Call for evidence response form

Your response

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

Question	Your response
Question 1: Please provide a description introducing your organisation, service or interest in Online Safety.	BSR is a non-profit organization working with companies to create a just and sustainable world. With offices in Asia, Europe, and North America, BSR provides over 300 member companies with insight, advice, and collaborative initiatives. Among other activities, BSR partners with technology companies (including internet companies) on human rights due diligence, including stakeholder engagement, human rights assessments, and advice on reporting and disclosure. We have also written reports about Online Safety, such as Human Rights Based Approach to Content Governance. Our response to this consultation is informed by our practical experience working on human rights due diligence with around a dozen companies and organizations relevant for the UK Online Safety Bill.
Question 2: Can you provide any evidence relating to the presence or quantity of illegal content on user-to-user and search services?	Not applicable to BSR
Question 3: How do you currently assess the risk of harm to individuals in the UK from illegal content presented by your service?	BSR undertakes human rights impact assessments for internet companies that, among other things, identify actual and potential adverse human rights impacts. BSR's primary reference point are the International Bill of Human Rights (Universal Declaration of Human Rights; International

	Covenant on Civil and Political Rights; International Covenant on Economic, Social and Cultural Rights) and other relevant international human rights instruments. In practice this means identifying specific ways in which a company's products services could be connected to adverse impacts on the full range of internationally recognized human rights. In this context, our focus is online content that may be associated with adverse human rights impacts, regardless of whether the content is legal or illegal. BSR typically uses engagement with affected stakeholders as the primary pathway towards the identification of content that may have an adverse impact on human rights, though we review other sources too, such as content moderation case data. Further, we prioritize adverse human rights impacts using criteria based on the UN Guiding Principles on Business and Human Rights (UNGPs), namely "scope" (the number of people impacted), "scale" (the gravity of the impact), and "remediability" (whether a remedy will restore a victim). We also consider "likelihood" (the chance of a harm occurring). BSR considers the concept of vulnerability when assessing human rights impacts, defined as those that face being marginalized, discriminated against, or exposed to other adverse human rights impacts with greater severity and/or lesser potential for remediation. Vulnerability depends on context, and someone who may be powerful in one context may be vulnerable in another. At a conceptual level we find the following four dimensions of vulnerability to be helpful: • Formal Discrimination—laws or policies that favour one group over another. • Societal Discrimination—cultural or social practices that marginalize some and favour others. • Practical Discrimination—marginalization due to life circumstances, such as poverty.

	• Hidden Groups—people who might need to remain hidden and consequently may not speak up for their rights, such as undocumented migrants. Examples of vulnerable groups frequently include children, women, indigenous peoples, ethnic minorities, LGBTQI+ people, or persons with disabilities, though vulnerability depends on context, and someone who may be powerful in one context may be vulnerable in another. Here are two key resources: • Human Rights Assessment sets out our generic approach to human rights assessment across all industries • Human Rights-based Approach to Content Governance sets out our approach to human rights in the content governance field. In BSR's work we emphasize the importance of context (e.g., existence of conflict; language; culture; politics; literacy, etc) in shaping the impact of internet companies. We especially emphasize the importance of conflict-affected contexts, where companies should undertake "heightened" due diligence. See: • Business in Conflict-Affected and High-Risk Contexts
Question 4: What are your governance, accountability and decision-making structures for user and platform safety?	The outputs of BSR human rights impact assessments are typically considered by a formal or informal cross-functional group within the company, often including functions such as human rights, content policy, stakeholder engagement, legal, government affairs, and public policy, but also including others when relevant, such as product teams, research, or country representatives. Going forward we anticipate that our human rights impact assessments, which are largely focused on specific products/services, markets, or issues areas, will inform company-wide "human rights salience assessments". These assessments may be raised at Board level, or

	input into “enterprise risk management” and “compliance” processes.
Question 5: What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements?	Important factors arising in BSR human rights impact assessments include language (i.e., having policies available in multiple relevant languages), specificity (i.e., providing sufficient detail for terms to be understood, such as accompanying implementation guidance), and accessibility (e.g., visuals and summary versions; versions targeted at younger users).
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?	BSR’s Human Rights Impact Assessment for the Global Internet Forum to Counter Terrorism (GIFCT) reviews how companies can better fulfil their legal duty to remove terrorist content from platforms in a manner that respects human rights. Of particular relevance is (1) the consideration of definitions of terrorism and violent extremism and (2) anti-Islamic bias that exists in the counterterrorism field, (p32 – p36).
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?	BSR believes that online services should undertake a gap analysis between their reporting and complaints mechanisms and the effectiveness criteria for non-judicial grievance mechanisms contained in Principle 31 of the UNGPs (e.g., legitimate, accessible, predictable, equitable, transparent, rights compatible, source of continuous learning, stakeholder engagement). We are not aware of companies having done this systematically to date, though the Facebook Oversight Board Human Rights Review (p51 – p56) provides an example of what this could look like.
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?	Not applicable for BSR

