Introduction

Google welcomes the opportunity to engage with Ofcom in relation to the additional duties that will apply to categorised services under the Online Safety Act (the 'OSA').

Categorisation thresholds

Before addressing the substance of the additional requirements applicable to categorised services, we want to outline our views on the recommendations made by Ofcom to the Secretary of State in the advice, published on 25 March 2024, on the threshold conditions for categorisation. As Ofcom notes in the overview of its advice, the additional requirements applicable to categorised services should reflect the nature of such services. However, we are concerned that due to the very broad nature of the threshold conditions recommended by Ofcom, a wide range of services will fall within the categories, particularly Category 1, and the nature of such services does not reflect the additional Category 1 duties (e.g. those services are highly unlikely to host news publisher content, journalistic content or content of democratic importance or content relevant to the user empowerment duties.) Nor do they reflect the policy intent of the categorisation thresholds, which is to focus on services with high reach and that contribute to the quick dissemination of content.

Definition of 'content recommender system'

We note that categorisation of user-to-user services largely depends on whether or not they have a 'content recommender system'. This term is not defined under the OSA and has not been subject to parliamentary debate or public consultation. Ofcom has described this as "*an algorithmic system that, by means of machine learning model or other technique, determines or otherwise affects the way in which content is encountered by users of a service.*" We note that the advice envisaged that 12-16 services across the industry would likely meet the threshold for Category 1. However, in practice, we think that many more services could meet the threshold conditions for Category 1 than we understand to be intended.

Indeed, the Category 1 definition, as proposed by Ofcom, could cover almost any user-to-user service, on the basis that almost all services use algorithmic technology to determine the way in which content is encountered by users. For example, photo storage products may have recommendations of photos from a user's own gallery and navigational tools can use machine learning models to rank photos and reviews of local businesses. However, it is unclear how obligations designed for Category 1 services (like those relating to news publisher content or journalistic content) could apply to navigational or photo tools.

User counts

In relation to the threshold conditions based on user numbers, we note that measuring the number of users on our services is complex, due to difficulties in defining "user count" and due to the different use and functionalities of Google's services, and there is no one methodology for measuring user counts across different products. Users can choose to access many of our services when they are signed into an account or when they are signed out. Given our systems and privacy policies, we cannot comprehensively deduplicate within these costs or between them, which results in significant overcounting. When

calculating average user numbers over a period of time, there are also issues with data retention time frames, as well as seasonal fluctuations in user counts.

We would therefore suggest that there is some flexibility built into user counting, to reflect the different ways in which services operate and ways in which users can be counted in a reliable way, based on the specifics of that service.

- Given data retention policies (which differ across products), there should be some flexibility over requirements for determining monthly user counts and over what period they should be measured.
- Given the complexities involved in collecting user counts, the extensive resources required to complete user counting across all services, and the regulatory burden for Ofcom in collating and reviewing the data, it would be proportionate for user counting to be provided as a one-off when a service reaches the relevant user count threshold or for a service to 'self-certify' that it meets the relevant threshold without needing to provide user counts, rather than ongoing regular user counting reporting for all in-scope services.
- Lastly, given the range of users and services, clear guidance of which services are required to proactively provide user counting would be helpful.

At the time of submitting this response, it is not clear which of Google's services would meet the threshold conditions for Category 1. We have prepared this response through the lens of YouTube, on the basis that it is the most likely service to be relevant to the Category 1 duties.

Executive Summary

Our response to Ofcom's questions about the duties applicable to categorised services are set out below. We have chosen not to respond to each individual question at this stage and, in relation to some duties, we have included a summary answer rather than responding to each sub-question.

We welcome the opportunity to offer our thoughts on how the Codes of Practice and guidance for categorised services should be designed and look forward to commenting on the detail of Ofcom's proposals next year, once the draft documents are published. In line with our response to the illegal harms consultation, our overarching request is that the Codes should strike an appropriate balance between being sufficiently precise so as to give clear directions to services on how they can comply with the Act's obligations, while providing flexibility for services that have long invested in online safety, to leverage and improve their existing systems, processes and technologies, to protect users from harm in line with the Act's objectives.

This flexibility is particularly important in relation to the duties on Category 1 services because some of those duties have the potential to conflict with each other and Ofcom should allow service providers the discretion to decide how to balance competing objectives. In particular, given the breadth of content that could be considered to be journalistic content, content of democratic importance or news publisher content, prescriptive measures in the Code which aim to protect such content could conflict with

the Category 1 services' safety duties and user empowerment duties, leaving service providers in an impossible position.

We would therefore caution against the inclusion of prescriptive measures in the Code which mandate the use of specific processes or tools. Such an approach could inadvertently set a "compliance ceiling", pushing companies like Google, which strive to use the most sophisticated and innovative approaches to compliance, to adopt less effective methods of mitigating risk in order to benefit from certainty of compliance with the Code and thereby the Act.

We have explained in detail below the steps we take on relevant services to ensure we are transparent with our users and that they have access to the highest quality content. We hope this helps inform Ofcom's design of the draft Codes and guidance for categorised services and are very happy to discuss in further detail. This response focuses on the following areas:

Terms of service: YouTube's terms of service incorporate Community Guidelines, which are policies that outline what type of content isn't allowed on YouTube. We use different documents for different types of content but we ensure these documents are accessible in one place and that there are signposts within each document to other relevant policies. We also produce different types of content (including short video explainers) which describe aspects of the terms of service and related policies, such as how they are updated, to make this information as accessible as possible.

