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? – Y / N (delete as appropriate)*
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? – Y / N (delete as appropriate)*
Question 3: How do you currently assess the risk of harm to individuals in the UK from illegal content presented by your service?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 4: What are your governance, accountability and decision-making structures for user and platform safety?	*Is this response confidential? – Y / N (delete as appropriate)*
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? – NO* Specific steps that providers of online services might take to enhance the **clarity** of terms of service and public policy statements include: • Providing users with a high-level summary of terms of service or public pol-

	icy statements (e.g., in the form of simple bullet points), with the option for users to seek more information and detail should they desire; • Providing clear definitions of important terms, such as “hate speech”, “violent content”, “graphic content”, wherever such terms are used, with examples of what is and is not included within the definition. This could involve explaining any thresholds that the online service applies when determining if a piece of content is prohibited, and/or providing examples or additional detail to demonstrate what is meant by each term; • Publishing lists of any organisations or individuals for which content affiliated with or supporting such entities would be in violation of their policies; • Providing information about what enforcement actions the online service may take in the case of each type of content violation and in case of repeat violations; • Informing users clearly of how their data will be used, both for routine use and operation of the online service, including for complaints or appeals that relate to the user or the user’s content; • Explaining clearly whether the company will treat public figures differently when it comes to enforcement of its terms of service and if so, how; • Explaining clearly what exemptions or allowances may be made for violations of the terms of service for journalistic purposes and how such cases are assessed; • Providing users with reasonable notice of any new policy documents or any changes to terms of service before they take effect; and • Requiring explicit acknowledgment of the changes in terms of service by users, beyond simple pop-up banners or windows, which are often ineffective means of relaying information as users

	often ignore or quickly bypass such mechanisms. Specific steps that providers of online services might take to enhance the **accessibility** of terms of service and public policy statements include: • Hosting all terms of service and public policy statements in a centralised location with clear signposting towards different types of documents and information; • Ensuring through interface design that the location of the terms of service is easily accessible, and that users can search for the relevant information within terms of service or public policy documents (e.g., through a help centre or chatbot function); • Using unambiguous and non-technical language for all terms of service and public policy statements that is understandable to the average user; • Using age-appropriate language for terms of service and public policy statements relevant to children using the online service, including graphics, videos, or other creative means of communicating terms of service where appropriate; • Translating the terms of service and public policy statements into all languages in which the online service is used and available, including those spoken by minority groups and immigrant communities; and • Ensuring that terms of service and public policy statements are hosted in a way which is compatible with assistive technologies used by individuals with disabilities, and/or creating audio or visual versions of the documents, as well as working in consultation with those with disabilities to find other effective solutions.

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?	*Is this response confidential? – Y / N (delete as appropriate)*
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? – NO* The providers of online services should make reporting and complaint routes available for both users and non-users, given the fact that harmful content may impact a broad range of individuals that are not users of a particular online service, particularly those which are smaller or medium sized. Any content which is available to or visible by non-registered users should be accompanied by relevant reporting and complaints systems which are also available to registered users. Below we address specific steps on transparency, accessibility, ease of use and awareness for registered and non-registered users. Specific steps that providers of online services might take to enhance the **transparency** of their reporting and complaints mechanisms include: • Explaining clearly to users submitting a complaint what will happen to the complaint at each stage and how long they can expect the process to take; • Notifying the individual or entity responsible for the cause of the complaint that a complaint has been made, and explaining how it will be reviewed and what the potential outcomes will be; • Providing the individual or entity responsible for the cause of the complaint a chance to rebut or provide counter evidence or context; • Providing a clear explanation and justification to all relevant parties for any decision made or action taken in response to the complaint, referring to the specific sections of the terms of service where a violation has been identified;

