Your response: Please indicate how much of your response you want to keep confidential. Delete as appropriate.	None
For confidential responses, can Ofcom publish a reference to the contents of your response?	Yes

Your response

Google appreciates the opportunity to provide its response to Ofcom's consultation regarding Technology Notices to deal with terrorism content and/or CSEA content. We take seriously our commitment to keeping our users safe by detecting and actioning CSEA and terrorism content on our platforms. To that end, Google develops and deploys a myriad of content moderation tools that are designed to identify and remove this policy-violative or illegal content. To further this effort not only on our platforms but across the online ecosystem, Google makes available industry-leading content safety tools to qualifying partners, including the Content Safety API and CSAI Match tools on our Child Safety Toolkit website.

Based on the information provided in the consultation and related annexes, as well as Google's extensive technical experience in developing content moderation tools, Google makes the following submission.

Question	Your response

Question 1: Do you have any views on our audit-based assessment, including our proposed principles, objectives, and the scoring system? Please provide evidence to support your response	Confidential? – Y **/ N** Google appreciates the principles-based approach proposed by Ofcom for evaluating CSEA and terrorism detection technologies. While Google does not have specific comments regarding the selection of the four outlined principles (Technical Performance; Fairness; Robustness; Maintainability), we encourage Ofcom to consider: **Alignment with current reporting requirements.** Google believes that in evaluating technologies for accreditation, Ofcom should endeavour to align its review process with principles and metrics that platforms are already required to track and optimise for under other online safety regimes. For example, Articles 15.1(e) and 42.2(c) of the Digital Services Act (DSA) require providers of intermediary services to report their use of automated content moderation mechanisms, indicators of the accuracy, and the possible rate of error of the automated means. **Ensuring sufficient flexibility in evaluation metrics.** When evaluating a technology against the four principles and related objectives, Ofcom should ensure that applicants for accreditation are offered sufficient flexibility to determine the relevant performance metrics for their technology in order to account for the context in which the technology is deployed. While Ofcom should allow for flexibility in evaluating each of the four principles (to reflect both the context the technology is operating in as well as ensuring innovation is not stifled), this may be particularly necessary when evaluating Technical Performance. Technical Performance evaluations should not be limited to submission of a specific set of metrics such as precision, recall, or latency rates, and instead should allow applicants

	to explain how submitted metrics are used to evaluate the technology in different environments and use cases. In our experience, for example, when taking into account YouTube's size and scale, the current optimum technical performance metrics relate to Violative View Rate. YouTube has the ability, for example, to collect and report information about Violative View Rate (i.e., how many views a violating piece of content received before being removed) and the rate of removed content that was first identified by YouTube's automated means (as opposed to human flags). It would be impractical and ineffective for YouTube to measure some "traditional" metrics (e.g., recall or latency) across the entire platform. Moreover, allowing flexibility in the evidence and metrics submitted will minimise unnecessary burden for applicants. Collecting data for producing additional metrics requires significant resources and time. In some cases, producing certain metrics may not even be possible or meaningful, as end-users of certain technologies may choose to opt-out, and there may be inconsistencies in the volume of data points. Producing metrics specifically tied to UK users may also pose practical difficulties. Therefore, applicants should be encouraged to submit meaningful metrics that exist in the organisation. Lastly, Ofcom should account for circumstances where certain principles may not be applicable to specific technologies. For example, when evaluating a hash-matching technology, it may not be possible or useful for a service provider to evaluate and submit evidence regarding the fairness of the technology given that the technology is comparing and matching unique hash values associated with the content.. It is not clear how the suggested objectives, such as "bias