Journalistic Content, News Publisher Content and Content of Democratic Importance:

On YouTube, we do not define, identify or categorise content as journalistic content, content of democratic importance or news publisher content. All content is evaluated against our Community Guidelines which are enforced consistently. However, our policies recognise that important and authoritative content, which might otherwise violate our Community Guidelines, may be allowed to stay on YouTube if the content offers a compelling reason with visible context for viewers. This is why we have exceptions for content that is educational, documentary, scientific or artistic - the "EDSA" exception. Journalistic content, content of democratic importance and news publisher content are all likely to benefit from EDSA exceptions. The EDSA exception, along with our YouTube News Features which help users connect with the latest news content, ensures that we are promoting and protecting authoritative journalistic and news publisher content and content of democratic importance. *We therefore request that the Codes incorporate flexibility for services in how they assess whether content is journalistic or of democratic importance, to ensure that services can leverage existing policies and practices.*

We also note that the Act envisages that complaints procedures relating to journalistic content should be 'expedited'. However, we would suggest that this time frame should not be by reference to the turnaround times of other complaints. In other words, if a service is able to resolve all complaints within a short time frame, it should not be penalised for this by being required to respond to journalistic complaints *even more* quickly.

We also note that the definition in s56 of the Act of "recognised news publisher" is extremely broad and lacks clarity (for example because bad actors could claim that the principal purpose of their content is news-related material, published in the course of a business). The entity being "subject to a standards code" does not give any additional comfort in this regard. We would therefore suggest that Ofcom provides a list of recognised news publishers that meet this criteria, in order to ensure that services have clarity over the circumstances in which the temporary must-carry obligation must be applied, and that bad actors cannot use the broad definition to game the system.

User empowerment: On YouTube, we have Community Guidelines that closely mirror and prohibit the types of harmful content considered as 'relevant content' under the Act, to which the user empowerment duties apply. We would welcome clarity that services are not required to provide an optional filter for content that is prohibited across the service as a whole..

User ID verification: In relation to User ID verification, we would again welcome flexibility from Ofcom in the acceptable methods of ID verification. We note that this would be consistent with the apparent policy intention given that the Act makes clear that the verification process may be "of any kind" and "need not require documentation to be provided". The range of acceptable methods must allow for verification to take place in a way that does not undermine user privacy.

Fraudulent advertising: We have sophisticated advertiser verification systems in place to ensure that advertisers have completed one or more verification programs, involving a series of steps that advertisers are to follow and complete. Under this program, they provide information about their business and identity, which we can then verify. As above, we request that any measures introduced in Ofcom's Code of Practice take into account the investment some platforms will already have made in preventing fraudulent advertising by giving them sufficient flexibility to leverage and improve the existing systems rather than starting from scratch.

Should you and your team have any questions about our response, we are more than happy to meet with Ofcom and discuss further.

Your response – Additional terms of service duties

Questions 1 – 5: Terms of service and policy statements

For all respondents

Question 1: What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements? Please submit evidence about what features make terms or policies clear and accessible.
Response: **Our terms of service and policies** We have universal Terms of Service, which apply across a range of Google services. We also have service-specific additional terms and policies for many of our services, which are available on one page (see here). Our terms of service explain, through the use of simple and plain language, how those services work and the user's relationship with Google. These terms of service constitute a legal agreement between Google and the user. They require the user to comply with our policies, which explain what behaviour and content is and is not permitted on our services. A range of supporting materials and resources, across our services, provides additional information to help further explain our policies to users. The use of separate content policies and supplementary materials can also help to ensure that our more formal terms of service do not become overly detailed. Given the pace of changes in policies related to content moderation, including every detailed update to our content policies in our terms of service would become overly burdensome for users, who would be constantly bombarded with updates. Across our services, we aim to make our terms of service and content policies clear and easily accessible to all users and content creators. We avoid using legal jargon wherever possible and, in some cases, we also use video explainers to make sure that our policies are as clear and accessible as possible. **What providers can do** From our experience, to enhance the clarity and accessibility of terms of service and public policy statements, providers can publish terms of service and content policies that are publicly available in clear, plain language and accessible formats. Service providers can also (where appropriate) produce different types of content (including short video explainers) which describe aspects of the terms of service and related policies. It is important that services are clear and transparent in their policies about what type of

content is prohibited and how they treat it. Users should be provided with information that is as precise and specific as reasonably possible.

Our approach on YouTube

YouTube has its own Terms of Service and Community Guidelines (distinct from other Google services). This is because of its unique features, including the ability for a large number of users to rapidly share and access video content, and the need to address the particular content issues that may arise on the service.

YouTube's Community Guidelines include clear statements on content that is not permitted on the platform. There are specific policies particularly relevant to illegal content, including on:

- Spam, deceptive practices & scams
- Child safety
- Suicide and self-harm
- Harassment and cyberbullying
- Harmful or dangerous content

Alongside our Community Guidelines**:**

- We provide a summary explanation of how our policies are updated, as well as a user friendly video explanation.
- We explain how policy violations are detected, including that we use a combination of people and machine learning, and how human flagging of content works.

- We tell our users how potential policy violations are considered against exceptions, such as content that is educational, documentary, scientific or artistic.
- We also explain what action we take in respect of policy violations, our "strike" system and how we age-restrict content, with links to more detailed resources.

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: N

Question 2: How do you think service providers can help users to understand whether action taken by the provider against content (including taking it down or restricting access to it) or action taken to ban or suspend a user would be justified under the terms of service?