	• Informing all parties if the review of the complaint or report has been undertaken by an automated tool, and allowing any party to request a human review of the merits of their complaint; • Ensuring that appropriate safeguards and verification measures are in place to protect complaints and appeals systems from misuse or abuse by malicious actors (e.g., in an attempt to censor content that they do not like); and • Explaining clearly how any data or content shared as a result of a complaint will be stored, assessed and deleted. This is particularly important with regards to complaints or appeals over content shared on private or encrypted services, or over complaints relating to certain forms of content such as the non-consensual sharing of intimate images. Online service providers should ensure that rigorous safeguards and protections are in place for user privacy throughout the complaints and appeals process. Specific steps that providers of online services might take to enhance the **accessibility** of their reporting and complaints mechanisms include: • Ensuring through software design that users can easily report or make a complaint about any content that they encounter, in any format, including comments, private messages, multimedia and content shared within closed groups, as well as public posts and public webpages; • Using unambiguous and non-technical language for all reporting and complaints mechanisms, instructions and supplementary information, that is understandable to the average user (and has been tested with users to ensure this is the case); • Translating all reporting and complaints mechanisms, instructions and supplementary information into all languages in which the online service is used and available, including those

	spoken by minority groups and immigrant communities (in consultation with local experts); and • Ensuring that reporting and complaints mechanisms, instructions and supplementary information are hosted in a way which is compatible with assistive technologies used by individuals with disabilities, and/or creating audio or video versions of the documents (working in consultation with those with disabilities to ensure effective solutions). Specific steps that providers of online services might take to enhance the **ease of use** of their reporting and complaints mechanisms include: • Providing users with pre-prepared options or categories for their complaint as well as an open complaint category (in cases where the user is not sure which category to use or feels that no categories are suitable); • Allowing users to provide more detail on the context or substance or their complaint if it is not clear from the original content itself; • Providing confirmation of receipt of the complaint, ideally with a reference number that users can use to follow up easily; • Offering users the option of downloading or having a copy of their complaint sent to them (provided that non-registered users consent to providing relevant contact details); and • Ensuring that appeals mechanisms are designed to be just as clear, accessible, transparent and easy to use as the primary complaints mechanisms, in the ways outlined above. Specific steps that providers of online services might take to enhance **users' awareness** of their reporting and complaints mechanisms include: • Including along with any decision issued clear information about each af-

	fected party's right to appeal the decision, including both internal and external appeals processes; • Regularly (e.g., once per year) reminding users through a pop-up or notice of how to use the reporting and complaints mechanisms (this may only be possible for registered users, and may have to be randomised frequency for non-registered users); and • Where a piece of content has been identified as suspicious, for example, by an automated tool or by viral activity, the online service provider might prompt users about their reporting and complaints mechanisms with regard to that specific piece of content (e.g., "Are you concerned about this content? Report it here."). **Reporting routes for children and adults** Online service providers should recognise that vulnerable users, in particular children, may not be competent or able to make use of the reporting and complaints mechanisms designed for adult users. This may be due to a lack of awareness that particular content is wrong (for example, in the case of child grooming), a lack of knowledge of the reporting and complaints mechanisms (for example, if the child does not know about this feature of the platform), or a lack of understanding of how to use the reporting and complaints mechanism (for example, if the child does not know which category their complaint falls into or does not understand the instructions). In order to help children access and use such mechanisms effectively, providers of online services might consider: • Creating age-appropriate content regarding digital safety for child users to learn from, either upon signing up for a service or regularly (e.g., once per month) during their use of the service, which could be accompanied by games or quizzes for the child to complete which tests their understanding; • Creating more simple and straightforward mechanisms for underage users

	to lodge complaints, including simpler or more clearly explained categories, simpler language, graphics and visuals to aid explanation and instructions; and • Enabling adults to make complaints on behalf of a child under specific circumstances, such as when the adult is a parent or guardian or otherwise responsible for the child, or if the child has given the particular adult permission to make a complaint on their behalf.
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?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 9: If your service has a *complaints* mechanism in place, how are these processes designed and maintained?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 10: What action does your service take in response to *reports* or *complaints*?	*Is this response confidential? – Y / N (delete as appropriate)*
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? – NO* Content moderation measures, when designed and implemented well, can be an effective and proportionate measure to manage and mitigate the risks of harm to individuals and the risks of individuals encountering harmful or illegal content. However, poorly designed, overbroad or biased content moderation measures risk not only restricting user activity, but also limiting users' ability to freely express themselves online, and potentially discriminating against marginalised or minority groups using the service.