	identification" or "bias mitigation" would apply for hash-matching technologies. Ofcom should build in flexibility in the evaluation process to ensure that important technologies are not unable to become accredited because they cannot demonstrate principles that may be inapplicable or incoherent to the particular technology or use case. **Definition of "Fairness".** Google agrees with Ofcom that CSEA and terrorism detection technologies must demonstrate fairness and limit bias across different groups of people. Bias and fairness are terms that can be difficult to define, however, and Ofcom's evaluation process might benefit from further explanation of the ways in which they will evaluate whether a technology is sufficiently fair and free of bias. As explained in the 'Review into bias in algorithmic decision-making' paper from the UK center for data ethics and innovation (CDEI), fair decision-making can be related to procedural fairness or outcome fairness, and these definitions may be complementary to or conflicting with each other. Allowing applicants to provide context about the ways they view fairness in their technologies will be beneficial for applicants. **The scoring system may benefit from additional tiers.** The proposed scoring system appears to be designed to incentivise extensive reporting, by awarding 5 points for "robust and comprehensive" evidence, 1 point for "limited evidence" and 0 points for "no evidence". The large gap between the top score and the next tier may act as a deterrent for technology service providers for seeking accreditation, as this may inadvertently penalise those who are close to meeting the highest standards but fall short by a small margin. Instead, those services would be considered much

	closer to those services that provide “no evidence” at all. The examples provided in Annex 11 underscore this problem. For example, for the ‘performance metrics’ objective under the Technical Performance principle, many technology providers may likely be able to submit evidence between Level 1 (“*internal test results on limited or insufficiently diverse datasets*”) and Level 2 (“*comprehensive results from large-scale, diverse, and representative datasets, including breakdowns by harmful and non-harmful content (content type, language, and scenario).*”) Ofcom may wish to reconsider the scoring system to provide a more gradual assessment. This could involve introducing additional tiers or adjusting the point distribution to better reflect the evidence provided by technology service providers. A more balanced scoring system may encourage broader participation without the fear of failing accreditation for minor shortcomings.

Question 2: Do you have any views on our proposals for independent performance testing, including the two mechanisms for setting thresholds; the approach to testing technologies in categories against particular metrics; and data considerations? Please provide evidence to support your response.	**In most cases, independent performance assessments would present significant practical concerns and may not be necessary for thorough evaluation.** Google believes that it would be extremely difficult to construct a consistent independent performance assessment process that sufficiently, efficiently, and fairly evaluates different content moderation technologies for terrorism and CSEA content. Given the varying technologies, types and modalities of violating content, mediums, and environments that such content could be hosted on, Google does not believe independent performance assessments should be incorporated in the accreditation process. These concerns are detailed further below. **Benchmarked thresholds may be unworkable given the unique environment each technology will be deployed in.** Benchmarked thresholds seek to evaluate technologies based on the performance of similarly situated technologies. However, comparing CSEA and terrorism detection technologies in this way is likely to be impractical. Benchmarked thresholds often overlook the deployment environments of each technology and how these environments can impact performance. For example, two technologies that identify terrorism content of the same modality (e.g., text or image) may operate in entirely different environments. If one technology is deployed at the server level and the other on-device, benchmarked performance thresholds cannot coherently be used to evaluate the technologies because of the differences that may stem from these deployment environments (e.g., in this hypothetical, server level technology may perform better on 'latency' as it would benefit from more compute resources to classify relevant illegal content, and further, the

	data being tested would vary significantly, as text messages tend to be short and filled with jargon (which may affect accuracy metrics), while social media posts may be longer and use more standard language). **CSEA and terrorism detection technologies are typically designed to detect content unique to their platform.** A similar issue arises from the reality that many CSEA and terrorism detection tools are designed to detect content specific to their platform. For example, YouTube's moderation technologies may be fine-tuned to detect content more often seen on YouTube, e.g., related to real individuals. On the other hand, gaming or live-streaming platforms may use similar technologies but that are fine-tuned for detecting gaming or anime-centric content, which is more prevalent on their platform. Therefore, measuring performance based on a certain benchmark dataset without understanding the nature of the content on a platform makes either benchmarking or prescribed thresholds impractical. **Benchmarked or prescribed thresholds do not account for a tool's enforcement context and structure.** When developing content moderation and detection technologies, providers must consider the needs of their particular enforcement environment and structure. For example, Google deploys multiple layers of automated and human review techniques to keep our platforms safe from terrorism and CSEA content. Understanding the accuracy of any one of these technologies in isolation does not give a view into the efficacy of the tool within the system and so may be misleading, as certain tools are developed to fit a particular piece of the entire enforcement structure.