In your response to this question please consider and provide any evidence related to the level of detail provided in the terms of service themselves, whether services should provide user support materials to help users understand the terms of service and, if so, what kinds of user support materials they can or should provide.
Response: See response to Question 1 above.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For providers of online services

Question 3: How do you ensure users understand the provisions in your terms of service about taking down content, restricting access to content, or suspending or banning a user from accessing the service and the actions you might take in response to violations of those terms of service? In your response to this question, please provide information relating to (a) – (d) where relevant.
(a) how you ensure your terms of service enable users to understand both what is and is not allowed on your service, and how you will respond to user violations of these rules;
Response: See response to Question 1 above.
(b) any relevant considerations about the risk of bad actors taking advantage of transparency around your terms of service and how they are enforced;
Response: It is important that services are clear and transparent in their policies about what type of content is prohibited and how they treat it. Users should be provided with information that is as precise and specific as reasonably possible. Services are required to balance the need to provide users with sufficient information to understand the terms of use, with the need to ensure that the terms of service do not become unwieldy or difficult to understand. The Codes should give services some discretion over how they strike this balance, depending on the nature of the service and content that users are likely to encounter. In Ofcom's approach to terms of service, we would recommend that it takes due account of the risk of inadvertently exposing sensitive information to bad faith actors, such as terrorists or hostile states. Expectations on services must reflect the risk of giving users descriptions of methods and tools used in content moderation at a level of detail that could allow bad faith actors to game platforms' systems, to the detriment of user safety.

(c) details about any user support materials or functionalities you provide to assist users to better understand or navigate your terms of service or related products;
Response: See response to Question 1 above.
(d) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: N

Question 4: Please describe the processes you have in place to measure user engagement with and comprehension of your terms of service and how you make improvements when required. In your response to this request, please provide information relating to (a) – (f) where relevant.
(a) how you measure user engagement with/comprehension of your terms of service and the metrics you collect;
Response: **How we monitor the effectiveness of our terms** We track traffic to the webpage policies.google.com which contains our universal Terms of Service. In 2021, there were 425 million visits to this page over a 90-day period. □ We continually review our internal standards on the language of our policies and how we can make our policies clear and intelligible to users, in areas such as how we write, format and present our policies externally. Internal guidelines are reviewed and made available to teams within Google responsible for policies. **How much resource we dedicate to this** We use both internal and external resources to design and maintain our terms of service. Numerous teams across the company - including our Trust & Safety, Public Policy and Legal teams, together with our Product teams - are involved in user safety and in the enforcement of our terms of service.
(b) any behavioural research you undertake to better understand engagement with and/or comprehension of your terms of service (including any research into reasons why users do not engage with terms of service);
Response: See response to sub-section (a) above.

(c) any measures you have taken to improve engagement with and/or comprehension of your terms of service, including (but not limited to) how the findings of any behavioural research influenced these measures and/or any design changes (e.g. prompts to remind users to read the terms of the service, changes to the structure of the terms of service or changes to how users access the terms of service etc.);
Response:
(d) costs of these processes (including the design, implementation and continued use of these processes or updated versions of these processes);
Response:
(e) how you evaluate the effectiveness of measures designed to improve engagement with and/or comprehension of your terms of service;
Response:
(f) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: □ partly

Question 5: Please describe any evidence you have about the effectiveness of using different types of mechanisms to promote compliance with terms of service or change user behaviour in the event of a violation, or potential violation, of terms of service. In your response to this request, please provide information relating to (a) – (d) where relevant.
Response:
(a) any evidence about the effectiveness of enforcement measures such as taking down content, restricting access to content, or suspending or banning user accounts in relation to encouraging users to comply with specific aspects of terms of service in the future
Response:
(b) any evidence about how effective non-enforcement mechanisms are at reducing violations of the terms of service or repeated violations, including the type of non-enforcement mechanism and how it is implemented (e.g. prompts for users to consider the appropriateness of their content before posting it to the service (with or without links to specific provisions within the terms of service), or prompts for users to review certain provisions within the terms of service when their content is found to violate these provisions)
Response:
(c) any information and/or evidence on the costs of designing and implementing different types of enforcement or non-enforcement mechanisms (including costs of the research behind the design, implementation and continued assessment/study of these mechanisms)
Response:

(d) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Questions 6 – 8: Reporting and complaints processes

For all respondents

Question 6: What can providers of online services do to enhance the transparency, accessibility, ease of use and users' awareness of their reporting and complaints mechanisms? In your response to this question, please provide evidence about what features make user reporting and complaints systems effective. In your response to this question, please provide information relating to (a) – (h) where relevant.
On **YouTube**: When a creator's video or channel is removed due to a policy violation, we provide a link with simple steps to appeal the decision. If a creator chooses to submit an appeal, that appeal is reviewed by a member of our Trust & Safety team, and the decision is either upheld or reversed We keep detailed records, including data on complaints, appeals and reinstatements, and we think it is important to be transparent about this information, which we publish on a quarterly basis, in our Community Guidelines Enforcement Report. This report includes data on video, channel, and comment removals; appeals and reinstatements; and human and machine flagging. Our "three strikes" approach to moderation on YouTube enables us to balance our aims of keeping users safe online while also preserving freedom of expression for our creators, as well as allowing us to educate creators about our Community Guidelines before removing their videos or channels. ○ We understand mistakes happen and creators don't mean to violate our policies - that's why the first violation is typically only met with a warning. ■ For a subsequent violation, we issue a "strike" against the channel. Strikes also come with upload freezes, meaning that creators who receive a strike are barred from uploading to the platform for one week on their first strike and two weeks on their second. Approximately 94% of people who receive a first strike never receive a second one.

■ If a creator receives three strikes within 90 days, their channel, and therefore all of its content, is removed from YouTube. ○ Users who repeatedly or maliciously report content or otherwise misuse our complaints mechanisms may have their accounts suspended and be prohibited from using them.
(a) reporting or complaints routes for registered users, non-registered users and potential complainants (being affected persons who are not users of the service)
Response:
(b) how to ensure that reporting and complaints mechanisms are not misused
Response:
(c) the key choices and factors involved in designing these mechanisms
Response:
(d) how users can or should be supported to report/complain about specific concerns (e.g., other users, certain types of content or, appeal content takedowns or account bans)
(e) how to ensure they are user-friendly and accessible to all users (e.g., disabled users, children)
Response:
(f) whether users are informed that their reports are anonymous (e.g., other users will not be informed about who has reported their content or account);
Response:
(g) any user support materials that explain how to use the reporting and complaints process and what will happen when users engage with these systems
Response:
(h) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For providers of online services

Question 7: Can you provide any evidence or information about the best practices for effective reporting and/or complaints mechanisms, and how these processes are designed and maintained? In your response to this question, please provide evidence relating to (a) – (j) where relevant.