	Content moderation may be achieved through a mixture of human content reviewers and automated processes, including hashing systems (most effective at identifying known images of illegal content) and machine learning systems (applied to images, text, videos, user activity data or user metadata with varying degrees of accuracy). Whilst in some limited cases automated tools do have a high degree of accuracy at flagging particular forms problematic content, in general their accuracy is limited. Broad steps that online service providers could take to improve their content moderation systems include: • Using automated systems only to augment or assist, rather than replace, human reviewers, and providing users with the option to request that content moderation undertaken by an automated tool be reviewed by a human; • Where appropriate, making use of techniques other than content removal (such as downgrading, deprioritising, labelling or otherwise modifying the content), still communicating clearly to the user the reasons for any action taken and offering them the opportunity to appeal; • Designing content ranking algorithms – which can be considered a form of content moderation – to promote high-quality and verified content rather than to promote content which has received a high number of engagements or reactions or which is viral due to its controversiality; • Ensuring that any "blacklists" of words that are either flagged by automated tools or used by human content reviewers to determine whether a piece of content is hateful or otherwise prohibited are regularly updated and are context sensitive, working in conjunction with local experts and advisors to devise relevant lists for each language in which the online service is available;

	• Ensuring that users can report or share of incidents of abusive language without such posts being censored; and • Ensuring that any content moderation systems, whether automated or human, which are purchased or contracted by external suppliers, meet the same rigorous standards as those applied to internal content moderation systems through rigorous and regular vetting and assessment processes. With regard to **human content moderators,** specific steps that online service providers could take to improve their content moderation systems include: • Where possible, employing content moderators directly rather than outsourcing to external agencies, in order to ensure consistency between content moderation teams, to ensure appropriate accountability, and to facilitate knowledge sharing; • Ensuring that there is sufficient coverage of human content moderators, both in terms of hours covered by shifts and numbers of employees, to allow moderators sufficient time to review each piece of content; • Ensuring that there is sufficient coverage of human content moderators in each language in which the online service is used and is available, with awareness of the social realities in which the service operates; • Refraining from imposing simplistic quantitative targets on human moderators to meet per day or per week. Such targets prioritise quantity over quality of decisions, overlook the complexity of certain cases, and prevent moderators from researching necessary context or information before making their decisions; • Employing a "tiered" system, whereby less experienced moderators can forward more difficult or nuanced re-

	quests to more experienced moderators without having to make a decision themselves if they are unsure; • Providing extensive and regular training to moderators, on the detail and application of the respective terms of service and ensuring that moderators are aware of any changes made ahead of their implementation; • Providing extensive and regular training to moderators on any relevant laws that will affect their moderation decisions, for example, on the types of high priority illegal content specified in the Online Safety Bill; • Providing extensive and regular training to moderators on how their decisions impact the rights of users; • Providing adequate support – financial, emotional, psychological, and any other form of support required – to moderators, particularly those reviewing highly distressing forms of content. Beyond taking care of the moderators, this support is vital to reduce turnover and burnout in content moderation teams, which limits institutional knowledge and consistency between decisions and lowers the overall accuracy of the content moderation systems; • Regularly reviewing the accuracy and consistency of human moderation teams, taking into account the number of decisions made which were subsequently appealed and overturned and comparing the accuracy of decisions made for different content types and formats. Such reviews should assess, in particular, any impacts of human content moderation decisions on users' right to freedom of expression; • Using the findings of these regular reviews to implement practical changes to human moderation systems, such as mandating regular breaks to aid concentration, providing extra training on content types which are frequently mislabelled, or implementing "shadowing" systems where content reviewers might sit in a different review team or

