categories to triage and escalate risk;* *• consideration must be given to the accessibility of flagging and reporting for younger users who may not be conscious of VAWG dynamics impacting their case;* *• regulated services must use the intelligence from the report or flag to prioritise its human and automated content moderation;* *• in the case where content, which has had a determination by automated technology, is continuing to be flagged or reported, it must be*

	assessed by a human moderator; *• there must be an appropriate number of VAWG-trained human moderators, taking into account the scale of any VAWG problem on the service;* *• human moderators must be supported in a holistic manner which recognises the psychological impact of the work;* *• harmful content or actions which have been flagged as having gendered nature must be expedited and considered by moderators with VAWG and child protection expertise;* *• regulated services must explain the outcome of a report or flag in clear and simple language, outline a user's right to appeal and explain the steps a user must take if they do not agree with the determination; and* *• recommender algorithms must consider content that has been recently flagged or reported and limit its spread until the content has been reviewed.*
Question 8: If your service has *reporting or flagging* mechanisms in place for illegal content, or users who post illegal content, how are these processes designed and maintained?	*N/A*
Question 9: If your service has a *complaints* mechanism in place, how are these processes designed and maintained?	*N/A*

Question 10: What action does your service take in response to *reports* or *complaints*?	*N/A*
Question 11: Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what?	*Is this response confidential? – N* In terms of user rights, we are very much of the view that interventions that have an effect before take down are more proportionate, as per the views of the UN Special Rapporteur on Freedom of Expression (A/HRC/38/35). With regard to content moderation, we would suggest the following approach should be followed: 1. The service provider's policies must be effectively and consistently enforced. A service provider must have in place expanded guidance explaining their terms of service/privacy policies/community standards (and how these are developed, enforced and reviewed, plus the role of relevant survivors' groups and civil society in developing them). Such further guidance must be in accordance with national law and international human rights. 2. A service provider must have in place sufficient numbers of moderators, proportionate to the service provider size and growth and to the risk of harm who are appropriately trained to review harmful and illegal content and who are themselves appropriately supported and safeguarded. 3. Where automated tools are used, a service provider must put in place processes to ensure those tools operate in a non-discriminatory manner and that they are designed in such a way that their decisions

	are explainable and auditable. Users should be informed of the use of such tools. Machine learning and artificial intelligence tools cannot wholly replace human review and oversight. (eg see OSCE policy manual on AI and freedom of expression.) 4. A service provider must establish clear timeframes or other benchmarks for action against non-compliant content. 5. Action in relation to a complaint must be proportionate to the severity of the harm likely to be caused; illegal content is to be dealt with swiftly. The terms of service must make clearly the nature of any such action and the circumstances in which it would arise, as well as details of any appeals process. Action could include: a) Label as inaccurate/misleading; b) demonetise content; c) Suppress content in recommender tools; d) Geo-blocking of content; e) Suspension of content; f) Removal of content; g) The existence of a strike system, if a strike system is in place; h) Geo-blocking of account; i) Suspension of account; j) Termination of account. 6. A service provider must have systems of assessment and feedback to the initial reporter and the owner of content that has been flagged and actioned to ensure transparency of decision making. Users should be kept up to date with the progress of their reports

	and receive clear explanations of decisions taken. 7. A service provider must put in place a right of appeal on all decisions made concerning illegal or harmful content, or content that has been flagged as illegal or harmful content. All users must be given a right to appeal any measures taken against them, whether in full or in part. Users must be able to present information to advocate their position. 8. A service provider must have appeals systems which must take no longer than seven days to assess appeals, except in exceptional circumstances which are unforeseeable and beyond the provider's control. 9. A social media provider must consider putting in place an appropriate trusted flagger programme, with due regard to the subject-specific qualifications that would equip them for the job, that maintains independence from the service provider and from governments. A service provider must: a) ensure trusted flaggers are not used as a sole provider of flagging content; b) ensure trusted flaggers are appropriately compensated, while not compromising their independence c) hold regular meetings with members of the trusted flagger programmes to review content decisions and discuss any concerns;

	d) provide support to trusted flaggers who are exposed to harmful content in line with the service provider's support to its own moderation teams.
Question 12: What automated moderation systems do you have in place around illegal content?	*N/A*
Question 13: How do you use human moderators to identify and assess illegal content?	*N/A*
Question 14: How are sanctions or restrictions around access (including to both the service and to particular content) applied by providers of online services?	*Is this response confidential? – N* We would suggest that this question is approached firstly from the perspective of safety-by-design principles and then with a focus on a principle-based approach to creation of content, both of which help introduce some important safeguards and protections for users further upstream of decisions on restrictions to access and sanctions. Ex-post interventions, especially those relying on 'bolt on' safety tech, are one part of the picture but they should not replace safety by design which allows a broader range of interventions that are potentially less problematic from a FoX perspective We set out our recommended approach to both below: <u>Safety by Design</u>: Bearing in mind the outcomes of the risk assessment, service providers should implement appropriate technical and organisational measures to embed safety by design in the running and the development of service and its features