Response:
(a) how users report harmful content on your service(s) (including the mechanisms' location and prominence for users, and any screenshots you can provide);
Response:
(b) whether there are separate or different reporting or complaints mechanisms or processes for different types of content and/or for different types of users, including children;
Response:
(c) how users appeal against content takedowns, content restrictions or account suspensions or bans;
Response:
(d) what type of content or conduct users and non-users may make a complaint about / report, including any specific lists or categories;
Response:
(e) whether users need to create accounts to access reporting and complaints mechanisms (if there are multiple mechanisms, please provide information for each mechanism);
Response:
(f) whether reporting and complaints mechanisms are effective, in terms of: **(i) enabling users to easily report content they consider to be potentially the types of content specified in the relevant terms of service, and how to determine effectiveness;**
Response:
(ii) enabling, supporting or improving the accuracy of user reporting in relation to identifying the types of content specified in the relevant terms of service, and how to determine effectiveness;
Response:
(iii) enabling, supporting or improving the provider's ability to detect and take timely enforcement action against content or users as specified in the relevant terms of service, and how to determine effectiveness;
Response:
(g) whether there are any reporting or complaints mechanisms you consider to be less effective in terms of identifying certain types of content and how you determine this;
Response:
(h) the use of trusted flaggers (and if reports from trusted flaggers should be prioritised over reports or complaints from users);
Response:

(i) the cost involved in designing and maintaining reporting and/or complaints mechanisms, including any relevant issues, difficulties or considerations relating to scalability; and
Response:
(j) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 8: What actions do or should services take in response to reports or complaints about content that is potentially prohibited or accounts engaging in potentially prohibited activity? In your response to this question, please include information relating to (a) – (g) where relevant.
Response:
(a) what proportion of reports are reviewed, and what proportion result in action taken including; **(i) any potential variation in the number and actionability (i.e., the proportion that result in a takedown or other action) of reports or complaints in relation to different provisions within your terms of service;**
Response:
(ii) any differences for cases involving multiple reports/complaints about a single piece of content or user;
Response:
(iii) the costs associated with reviewing reports;
Response:
(b) whether any reports or complaints are expedited or directed to specialist teams, including: **(i) the criteria for this;**
Response:
(ii) the cost involved in facilitating this;
Response:
(c) the extent to which relevant individuals (content creators, users, and non-registered or logged-out users) are informed about the progress of their report or complaint, including: **(i) if they are not, the reasons why;**
Response:

(ii) if they are, what is included when users are informed about the progress of their report (e.g. receipt of the report, the progress of the report through the service's review process, and/or the outcome of the report);
Response:
(iii) the technical mechanisms/process to inform any relevant individuals about the progress of their report (e.g., whether non-registered users are provided an opportunity to provide an email address);
Response:
(iv) any differences in responses to different types of reports (e.g., reports about content or an account a user believes violates the terms of service, about the provider not operating in line with its terms of service, or about the accessibility, clarity or comprehensibility of those terms of service);
Response:
(v) the costs associated with responding to reports;
Response:
(d) what happens to the content while it is being assessed/processed (e.g., if and how it may still be found or viewed by other users);
Response:
(e) any internal or external timeframes or key performance indicators (KPIs) for reviewing and/or acting on reports or complaints;
Response:
(f) any user support materials that are used or should be used to support users understand the service's responses to reports, or how users can appeal moderation decisions about their content or accounts, or about decisions taken in response to reports they have submitted about other users' content or accounts;
Response:
(g) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Questions 9 – 15: Moderation

For all respondents

Question 9: Could improvements be made to content moderation to deliver more consistent enforcement of terms of service, without unduly restricting user activity? If so, what improvements could be made?

In your response to this question, please provide information relating to (a) –(c) where relevant.
Response:
(a) improvements in terms of user safety and user rights (e.g., freedom of expression), as well as any relevant considerations around potential costs or cost drivers;
Response:
(b) evidence of the effectiveness of existing moderation systems including any relevant examples of the accuracy, bias and or effectiveness of specific moderation processes;
Response:
(c) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For providers of online services

Question 10: Please describe circumstances where you have taken or would take enforcement action against content or users outside of what is set out publicly in your terms of service and the reasons for taking this action. In your response to this question, please provide information relating to (a) – (e) where relevant.
Response: We do not take enforcement action against content or users outside of what is set out publicly in our terms of service. To the extent that a new type of harmful behaviour was identified, for example content arising out of an unexpected event such as the Covid-19 pandemic, we would update our terms of service to reflect this.
(a) the types of action taken, and frequency of these actions (including per type of action);
Response:
(b) how relevant content or users were or would be brought to your attention;
Response:
(c) any policies, approaches or processes you have used or would use to guide moderation decisions in these cases;
Response:
(d) whether new policies are or would be written in response to these cases, and if so:

(i) whether and when these new policies are written before enforcement action is taken or after;
Response:
(ii) when and how these new policies would be added to or included in your publicly available terms of service;
Response:
(e) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 11: If you are made aware of content or an account that potentially violates your terms of service, please describe any relevant circumstances which might not result in enforcement action, immediately or at all. **In your response to this question, please provide describe (with examples) any relevant circumstances relating to (a) – (e).**
In relation to **YouTube**: As explained above, alongside our Community Guidelines, we tell our users how potential policy violations are considered against exceptions, such as "EDSA" content that is educational, documentary, scientific or artistic. In particular, we explain to our users that sometimes videos which might otherwise violate our Community Guidelines may be allowed to stay on YouTube if the content offers a compelling reason with visible context for viewers. We often refer to this exception as "EDSA", which stands for "educational, documentary, scientific or artistic'. To help determine whether a video might qualify for an EDSA exception, we look at multiple factors, including the video title, descriptions and the context provided. EDSA exceptions are a critical way in which we make sure that important speech stays on YouTube, while protecting the wider YouTube ecosystem from harmful content.
(a) circumstances that relate to issues or challenges within your content moderation system (e.g. moderator error, language or local knowledge gaps, content is no longer available (e.g. livestream), nuance/context of content means it is found non-violative, further investigation needs to be done before action can be taken);
Response:
(b) circumstances that relate to issues or challenges within your terms of service and/or associated policies (e.g. new iterations of a harm falls outside the scope of internal moderation policies, individual piece of content is only of concern at scale (but itself does not violate policies);

Response:
(c) circumstances that relate to competing priorities (e.g., freedom of expression, public interest concerns);
Response:
(d) circumstances that would be understood by a user who has read the terms of service and why or why not, (e.g., the terms of service sets out exception for not removing violating content (e.g. news content), or transparency is not provided to avoid empowering bad actors);
Response:
(e) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 12: What automated systems do you have in place to enforce terms of service provisions about taking down or restricting access to content or suspending or banning accounts? **In your response to this question, please provide information relating to (a) – (d).**
Response:
(a) the suitability/effectiveness of automated systems to identify content or accounts likely to violate different provisions within your terms of service, including the factors that materially impact suitability/effectiveness (e.g. language of content, type of content) including: **(i) the suitability/effectiveness of automated systems to take down content, apply access restrictions or ban accounts in relation to any or certain provisions within your terms of service without further assistance from human moderation;**
Response:
(ii) how you use your recommender systems to restrict access to certain content, and how you measure the effectiveness and any unintended consequences of using the recommender system in this way;
Response:
(iii) whether and how automated moderation systems differ by type of content (e.g., audio, video, text) or type of violation (of provisions within your terms of service) and any relevant information about costs of these different systems;
Response:
(iv) how data is used to develop, train, test or operate content moderation systems is sourced for different provisions within your terms of service;

Response:
(v) how performance/effectiveness/accuracy of automated systems are assessed and improvements then made, including any relevant considerations or differences for different provisions within the terms of service (e.g., tolerance level for false negatives and false positives between different provisions);
Response:
(vi) how and when automated systems are updated, and the trigger for this (e.g., in response to changing user behaviour or emerging harms);
Response:
(vii) what safeguards are employed to mitigate biases or adverse impacts of automated content moderation (e.g., on privacy and/or freedom of expression), and any relevant considerations or differences for different provisions within the terms of service;
Response:
(b) the range and quality of third-party content moderation system providers available in the UK, particularly for different provisions within your terms of service;
Response:
(c) the process and costs associated with expanding use of existing automated moderation systems for additional provisions in your terms of service, and any relevant barriers or challenges in deploying these automated moderation systems or expanding or upgrading these systems to cover new or additional provisions;
Response:
(d) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 13: How do you use human moderators to enforce terms of service provisions about taking down or restricting access to content, or suspending or banning accounts? **In your response to this question, please provide information relating to (a) – (c).**
Response:
(a) how you determine your services' resource requirements in relation to human moderation, and the factors (or key factors) that impact these requirements (e.g., increases in content or users, the range or types of content prohibited in your terms of service or technological advances in your automated system) including;

(i) which languages are covered by your moderation team and how you decide which languages to cover;
Response:
(ii) whether moderators are employed by the service or outsourced, or are volunteers/users and any differences regarding how different provisions within the terms of service are moderated;
Response:
(iii) whether and how moderators are vetted, and any relevant consideration for how moderators are assigned to different roles relating to different provisions within the terms of service;
Response:
(iv) the type of coverage (e.g., weekends or overnight, UK time) moderators provide and any relevant considerations for different provisions within the terms of service;
Response:
(b) the process and costs associated with extending the use of human moderation for new/additional provisions in your terms of service, and any relevant barriers or challenges to adding new/additional provisions in your terms of service in relation to your human moderation resources;
Response:
(c) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 14: What training and support is or should be provided to moderators, and what are the costs incurred by providing this training and support? **In your response to this question, please provide information relating to (a) – (g).**
Response:
(a) whether certain moderators are specialised in certain harms or subject material relating to different provisions in the terms of service;
Response:
(b) how services can/should/do assess the accuracy and consistency of human moderation teams;
Response:

(c) the impact of mental health or well-being support for moderators on the effectiveness of content moderation (including impacts on turn-over in moderation teams);
Response:
(d) whether training is provided and/or updated (including for emerging harms), and the frequency of these updates;
Response:
(e) the costs of creating training materials and support systems, and then the costs of updating or expanding these materials and systems (when relevant/required);
Response:
(f) how training, guidance and/or any relevant support systems and/or materials are provided to moderators including which moderators it is provided to (internal, contract, volunteer etc);
Response:
(g) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 15: How do human moderators and automated systems work together, and what is their relative scale in relation to each other regarding how you ensure your terms of service are enforced? **In your response to this question, please provide information relating to (a) – (e).**
Response:
(a) how and when automated systems or human moderators are deployed in the moderation process;
Response:
(b) the costs of different systems or processes and of using different combinations of these systems and processes. In the absence of specific costs, please provide indication of cost drivers (e.g., moderator location) and other relevant figures (e.g., number of moderators employed, how many items the service moderates per day);
Response:
(c) how the outputs of human moderators, or appeal decisions are used to update the automated systems, and what steps are taken to mitigate bias;
Response:
(d) whether there are any relevant differences or considerations for costs or quality assurance processes for moderating different provisions within the terms of service; and

Response:
(e) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Your response – News publisher content, journalistic content and content of democratic importance

Questions 16 - 17: Identifying, defining, and categorising journalistic content, news publisher content and content of democratic importance