	There are numerous hurdles to creating and maintaining datasets to evaluate CSEA and terrorism content. Google is not aware of any existing datasets for evaluating CSEA or terrorism detection tools. Google believes that attempting to create and maintain such datasets for independent performance testing will create several issues. 1. **Having a benchmark dataset requires technology providers to share a common definition of "CSEA content" or "terrorism content".** Technology providers currently develop and evaluate their content moderation tools using their own definitions and understanding of illegal or violative content. Given that different providers and jurisdictions have varying definitions for "CSEA content" or "terrorism content," it will be very difficult to compare the effectiveness of different technologies, as those technologies are likely to be designed to identify slightly different forms of content. 2. **Testing a technology on illegal content would require specific controlled environments.** Maintaining or storing a benchmark dataset, and testing technologies against that dataset, poses significant security risks. Independent evaluations of content moderation tools would become a prime target for malicious actors seeking to identify vulnerabilities in platform security systems. To mitigate these security concerns, testing would need to occur in a controlled environment or through another process with significant privacy and security protections, potentially hosted by an organisation like NCMEC. This would pose an additional burden on Ofcom

	to develop (or procure the development of) such specific environments. 3. **Difficulty in data collection.** Annex 13 suggests that Ofcom would collect datasets from a variety of sources by using Ofcom's information gathering powers. However, most service providers do not retain copies or datasets of CSAM on their standard infrastructure. For example, when Google detects any kind of CSAM on our services, actions taken include reporting, preserving the content in secure infrastructure, and deleting the content from standard infrastructure and taking account level enforcement actions where appropriate. In addition, for datasets retained in other countries, there may be legal restrictions on whether or not these can be provided to a UK based organisation. For example, we understand US federal law would prevent this type of dataset being transferred from US based servers to the UK. Given these points, it may be difficult for Ofcom to compile relevant datasets. 4. **Outdated data and threat evaluation.** Even if Ofcom could compile and maintain a functional benchmark dataset, this dataset may quickly become outdated. As Ofcom notes in various guidance documents, users develop various techniques to bypass illegal content detection technologies in real-time, rendering any static dataset out-dated in short order. Platforms take measures to account for this reality. For example, while YouTube's automatic content moderation tools are very effective, YouTube also uses reactive moderation (i.e. moderation in response to user reporting) in tandem with tools to ensure that YouTube captures and removes as much violative content as

	possible. This kind of moderation enables YouTube to identify new trends and emerging harms. Using a defined dataset for evaluation inevitably means that the benchmarked data will not be indicative and reflective of current traffic. Consequently, if Ofcom publishes an evaluation set not derived from real or current traffic, the results may be both gameable and less relevant. **Independent Performance Testing may be feasible in limited contexts, such as when evaluating hash-matching technologies.** While it may be difficult to establish an independent performance evaluation regime for many technologies, Ofcom may wish to consider whether such a system may be feasible for specific technologies. For example, it is easier to conduct performance evaluations and construct related performance thresholds and datasets for hash-matching technologies, which are more readily compared given the 1:1 matching nature of the technology. While independent performance evaluations may not be practical or effective for many moderation technologies, Ofcom may consider establishing additional performance evaluations for some tools and not others, where practical.

Question 3: Do you have any comments on what Ofcom might consider in terms of how long technologies should be accredited for and how often technologies should be given the opportunity to apply for accreditation? Is there any further evidence we should consider?	Ofcom may wish to consider the necessity of re-accreditation, given that the audit-based assessment includes forward-looking objectives such as “ongoing bias assessment,” “proactive risk management,” and "detection and mitigation of threats”. Given these objectives already review whether the applicant’s technology can be robustly maintained, it would be beneficial to clarify the role of re-accreditation in light of these objectives. Providing evidence for re-accreditation is likely to involve significant time and costs for applicants. As explained in our responses to Questions 1 and 2, there are significant costs and operational challenges associated with gathering the necessary data to produce evidence for submission. Ofcom may wish to consider that any re-accreditation regime does not need to follow a one-size-fits-all approach, but instead could be tailored to different types of technologies. Ofcom may want to consider a wide range of factors such as modality and data type. For example, the rate of technological development in generative AI technologies is so rapid that accredited technology is likely to be superseded by new advancements over the proposed four-year re-accreditation period, whereas the rate of advancement in other technological contexts may be slower.