	with a different review agency for a short-term period in order to knowledge share and ensure consistency; and • Further evaluating the potential of automation bias and how this may impact human moderator decisions, and taking appropriate mitigation efforts. Improvements could also be made to online service providers' **automated content moderation systems**. Improvements that could be made to **hashing systems** include: • Using hashing systems only for content types which are manifestly illegal regardless of content type, such as known child sexual abuse material or terrorist propaganda images; • Ensuring that the databases of known illegal content scanned by the hashing algorithm are either verified by a trustworthy, independent party (such as the Global Internet Forum to Counter Terrorism's hash-sharing database for terrorist content), or are securely maintained by the online service provider itself, subject to regular audits to ensure that all matches generated by the hashing system are for genuinely illegal content; and • Ensuring that hashing systems can still flag known illegal images that have been cosmetically altered, for example by cropping or changing image contrast, through perceptual hashing techniques. Improvements to **machine learning techniques and tools,** such as natural language processing or image recognition software, include: • Using such tools only ever in conjunction with human review processes, given their limitations regarding accuracy and context-sensitivity;

	• Ensuring that each determination or output generated by a machine learning tool is accompanied by a certainty score, for example, determining that a piece of content is hate speech with 76% certainty, and using these certainty scores to inform the course of action taken; • Using datasets of authentic examples which have been labelled by content experts to train any machine learning tools (any datasets which have been automatically augmented should be assessed rigorously for amplification of biases in the original authentic data through the augmentation process); • Using language-specific datasets, rather than translating examples from one language to another, to train any machine learning tools (any datasets which have been translated, whether by hand or automatically, should be sense-checked with language experts to ensure that the examples are still valid and are labelled correctly in the new language); • Establishing minimum thresholds for precision and recall as acceptable for each tool. These may vary according to content type (for example, for image recognition of child pornography, it may be necessary to prioritise high recall (high percentage of actual positives identified) so that the tool can detect all instances of child pornography quickly, even where doing so results in a higher proportion of false positives. For determination of abusive or hateful speech, on the other hand, it may be more appropriate to prioritise high precision (high percentage of correct positive identifications) to ensure that users who are not sharing abusive or hateful speech are not unduly censored by a tool which frequently results in over takedown); • Designing the decisions that an automated tool can take in accordance with its accuracy, certainty and potential risks of erroneous decisions. For example, where a determination relates to a

	high-priority form of illegal content which could cause considerable harm if left online, but the automated tool has a low certainty score, the case should be passed to a human moderator; whereas if a determination relates to a lower priority form of illegal content, but the machine has a high level of certainty, it could result in a warning or re-direction being applied to the content; • Extensively testing any machine learning tools prior to roll-out across a range of real-life scenarios, assessing their performance on both precision and recall as well as any potential risks to users' human rights risks posed by erroneous determinations or decisions by the tool, amending the tool until such risks have been mitigated; and • Regularly reviewing any machine learning tools utilised for content moderation for their performance on precision and recall and any impacts they have had on users' human rights, making practical amendments to the tool or its application wherever necessary.
Question 12: What automated moderation systems do you have in place around illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 13: How do you use human moderators to identify and assess illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*
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? – NO* Sanctions or restrictions around access are applied by providers of online services through various means. These may include restricting accessibility or shareability of content, de-amplifying or deprioritising it in ranking algorithms of content, providing warnings or flags over the