For all respondents

Question 16: What methods should service providers use to identify and define journalistic content and content of democratic importance, particularly at scale? In your response to this question, please provide information relating to (a) where relevant.
We consider that the nature of the duties and requirements in respect of journalistic content, news publisher content and content of democratic importance is highly instructive as to the services that Parliament envisaged should be Category 1. On all services other than YouTube, these broad categories of content are simply not prevalent and therefore it would be inappropriate and disproportionate for such services to be categorised. On YouTube, we do not define, identify or categorise content as journalistic content, content of democratic importance or news publisher content. All content is evaluated against our Community Guidelines which are enforced consistently, regardless of the subject or the creator's background, political viewpoint, position or affiliation. Given the breadth of content that could be considered to be journalistic content, content of democratic importance or news publisher content and the interrelation between the duties in respect of this content, the safety duties and the cross-cutting duties in the Act, it is **critical** that service providers are afforded a broad degree of flexibility in how they seek to comply. If Ofcom is too prescriptive as to how service providers must comply in respect of the duties concerning journalistic content, content of democratic importance and news publisher content, that could lead to the delicate balance between protecting users from harm and fostering healthy democratic discourse being disturbed. Misuse of protections and publisher privileges is a significant concern for our platform, as are broad legislation and frameworks that look to privilege publisher content in a way

that could result in misuse. Broad definitions, lack of clarity, and regulatory overlap create a complicating environment for a platform like YouTube to operate in -- especially given any protections offered within one jurisdiction due to local laws are not universal, such that an obligation to not moderate news publisher content in the UK until an appeals process has been complete does not remove our obligation to users in other markets to apply our Community Guidelines fairly and consistently. In some jurisdictions (e.g. EU) there is recognition that users need to be protected from harm as a matter of priority which is why there are certain safeguards to ensure the provisions are more nuanced and allow platforms to address systemic risks first.

What is more, not all regulatory frameworks are created the same and there are real risks that jurisdictions looking to copy the Ofcom Codes may not be motivated by the same impartial objectives as the UK OSA regime.

What we do at YouTube

At YouTube, we believe that people should be able to speak freely, share opinions, foster open dialogue and that creative freedom leads to new voices, formats and possibilities. We are therefore committed to connecting users with authoritative news sources and recognise the significance of users accessing content that could fall within the definitions specified in the Act. We seek to achieve this through the following measures.

<u>YouTube News Features</u>

We have developed a number of features to help users connect with the latest news on YouTube. These features rely on signals to identify news moments and top stories of the day and the content comes from **authoritative** news sources. The specific videos that surface are algorithmically generated based on several signals, which may include, but are not limited to: local relevance and language; relevance to news topic or event; freshness of content; reporting intent, or videos with journalistic context; and YouTube's standard recommendation signals.

These features include:

- **Breaking News Shelf on the Homepage**: When there's a significant news event, a breaking news shelf may surface on the YouTube homepage. Examples of significant news events may include large-scale tragedies or momentous political occurrences. The shelf will surface for signed-in users that are 18 years old and older. If a user is not interested in this shelf, they can dismiss it.

- **Top News Shelf in YouTube Search:** If you search for a specific news story on YouTube, you may notice a top news shelf in your search results. If it's one of the top news stories of the day, the top news shelf will showcase content related to the news story. The Top News shelf will show up regardless of age and can't be dismissed.

- **Top News Shelf on the Homepage**: If a user often watches or searches for news on YouTube, a top news shelf may show on their homepage. The top news shelf includes content related to the top news stories of the day. This shelf will show up for signed in users regardless of age.

- **YouTube.com/news:** YouTube.com/news highlights the top news stories and videos of the day.

- **News Watch Page**: The news watch page brings together news stories from authoritative and diverse sources on YouTube to help viewers deep dive and explore different sides of a news story. The watch page feature stories across multiple formats including:

    - Latest Updates, with the most recent video coverage of the news story.
    - Explanations and Commentary, with additional context on the news topic.
    - Live News, with live streams showing what's happening at the moment.
    - Shorts, to quickly catch up on the news story's latest updates.

  Viewers can open the watch page for a specific news story by clicking on a video with the newspaper icon on the YouTube homepage, news destination page, or in search results.

The application of EDSA exceptions

The application of EDSA exceptions is described in detail here and here. However, the key points are as follows:

- Sometimes, videos that might otherwise violate our Community Guidelines and associated policies may be allowed to stay on YouTube if the content offers a compelling reason with visible context for viewers. We often refer to this exception as "EDSA," which stands for "Educational, Documentary, Scientific or Artistic."

- The application of an EDSA exception is context and content specific. EDSA exceptions are frequently applied in three circumstances:

    - First, where the content contains basic facts about an event that is happening. For instance, it identifies who is in the content, describes what the content shows or when and where it takes place, or explains why certain content is present.
    - Second, where the content condemns certain claims, includes opposing points of view or is satirical.
    - Third, where content discourages dangerous behaviour.

- We provide numerous examples of content that is more likely to secure an EDSA exception and content that is not likely to secure an EDSA exception here and here.

- EDSA exceptions are applied on a case-by-case basis by our content reviewers and irrespective of who the publisher is.

- In certain cases, in order to secure an EDSA exception, the context provided by the creator must be provided in the audio or video content itself and not, for example, in the description of the content. Further, there are certain types of content that can never benefit from an EDSA exception.

Journalistic content, content of democratic importance and news publisher content is likely to benefit from EDSA exceptions. We consider that the application of EDSA exceptions to specific content (rather than at an account level) is a tailored and proportionate way of ensuring that our users are able to access the content but not in a way which unduly exposes them to harmful content.

Priority flaggers

Flags from human detection can come from a user or a member of YouTube's Priority Flagger program.