Question 9: If your service has a *complaints* mechanism in place, how are these processes designed and maintained?	Not applicable for BSR
Question 10: What action does your service take in response to *reports* or *complaints*?	Not applicable for BSR
Question 11: Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what?	In BSR's experience the biggest improvements that can be made to content moderation relate to the ability to understand context (e.g., language / dialect, culture, politics, etc) and dedicate sufficient resources (e.g., human reviewers, reliable classifiers) to implement content policy. Generally speaking, the more context that is needed to assess whether a piece of content is harmful, the more challenging it is for companies to have effective moderation at scale. Classifiers tend to struggle with reliability for contextually dependent content, particularly in non-English languages, and therefore more human resources are needed to review flagged content. This can lead to adverse human rights impacts when legal liability risks incentivize companies to over-enforce. This has been demonstrated with terrorist and violent extremist content, which often require context to appropriately assess, and has led to undue restrictions on free expression, access to information, association, and other rights of Muslim and Arabic-speaking communities. See our Human Rights Impact Assessment of GIFCT for more information. Given the need to moderate content at scale, prioritizing both human and engineering resources based on the severity of risks to people (i.e., scope, scale, remediability, likelihood) is essential. We also emphasize the role of effective, meaningful, and mutually beneficial relationships with stakeholders who can provide important context to inform better

	content policy and enforcement, and alert companies to content trends.
Question 12: What automated moderation systems do you have in place around illegal content?	BSR's Human Rights Impact Assessment for the Global Internet Forum to Counter Terrorism (GIFCT) reviews the use of the GIFCT hash sharing data base for content removals from a human rights perspective. We recommend a clear taxonomy for content that qualifies for inclusion in the hash sharing database, oversight mechanisms, human review, not allowing governments to add hashes directly, appeals mechanisms, third party review, researcher access, and multi-stakeholder governance. We make these recommendations to protect the rights to freedom of expression, association, and assembly, and non-discrimination by ensuring hashes added to the database are limited to clearly defined terrorist and violent extremist content and do not result in the removal of borderline and/or legitimate content. We emphasize that ultimate accountability for content removal rests with companies using the hash sharing database, rather than GIFCT. In BSR's Human Rights Impact Assessment of Meta's Expansion of End-to-End Encryption we broadly discuss hash-based systems for automated content moderation. In order to work properly, only clear-cut, definable instances of illegal content can be hashed. This includes, for example, known instances of CSAM or terrorist group manifestos. Hash-based systems are not appropriate for content that requires contextual analysis.
Question 13: How do you use human moderators to identify and assess illegal content?	Based on insights gained during human rights impact assessments, we emphasize the importance of (1) hiring a sufficient number of human moderators with the ability to understand context relevant for the content being reviewed (e.g., language / dialect, culture, politics, etc), (2) investing in the

	capability to scale-up / scale-down on short notice to respond to crisis events that can result in sudden spikes in illegal content, and (3) maintaining quantitative metrics that assess the accuracy of human moderator decisions—not least because these are often used to train machine-based classifiers, and so human moderator errors can be reproduced in classifiers if not addressed.
Question 14: How are sanctions or restrictions around access (including to both the service and to particular content) applied by providers of online services?	Not applicable for BSR
Question 15: In what instances is illegal content removed from your service?	Not applicable for BSR
Question 16: Do you use other tools to reduce the visibility and impact of illegal content?	BSR emphasizes the importance of companies using human rights principles (e.g., necessity, proportionality, non-discrimination) when acting against content where the legal status of that content is unclear, or when it is unclear whether content violates the company's content policy. For this reason, actions to reduce the visibility of content, rather than removing content altogether, can be an appropriate and helpful course of action. However, companies should be cognizant of the human rights risks (e.g., freedom of expression, non-discrimination, democratic participation) when taking this action, and identify scenarios where reducing visibility has adverse impacts on public dialogue.
Question 17: What other sanctions or disincentives do you employ against users who post illegal content?	In BSR human rights assessments we emphasize the importance of enforcement actions (e.g., strikes, and the reduced visibility / functionality often associated with strikes) being proportional to the violation and clearly communicated to users (e.g., category of violation, action taken) so that they can choose whether to appeal.

	In BSR's past work with a variety of online platforms, we have generally found most platforms to lack sufficient explanation to users related to alleged content policy violations and associated enforcement actions. This lack of information makes it challenging for users to understand what they may have done wrong and if they should appeal an enforcement decision.
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?	Not applicable for BSR
Question 19: To what extent does your service encompass functionalities or features designed to mitigate the risk or impact of harm from illegal content?	Not applicable for BSR
Question 20: How do you support the safety and wellbeing of your users as regards illegal content?	Not applicable for BSR
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?	Not applicable for BSR
Question 22: What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them?	In BSR's experience, age assurance / age verification is inherently difficult. Most existing age verification efforts have been found to be minimally effective, while those that are effective in identifying the age of the user often have adverse impacts on user privacy. Experts consulted as part of BSR-led child rights impact assessments have described this challenge as one of the most significant child rights issues for the coming years. Without confirmation of the age of the user, it can be difficult to identify risks to children, take appropriate action, and provide them with a