Question 4: Do you have any views on how to turn these proposals into an operational accreditation scheme, including the practicalities of submitting technology for accreditation? Is there any additional evidence that you think we should consider? Please provide any information that may be relevant.	**Submitting technologies require significant confidentiality and security measures.** The accreditation evaluation process would require platforms to provide sensitive and thorough information about the methods and systems they use to keep their platforms safe. If accessed or made publicly available, this information could be used by bad actors to game systems and evade content moderation efforts, obtain commercially sensitive information and trade secrets, and potentially expose user information. Ofcom must consider how they will ensure that this information will be securely stored and protected throughout the entire application and evaluation process**.** **Verification of the appointed third party.** These concerns would be significantly enhanced if Ofcom were to delegate the evaluation process to a third party. The consultation does not explain how the "nominated third party" would be appointed by Ofcom. Given the potential technical and security challenges, Ofcom and the nominated third party should, at a minimum: ● Provide and make available documentation and evidence that the nominated third party is not actively developing, or associated with any developers of, illegal content detection and moderation technology; and ● provide evidence that they are able to appropriately access and make use of data in a manner that effectively safeguards any accessed information and ensures an appropriate level of confidentiality, including through implementing technical controls.

Question 5: Do you have any comments on our draft Technology Notice Guidance?	**<u>Section 2: Introduction</u>** **Unclear timeline for compliance with a Technology Notice.** Section A2.27 and A2.28 of Annex 5 suggest that a Technology Notice will give service providers a "reasonable period" of time to comply with a Technology Notice. Both incorporating external technologies into existing processes or developing or sourcing new technologies may take significant amounts of time for a service. Any timeline for compliance should be discussed and negotiated with service providers in advance to ensure that compliance is feasible. For example, Ofcom could consider adopting an approach similar to that proposed in Section 4.24 and 4.25 of the Information Power Guidance, which provides that Ofcom would first issue a draft notice and allow the relevant stakeholder to provide comments on the practicality of the notice. **<u>Section 3: Assessing whether a Technology Notice is necessary and proportionate</u>** **The guidance is unclear about metrics Ofcom will use to assess whether a Technology Notice is necessary and proportionate.** Section A3.5(d) of Annex 5 reflects the requirement in s124(2) (d) and says that Ofcom will consider the "prevalence of relevant content on the service, and the extent of its dissemination by means of the service." Ofcom should define "prevalence of relevant content", and clarify how it will determine prevalence and the extent of dissemination of such content. Additionally, these metrics may not be the most relevant or impactful metrics to use to determine the effectiveness of a service provider's terrorism or CSEA content moderation infrastructure. Ofcom should clarify that it will work with any service provider that it is investigating under an initial assessment to determine what

	metrics may be most appropriate to determine whether a Technology Notice would be necessary and proportionate to that service provider. **Unclear independent compatibility testing process.** Section A3.14 suggests that Ofcom may require independent compatibility testing of an accredited technology that it is considering imposing on a service, and Section A3.18 says this would generally occur before Ofcom decides whether to issue a Warning Notice. However, the guidance does not clarify under what circumstances such compatibility testing may be required or how Ofcom will determine that it is necessary to conduct such testing in the course of exploring whether a Technology Notice would be necessary and proportionate. Conducting independent compatibility testing would be onerous for service providers, as it would present a series of technical, security, legal, and governance challenges. Given these costs, Ofcom should only be able to require independent compatibility testing after it has determined that issuing a Technology Notice will be necessary and proportionate, otherwise the process risks subjecting service providers to indefinite and unbounded testing prior to any legal conclusion that their content moderation architecture is deficient under the statute. As one example, Section A3.16 suggests that services may have to provide "bespoke datasets representative of content the technology would expect to encounter on the service" for independent compatibility testing. As discussed in response to Question 2, it is very difficult to compile datasets of CSAM or terrorism content, as this comes with significant legal, security, and privacy concerns. Additionally, testing external technologies on a service's internal datasets would