	content, redirecting users away from the content, removing the content entirely, and temporarily or permanently removing a user or entity or removing certain functionalities or services available to them. In some cases, particular content types may be referred to law enforcement. These sanctions may be determined either by a human moderator or human review team, or by an automated tool. These sanctions and enforcement mechanisms may have considerable adverse impacts on users' human rights. Online platforms designing sanctions and enforcement mechanisms should be aware that: • Unwarranted sanctions, whether imposed by an automated tool or by a human moderator, may result in the temporary or permanent disabling or removal of content which is not actually unlawful or against terms of service, which may have a detrimental impact on the ability to impart and receive information of all kinds; • Passing on suspected illegal content to law enforcement poses significant risks to user privacy, and should only be justified where explicitly required by law and in relation to the most serious forms of illegal online content. In each case, the decision should be made by a human reviewer, and the relevant user notified of the action being taken; • Implementing "three-strike rules" or similar means of assessing repeat offenders on the online service before taking action against a particular user requires the retention of user data and violative content shared by the user, as well as data on those who have submitted complaints relating to the user in question. This may require the processing and storing of personal information on individuals wishing to remain anonymous, posing risks for individuals' privacy and personal data; and • Enforcing sanctions inconsistently across different users or groups may

	result in a disproportionate level of removals or deplatforming of particular groups, particularly in cases where the sanctions are erroneous. This may threaten individuals' right to non-discrimination. All of these human rights risks should be carefully assessed in accordance with the potential harms caused by not implementing such sanctions and enforcement policies, in consultation with experts on free expression, privacy and other affected human rights. Wherever an automated tool cannot make a determination with a high degree of certainty, it should be passed on to a human moderator. Similarly, wherever a human moderator is at all uncertain of the correct course of action or how to apply the terms of service in a particular case, there should be the possibility of passing the case on to a more experienced or specialist moderator, to reduce the likelihood of unwarranted sanctions.
Question 15: In what instances is illegal content removed from your service?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 16: Do you use other tools to reduce the visibility and impact of illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 17: What other sanctions or disincentives do you employ against users who post illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*

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? – NO* There are a number of functionalities and design features – whether embedded in the online service or provided through third-party software or middleware – which evidence shows can effectively prevent harm by online service providers. These include: • Deploying counter speech against harmful speech, whether through funding or supporting counter speech projects and initiatives, or through developing automated tools which can generate effective counter speech; • Redirecting users who are searching for or consuming illegal or damaging content, such as terrorist content or child pornography, towards alternative content such as helplines or resources; • Ensuring that private or encrypted services have clear and accessible user complaints mechanisms allowing users to report content shared on the private or encrypted channel that they think is violative of the terms of service. This ensures that online service providers can continue to provide end-to-end encryption, which provides security to online activities and communications and protects data from potential malicious actors – which particularly important for the protection of vulnerable groups, including LGBTQ+ persons, survivors of domestic violence and human rights defenders – while also ensuring that illegal or harmful content is not left unchecked on those channels; • Allowing users to customise their own moderation rules beyond what is prohibited in the terms of service, such as Twitter's Bodyguard tool, which allows users to set their own moderation rules; • Allowing users to block content from particular people or groups or on particular topics, or content from unverified or anonymous accounts, such as Twitter's Block Party tool;

	• Allowing users to limit their own discoverability, or to have invisible or anonymous accounts; • Developing software that helps users to review, document and export repeated instances of illegal or harmful content online, such as Google Jigsaw's Harassment Manager tool; • Allowing users to flag what they believe are underage accounts; • Implementing additional privacy-by-default settings for children's accounts, such as only allowing their content or profile to be visible to or engaged with by their friends or contacts; • Limiting certain functionalities for childrens' accounts, such as disabling search or posting features or implementing additional content moderation systems for adult content; • Developing parental controls to allow adults to have control over what types of content is encountered, particularly for younger or vulnerable children; and • Empowering users to add an age-rating or suggestion to content they create or view, provided such an approach is assessed for potential impacts on individuals' ability to receive and impart information. This approach, currently being tested by TikTok, would restrict children's access to live streamed content and to other content which is labelled 18+.
Question 19: To what extent does your service encompass functionalities or features designed to mitigate the risk or impact of harm from illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 20: How do you support the safety and wellbeing of your users as regards illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*