	safe and age-appropriate experience. Conversely, attempts to verify the age of users, particularly children, come with their own risks, including violations of privacy and inaccuracies based on race, gender, ethnicity, culture, or other factors. Age verification mechanisms are typically deployed as part of a company's broader approach to protecting children's safety and security online. These approaches often address specific child safety risks and fail to consider the full range of child rights. Specifically, age assurance / verification mechanisms may prevent children from engaging with the digital environment anonymously, potentially impacting a range of other rights, including the right to civic participation, access to information, participation in the cultural life of the community, and potentially other rights such as the right to health and education. Companies should assess and take action to address all child rights / human rights impacts (including considerations around child participation, freedom of expression, access to information and culture, etc.) as part of their due diligence processes. There is a need for cross-industry collaboration on rights-based approaches to age assurance that address these issues, as well as equity concerns related to age verification processes.
Question 23: Can you identify factors which might indicate that a service is likely to attract child users?	High quality content that is positive and appropriate for young audiences keeps children engaged on platforms specifically designed for them. It is also key to children having enriching, educational and empowered experiences. Low quality content may hinder learning and encourage negative and even damaging behaviors and attitudes among children. It may also encourage children to leave digital experiences / safe spaces designed specifically for children in search of more interesting, higher-quality content on platforms or services without specific guardrails or protections for children.

	Currently, content quality is inconsistent across geographies and languages, and content may not be developed or designed for the full range of child users, such as children with disabilities, learning disorders, health issues, etc. Companies have expressed difficulties in finding and incentivizing content creators to create high quality children's content, particularly across languages, cultures, geographies, etc. This impacts the diversity, quality, inclusion, and equity of content. Companies and government actors providing support for children's media may need to collaborate to address this. Child users have also indicated greater interest in platforms that allow them to engage with content and their peers (through comments, chat features, etc.).
Question 24: Does your service use any age assurance or age verification tools or related technologies to verify or estimate the age of users?	Not applicable for BSR
Question 25: If it is not possible for children to access your service, or a part of it, how do you ensure this?	BSR assessments have found that even when expressly prohibited, children often find ways of accessing platforms and services of interest. Once on the platform, children may experience difficulties in reporting hateful, harmful, illegal, or otherwise problematic behaviour, including harassment and bullying, sexting, or the sharing or distribution of sexual imagery. They may also have difficulties requesting removal of content they are in, including sexual imagery (self-generated or otherwise). This may be due to a lack of knowledge on where to find reporting channels, practical difficulties they encounter when trying to submit a report, or fear of punitive actions, including being kicked off the platform. Reporting structures need to be accessible and understood by all users, including children. The most effective reporting mechanisms are visible, easily discoverable, recognizable, accessible and available at all times, to all users, with a clear infrastructure and

	established process to ensure speedy review and appropriate action. According to a recent report by Thorn, "children are more than twice as likely to use platform blocking and reporting tools than they are to tell parents and other caregivers about what happened."
Question 26: What information do you have about the age of your users?	Companies should determine actions to address child rights-related risks by context and the age of the child, not children as a generalized category. Products and services that take a blanket approach to content restrictions for all children (e.g., all users under the age of 13, 15, or 18 depending on the country), may limit a child's rights and ability to access information and participate in their community / cultural life and the arts. Similarly, wide age categories may not meet the needs of children in a specific phase of childhood development. Age categories may need to be developed and/or reviewed to ensure that policies and approaches include considerations of children's rights and reflect children's developmental stage at different ages and the specific risks they may face on the service.
Question 27: For purposes of transparency, what type of information is useful/not useful? Why?	Consistent with Principle 21 of the UNGPs, we believe that companies should publish sufficient information for their content moderation approach to be effectively evaluated by stakeholders. In this context, we believe that companies should publish the results of their human rights due diligence, the actions taken (alone and with others) to address adverse human rights impacts, and how they review the effectiveness of their approach. These disclosures are broader than data relating to illegal content, but we believe they provide essential context for company evaluation. BSR believes that the most useful reports are a mix of quantitative data and qualitative analysis, and that today's transparency reports are too skewed towards the former.

	BSR notes: (1) that internet companies are about to be subject to a wide range of transparency requirements relating to content moderation, including the EU Digital Services Act, the EU Corporate Sustainability Due Diligence Directive, the EU Corporate Sustainability Reporting Directive, the EU AI Act, and the UK Online Safety Bill; (2) that international reporting standards, such as the Sustainability Accounting Standards Board / International Sustainability Standards Board (SASB/ISSB) and the Global Reporting Initiative, are also developing standards relevant for content moderation; and (3) voluntary initiatives are underway in the industry, such as the Digital Trust and Safety Partnership and the World Economic Forum's Global Coalition for Digital Safety. For this reason, we emphasize the importance of harmonization, alignment, interoperability, and/or substitutability across these standards and initiatives.
Question 28: Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content?	See answer to question 27.

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