The YouTube Priority Flagger program members include NGOs and government agencies that are particularly effective at notifying YouTube of content that violates our Community Guidelines.

Our Priority Flagger program helps provide robust tools to government agencies and non-governmental organisations. These agencies and NGOs are particularly effective at telling YouTube about content that violates our Community Guidelines. Content reported by Priority Flaggers is not automatically removed or subject to any differential policy treatment — the same standards apply for flags received from other users. But, because of their high degree of accuracy, our teams prioritise flags from Priority Flaggers for review.

Transparency

We do not track enforcement action by who the speaker is, their affiliation or whether content falls within a particular definition or category of content. However, we do release quarterly transparency reports on our enforcement actions. These reports are broken down by (among other things): how we detect harmful content; the reasons for removal of channels, content and commentary; and data of video views in advance of detection.

In our most recent reporting quarter, from January 2024 to March 2024, of the approximately 8.2 million videos removed, 7.9 million were through automated flagging. Of the remaining, 238,058 were first flagged by users, 60,676 were by organisations and 6 were by Government agencies. The report demonstrates the impact of our highly effective automated flagging systems - with 96% of our removed videos first detected by our machines. We are able to react more quickly and accurately to enforce our policies, meaning more than 57% of videos were removed before they received any views.

Our quarterly transparency reports can be found here.

Complaints, counter-notice and appeal processes

All creators enjoy the same rights on YouTube regarding complaints and appeals. We do not have separate processes depending on the nature of the content that the complaint or appeal concerns. For a description of that complaints and appeals process, see the response to question [x] above.

Elections

Users may come to YouTube for news and information about elections, seeking information on, for instance, where to vote and what candidates are saying. We believe we have a responsibility to support an informed citizenry and foster healthy political discourse. For this reason:

- We have a set of misinformation policies specific to election contexts. In particular, we do not allow any content on YouTube which interferes with democratic processes. This includes the following content:
    - **Voter suppression**: Content aiming to mislead voters about the time, place, means, or eligibility requirements for voting, or false claims that could materially discourage voting.
    - **Candidate eligibility**: Content that advances false claims related to the technical eligibility requirements for current political candidates and sitting elected government officials to serve in office. Eligibility requirements considered are based on applicable national law, and include age, citizenship, or vital status.
    - **Incitement to interfere with democratic processe**s: Content encouraging others to interfere with democratic processes. This includes obstructing or interrupting voting procedures.
- Policies across other areas are applied to election-specific content. For example, our harassment policy prohibits content that threatens poll workers.
- Our Intel Desk monitors trends and stays ahead of potential threats to democratic processes that may appear on our platform.
- When voters look for election-related information on YT, our systems prominently raise content and relevant context from authoritative sources in search results, the homepage, and the “watch next” panel.
- During key civics and election moments, we make additional efforts to raise information from authoritative sources, which may include information panels that point users to authoritative sources about certain election-related topics.
- We’ve also invested heavily to strengthen transparency around AI-generated content. We recognize that bad actors may try to exploit Generative AI tools during an election — which is why we’re ensuring that several layers of transparency and protections are in place, including current and upcoming elections:
    - We now require creators to disclose whether the video they’re uploading contains altered or synthetic content and is realistic.
    - We will then add a transparency label to realistic synthetic or altered content to provide this important context to viewers. For election content, the label will appear on the video itself and in the video description.
    - In certain cases, we may also add a label even when a creator hasn't disclosed it, especially if the use of altered or synthetic content has the potential to confuse or mislead viewers.

- Coordinated influence operations are not allowed on YouTube, regardless of the political viewpoints they support. We work closely with Google's Threat Analysis Group (TAG) to spot these types of campaigns on YouTube and terminate their channels. This can include attempts to interfere with elections. For example, as TAG shared, before the recent Portuguese elections, we terminated 7 YouTube channels linked to individuals in Argentina as part of our ongoing investigations into disinformation campaigns. Through TAG, we also share threat information with law enforcement agencies.

(a) how journalistic content and content of democratic importance can be described in the terms of service so that users can reasonably be expected to understand what content falls into these categories.

As mentioned above, in the context of removals, we assess whether the content has educational, documentary, scientific or artistic (EDSA) context.

The type of context that a creator must include to get an EDSA exception depends on what's in the content; and we make most EDSA exceptions when the content has one or more of the following:

1. Basic facts about what's happening in the content: Identify who's in the content, describe what the content shows or when and where it takes place, or explain why certain content is present.

2. Condemnation, opposing views or satire: Communicate that your content condemns certain claims, includes opposing points of view or is satirical.

3. Discouragement of dangerous behaviour: Tell viewers not to imitate what's in the video.

Each of these is explained in detail in our Help Center and include examples, reminding creators that including different types of content (like basic factual information, multiple points of view, and clear and informative discouragement against imitating dangerous or harmful behavior) is the best approach to ensure the content uploaded will meet the bar for the EDSA exception. The Help Center also reminds creators to add context to their EDSA content, and reminds them that we don't make EDSA exceptions for context that may appear in comments, tags, channel descriptions, pinned comments, or other surfaces since that content is not always visible to viewers.

We are also clear with our creators about where EDSA context should be added, and encourage them to add EDSA context to the video, audio, video title, and video description.

In some instances we may apply an age restriction or warning to content even if it gets an EDSA exception because some viewers may find it sensitive or inappropriate (e.g. war zone footage). Additionally, we may make EDSA exceptions based on public interest. This could include, for example, controversial content featuring national political candidates

on the campaign trail, graphic footage from active warzones or humanitarian crises, comments disputing health authority guidance made during a public hearing or nudity in the context of sex education.