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?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 22: What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them?	*Is this response confidential? – NO* There are several age assurance and age verification technologies available to online services, including requiring upload or verification of some form of official proof-of-age, such as a photo ID, social vouching (whereby a certain number of adult users must vouch that the individual in question is indeed over 18), or automated age verification technologies, which assess facial features and other cues from a video or photo input to determine the estimated age of the individual. Some services also offer background checks, whereby a users' details are verified against public records, and some online services consider additional data – like numbers contained in birthday messages, or information shared by operating system providers and internet providers – in determining the age of a particular user. At present, to our knowledge, little information is publicly available as to the exact mechanisms of age verification employed by each online platform, and on the degree to which such measures effectively prevent children from accessing harmful content. However, evidence is available as to the potential adverse impacts on individuals' human rights that such mechanisms may pose: • Any mechanisms which require the sharing or upload of official identification documents or of sensitive biometric data pose risks to user privacy. Even where the online service provider does not retain copies of these documents, the risk of malicious actors hacking or otherwise intervening in such data exchanges remains salient, and could result in abuse of personal information. This could also adversely

	affect vulnerable groups, including children; • Any mechanisms which require the sharing or upload of official identification documents or of sensitive biometric data would remove the possibility of individuals being able to use services anonymously, which may be vital for certain vulnerable or persecuted groups to be able to access and share information online without fear of reprisal; • Any mechanisms which require the sharing or upload of an official or up-to-date ID may adversely affect the freedom of expression of some of the most vulnerable users, who may not have access to an ID due to financial limitations, homelessness, or due to being the victim of human trafficking or controlling partnerships; • Any mechanisms which rely on machine learning tools for age estimation will contain a margin for error which, even if small, would adversely impact individuals' right to freedom of expression by preventing them from accessing or sharing information when they should be able to do so; • Any mechanisms which rely on machine learning tools for age estimation or verification may pose risks to individuals' right to non-discrimination, as such tools have been shown to be less accurate for particular racial groups or genders; and • Any age verification systems run the risk of creating a two-tiered internet, as well as serving as a deterrent for many adults accessing legal content. If online service providers are still required to use age verification measures, whether these are designed in-house or outsourced to an external company, the provider should ensure that: • The highest standards of data privacy are in place for users sharing personal IDs or sensitive biometric data, and that no such data is retained longer

	than the period necessary to conduct the age check; • Individuals who do not wish to, or cannot share, a personal ID or biometric data are provided with alternative means of verifying their age, or are provided with alternative means of accessing adult portions of the site; • Users are able to appeal any determinations or estimations of age made by an automated tool, and are provided with alternative means of verifying their age where they claim that the decision of the automated tool is incorrect; and • All age verification measures are assessed for potential impacts on human rights and potential biases, and any such impacts or biases are addressed prior to roll-out. It would also be valuable for online service providers to collect and publish data on exactly how effective age verification measures are at preventing children from encountering harmful or illegal content online, as well as any information on how underage users may be circumventing the age checks to access adult content intentionally.
Question 23: Can you identify factors which might indicate that a service is likely to attract child users?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 24: Does your service use any age assurance or age verification tools or related technologies to verify or estimate the age of users?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 25: If it is not possible for children to access your service, or a part of it, how do you ensure this?	*Is this response confidential? – Y / N (delete as appropriate)*