	and to drive ongoing improvement. Safety by design does not mean the elimination of all risks but rather to inculcate an approach where appropriate choices about understanding, minimising or allocating risk can be made. We further recommend that a service provider must take steps to ensure that the design process takes into account the different characteristics of users, aiming to design inclusively; and that the provider should review, consulting with external experts where necessary, and where appropriate revise those technical and organisational measures in the light of that review. As part of its risk assessment and mitigation processes, the service provider should carry out or arrange for the carrying out of such testing and examination of its service and business systems (including any advertising systems) to assess the safety of the service by reference to the harms caused in the relevant content domain. This testing should include systems and tools for recommendation, content curation and moderation, especially automated tools. <u>Creation of content</u>: A service provider must consider the appropriate levels of friction in the content-posting process in the light of its risk assessment – for example prompts about language used; number of posts permitted over a given period. A service provider must also consider whether any monetisation or revenue-sharing arrangements with content providers provide incentives for or provide financial support to harmful content, and take appropriate steps to mitigate any such risk. We also recommend that a service provider should risk assess the tools for the creation of content – this includes but is not limited

	to bots (including chatbots), bot networks, deepfake or audiovisual manipulation materials, the ability to embed content from other platforms and synthetic features such as GIFs, emojis and hashtags. A service provider must also have terms of service and/or community standards in respect of its users that are fit for purpose taken against its values, local laws and international human rights. The provider should also undertake regular systemic reviews of its terms of service and/or community standards to ensure that they remain up-to-date, effective and proportionate, and amend them when appropriate, for example to take account of findings from risk assessments.
Question 15: In what instances is illegal content removed from your service?	*N/A*
Question 16: Do you use other tools to reduce the visibility and impact of illegal content?	*N/A*
Question 17: What other sanctions or disincentives do you employ against users who post illegal content?	*N/A*
Question 18: Are there any functionalities or design features which evidence suggests can effectively prevent harm, and could or should be deployed more widely by industry?	*Is this response confidential? –N* We would say upfront here that tools/functions/features are not automatically good or bad - eg banning anonymity. There are also specific interventions that may be relevant for particular types of harms: eg service providers might consider whether

	nudification apps have any legitimate purpose and consider how their use and development is generally design to affect and demean women. In 2021, Carnegie UK undertook some work in collaboration with Prof Ellen Goodman from Rutgers Institute in New York to look at the role of algorithmic auditing as a means to identify and prevent harm online. We convened a workshop with academics, researchers, regulators (including a representative from Ofcom) and civil society representatives to consider the components of such an auditing approach and came up with the modular approach to auditing, as set out below. • Input and output data: data fields that will aid in understanding the type of information that is submitted to and produced by the software (eg explanation of table ids or developer documentation) • Documentation about model development and structure • Pre-implementation self or independent audit • Post-implementation self or independent audit • Training materials • Implementation: automated and human-mediated decisions connected to the algorithmic system We would be happy to share our background materials from this workshop, including the reference paper and the minutes of the meeting, with Ofcom if helpful.
Question 19: To what extent does your service encompass functionalities or features designed to mitigate the risk or impact of harm from illegal content?	*N/A*

Question 20: How do you support the safety and wellbeing of your users as regards illegal content?	*N/A*
Question 21: How do you mitigate any risks posed by the design of algorithms that support the function of your service (e.g. search engines, or social and content recommender systems), with reference to illegal content specifically?	*N/A*
Question 22: What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them?	We would refer Ofcom to the extensive work undertaken by 5 Rights Foundation here.
Question 23: Can you identify factors which might indicate that a service is likely to attract child users?	*N/A*
Question 24: Does your service use any age assurance or age verification tools or related technologies to verify or estimate the age of users?	*N/A*
Question 25: If it is not possible for children to access your service, or a part of it, how do you ensure this?	*N/A*
Question 26: What information do you have about the age of your users?	*N/A*

Question 27: For purposes of transparency, what type of information is useful/not useful? Why?	*Is this response confidential? – N* We would refer Ofcom to the evidence submitted to the Online Safety Bill Public Bill Committee from Reset, which focuses on transparency which considers the different approaches taken internationally with regard to transparency (https://committees.parliament.uk/writtenevidence/39851/pdf/) In addition, we provide the following extract from our VAWG code of practice which sets out how this should work in relation to that particular type of online harm. (p31) *Transparency reporting and information release must contain three main elements:* *• collaboration and information sharing with relevant regulators;* *• collaboration and information sharing with relevant civil society bodies that support the prevention and mitigation of VAWG; and* *• public data sharing in line with transparency guidelines that is accessible and easily digestible for all service users. Clear transparency allows civil society and the public to monitor online services progress in tackling gender-based harms and hold online services to account. There is a public benefit to transparency concerning online safety. Transparency enables society to monitor the progress of the sector. It also builds confidence in the industry.* *Online services are strongly encouraged to collaborate with experts on VAWG topics and achieve better outcomes for their users. Online services that effectively collaborate with other platforms will be able to consider gender-based harms in the round and tackle issues before they appear on a platform. It is recommended that a UKCIS working group on VAWG is established which could bring regulated services, the regulator, VAWG sector and*

	government together and be used as a means of sharing reports and data.
Question 28: Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content?	*N/A*

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