	require significant investments in secure infrastructure for testing, development of an appropriate evaluation protocol, effective legal and governance protections regarding the testing process with third parties, and other resource commitments. **<u>Section 4: Initial Assessment</u>** **Ofcom should be required to engage with service providers during the course of an initial assessment.** Section A4.9 of Annex 5 says that when Ofcom is engaged in an initial assessment of a service provider regarding whether a Technology Notice may be necessary and proportionate, that Ofcom "may...engage with the service provider to give them an opportunity to comment on the issue(s), and to provide information to assist us in determining what action, if any, we should take." Ofcom should engage with service providers at the initial assessment phase during each initial assessment. This will help Ofcom better understand a service provider's content moderation tools from the outset, which will better inform any subsequent investigative steps and give service providers the opportunity to **best respond to Ofcom's concerns.** **<u>Section 5: Next steps and approach to information gathering</u>** **Service providers should be given the chance to review the appointment of a skilled person.** Section A5.12 of the Annex 5 suggests that where Ofcom will seek a skilled person's report, the skilled person will be appointed by Ofcom. Further, Section A5.12 indicates that Ofcom will incorporate guidance from the draft Information Powers Guidance with respect to the process for appointing a skilled person. While Google appreciates that such guidance suggests that a

	skilled person will likely only be appointed if Ofcom is satisfied that the person has appropriate safeguards in place to protect confidential information, it is critical that service providers have the opportunity to review and raise objections about the appointment of the skilled person on the grounds of conflict of interest, confidentiality or security issues. **Undefined payment of a skilled person.** Ofcom's guidance suggests that the service provider has to pay for the skilled person appointed by Ofcom. The payment arrangement should be time bound and governed by a set fee arrangement, or service providers should be allowed to object to fees outside a certain range where it is not proportionate. **Testing conducted by the skilled person should be clarified.** Section A5.13 of Annex 5 says Ofcom "may also request that the skilled person conduct separate testing." The extent of this authority should be clarified. Specifically, Ofcom should seek to clarify the scope of any additional testing and how such testing will occur. Undertaking separate testing would raise the various security and technical concerns that we have raised in our responses to Questions 1, 2 and 4. Ofcom should further explicitly reference the considerations listed in Section 4 of the Information Powers Guidance, in particular the data protection considerations in Section 4.64. In addition, Google would appreciate further clarity and explanation on the scope of assistance to be provided by service providers to skilled persons. Google understands that the service provider is under a "duty to give the skilled person all such assistance as they may reasonably require to prepare the report." However, Google is concerned that this language effectively provides a skilled

	person with an unbounded audit right" over Google's technologies. Section 5.17 of the Information Power Guidance helpfully clarifies that service providers are not required to "provide information subject to legal professional privilege to the skilled person." Google would appreciate additional limitations, guardrails, or explanations on what assistance may need to be provided for the skilled persons' report, including how a service provider may raise objections if they believe certain assistance is not reasonably required in preparing the report. **Section 6: Deciding whether to issue a Technology Notice** **Service Providers should be given full information about a skilled person's report.** At Section A6.5 of Annex 5, Ofcom states that the Warning Notice will contain a "summary" of the skilled person's report. In order to provide fulsome representations in response to a Warning Notice, Ofcom should make the entire skilled person's report and related information available to a service provider. This additional information should include, for example, information that Ofcom disclosed to the skilled person for the preparation of the skilled person's report, the specific requests that Ofcom asked to be included in the skilled person's report, or any evidence relied on by the skilled person in preparing the report. **Section 7: Next steps after issuing a Technology Notice** **Further Technology Notice process should incorporate representations from Service Providers.** Section A7.13 of Annex 5 states that when Ofcom decides to issue a further Technology Notice, Ofcom is not required to "obtain a further skilled person's report or issue a further Warning

	Notice". As drafted, this process would not require Ofcom to provide service providers with an opportunity to submit representations with respect to a new Technology Notice. Google believes that for all Technology Notices, including for further Technology Notices, service providers should have the right to make representations and provide evidence of their compliance to Ofcom.

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