Question 26: What information do you have about the age of your users?	*Is this response confidential? – Y / N (delete as appropriate)*
Question 27: For purposes of transparency, what type of information is useful/not useful? Why?	*Is this response confidential? – NO* Transparency is an essential means of understanding how online service providers operate, adhere to relevant regulations and safeguard their users and human rights online. However, specific transparency requirements must be approached in a proportionate manner, ensuring that different types and sizes of online services are required to submit appropriate information and to whom. In terms of information which could be provided to **users** to positively affect their safety or behaviours, wherever possible online service providers should: • Explain to users in understandable language how any content ranking algorithm works and what data points it uses in order to recommend content, allowing the user to disable particular data points or to switch to a chronological or non-personalised feed should they choose to; • Make clear to users where content has been prioritised due to paid search or ranking, clearly differentiating this from content which is ranked organically; • Make clear to users where content is produced to advertise or sell a particular product or service, and distinguishing this clearly from organic content; • Explain to users in understandable language why they have been shown a particular advert and what data points have been used to target the user;

	• Explain to users in understandable language how any content moderation decisions are made and what processes are in place to protect them from harm, including how the online service uses automated tools; and
	• Explain to users how their personal data is used across all functions of the service, including any verification technologies or any data shared by web browsers or operating systems.
	
	In terms of other information which could be made public for the purposes of transparency – for example, informing the work of researchers or policymakers on how best to address particular types of content – wherever possible, online service providers should:
	
	• Publish regular and detailed qualitative reports of measures taken to address different categories of online content, changes to terms of service, ranking algorithms or content moderation policies, and any other steps that the online service has taken to improve user safety;
	• Publish regular and detailed quantitative reports on content moderation efforts, broken down at the very least by content category (according to the platform's terms of service and any relevant local laws) and geographic region. These reports should include at least the following information:
	  ○ The number of complaints received and the number of pieces of content flagged by automated tools;
	  ○ The number of complaints or flags acted upon;
	  ○ The number of different responses taken (e.g., takedown, deprioritise, labelling);
	  ○ The number of content moderation decisions appealed by users;
	  ○ The number of content moderation decisions later reversed;

	○ The average time taken to respond to user complaints; ○ The average time taken to identify and remove illegal content; ○ The number of requests received from public bodies, including requests to remove particular pieces of content and to hand over user data for the purposes of investigations; • Publish regular and detailed quantitative reports on revenue generation, including revenue earned from advertising or from sale of user data; and • Ensure that qualitative and quantitative reports are hosted in a central location, and are downloadable. Under the Online Safety Bill, Ofcom would be required to publish annual transparency reports summarising the conclusions and trends from the transparency reports it has received from online service providers, examples of best practice, and any other relevant information. In this report, it would be particularly useful for Ofcom to: • Compare and contrast the different standards or metrics employed by different online service providers in their transparency reporting, and indicate best practice for other online services to follow; • Assess and explain, where possible, potential contextual reasons for particular trends in online service transparency reporting; for example, a spike in complaints about disinformation may be observed during an electoral period. This contextualisation will assist online services to predict – and implement more comprehensive responses to – future online harms within the UK context; • Summarise any penalties imposed by Ofcom on any online service providers during the reporting period and explain how such penalties were determined;

	• Include quantitative data on the number of content removal or content moderation requests made by Ofcom to each online service provider, and the actions taken; and • Summarise the policy changes made by any online service providers in response to penalties or requests from Ofcom. Online service providers may also choose to make public, or even open-source, particular functionalities of their platforms, such as ranking algorithms, content moderation techniques or age verification mechanisms. While transparency *about* such systems and processes is virtually always positive, online service providers must also assess the risks of providing too much information, or of providing source code, where doing so may allow malicious actors to exploit vulnerabilities in the systems or to reverse-engineer e.g. hashing databases to generate illegal content at scale. Such transparency may also allow users to circumvent moderation strategies or age protection mechanisms, or to escape detection while breaking the terms of service. As such, each transparency decision of this type should be assessed on a case-by-case basis, ensuring that any information the publication of which may result in harm is redacted or removed from the disclosure.
Question 28: Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content?	*Is this response confidential? – Y / N (delete as appropriate)*