However, certain content isn't allowed on YouTube, even if it has context added. For these reasons we are clear in instructing users not to post the following content, no matter what:

- Child sexual abuse media (CSAM)
- Video, still imagery or audio of violent physical sexual assaults
- Footage filmed by the perpetrator of a deadly or major violent event that shows weapons, violence or injured victims
- Unmodified reuploads of content created by or glorifying violent terrorist or criminal organisations
- Instructions on how to self-harm or die by suicide
- Instructions on how to build a bomb that's meant to injure or kill others
- Instructions on how to manufacture a firearm or prohibited accessories
- Offers of prohibited sales
- Instructions on how to use computers or information technology to compromise personal data or cause serious harm to others
- Content that reveals an individual's private information, such as their home address, email addresses, sign-in credentials, phone numbers, passport number or bank account information (doxxing)
- Hardcore pornography
- Spam

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

For providers of online services

Question 17: What, if any, methods are in place for identifying, defining or categorising content as journalistic content, content of democratic importance or news publisher content on your service?

In particular, please provide any evidence regarding the effectiveness of any existing methods.

Our content moderation and content enforcement actions are applied to creators equally—regardless of the subject or the creator's background, political viewpoint, position, or affiliation.

Importantly, YouTube is not in a position to assess whether a creator is a News Publisher under the Act's definition, given, we do not hold information on:

(i) whether the creator has as its principal purpose the publication of news-related material;

(ii) whether the content it posts is subject to editorial control;

(iii) whether the creator is working with a 'view to a profit';

(iv) whether it is subject to a standards code;

(d) whether it has policies and procedures for handling and resolving complaints; and

(e) whether it is the person with legal responsibility for the content published.

In order to give News Publishers and Category 1 services certainty over which entities should be treated as a 'Recognised News Publisher', it would therefore be helpful for Ofcom to maintain an industry-wide list of these entities.

Please **see response in Question 16** for more detail on how our systems use various signals to identify news sources for our news features.

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

Question 18: Moderating journalistic content, news publisher content and content of democratic importance

For providers of online services

Question 18: What considerations are taken into account when moderating journalistic content, news publisher content and content of democratic importance?

In your response to this question, please provide information relating to (a) – (e) where relevant.

Response:

On YouTube, News content from authoritative sources is showcased in prominent ways through our News features - as explained in **Questions 16 and 17.** That said, irrelevant of who the publisher is, all content is evaluated against our Community Guidelines and complementary sets of policies.

(a) once identified, how journalistic content, news publisher content and content of democratic importance is actioned and what kind of action is taken; and how that differs from the moderation of other types of content

Response:

As mentioned above in **Question 18**, YouTube's Community Guidelines are enforced consistently across the globe, irrespective of the speaker. This is true for the way our systems and classifiers operate, as well as our human evaluators. This means that our

content moderation and content enforcement actions are applied equally—regardless of the subject or the creator's background, political viewpoint, position, or affiliation.

YouTube is committed to raising authoritative content and reducing borderline content. There are a lot of signals - such as relevance and popularity - that matter in determining what videos you typically see in YouTube search and recommendations. However, when it comes to topics such as news, politics, medical, and scientific information, we know there is no substitute for authoritativeness. That's why we have introduced a range of features to tackle this challenge holistically, set out in detail in **Question 18**.

For example, in search results and recommended videos, we raise authoritative voices for newsworthy events and topics prone to misinformation. We also have dedicated product features such as the Breaking News shelf and Top News shelf that feature relevant videos from authoritative news sources. Context is critical when evaluating information, so we also provide information panels that feature text-based information alongside certain search results and videos to help users make their own decisions about the content that they find on YouTube.

(b) the factors that are or should be considered when taking action (e.g.: downranking/removal/suspension/ban or other) regarding this content

As previously mentioned, our enforcement against our Community Guidelines is applied equitably across all content uploaded, irrespective of the speaker. This helps ensure that we do not inadvertently moderate content in an unfair or biased manner.

When it comes to taking action on content, our classifiers are agnostic and in no way account for the creator or creator's affiliation. Reviews of all content begin by assessing whether content uploaded to the platform is in violation of our Community Guidelines. If a violation is identified, then the content is considered against whether there is sufficient EDSA context to merit an EDSA exception. As shared publicly in this blog post, we do not automatically give exceptions to a video just because it is being presented as part of a news broadcast. The educational or documentary intent needs to be clear by providing context.

As mentioned above **in Question 16 and 18a,** at the heart of our approach are the four Rs, including removing violative content and reducing the spread of harmful misinformation.

This moderation does extend to news publishers if their content is policy violative. There have been a number of high profile examples in recent years where recognized news outlets (even, for example, those with Ofcom broadcasting licences) have published content that does not comply with our Community Guidelines. We would be happy to follow up with further sensitive case studies; given the protections offered to a vaguely defined cohort of news publishers by the OSA legislation, it is critical any code does not inhibit platforms' ability to protect UK users from harmful content.

(c) the proportion of all journalistic content, content of democratic importance and news publisher content actioned upon by you that is actioned based on algorithmic decision making
We do not track enforcement action on content based on the speaker or speaker affiliation (e.g. journalist or news publisher). We release quarterly transparency reports on our enforcement actions based on channel removal and reason, video removal and reason, commentary removal and reason, and data on video views by detection method (including automated flagging or human detection) and in advance of removal. From October 2023 to December 2023, of the approximately 9 million videos removed, 8.6 million were automatically flagged for removal. Automated flagging enables us to act more quickly and accurately to enforce our policies. This chart within our transparency report shows the percentage of video removals that occurred before they received any views versus those that occurred after receiving some views. While we do not specifically track the proportion of journalistic content or news publisher content that is actioned upon, we do have specific examples where our enforcement action has been in response to known publisher uploaded content. For example in March 2022, we age-restricted a video uploaded by a German public service broadcaster for violating our graphic violence policies where the content included footage of corpses from the Russo-Ukrainian war. Although much news publisher content is non-violative of our content policies or benefits from an EDSA exemption, there are occasions when we have had to age restrict or block UK news publisher videos. We would be happy to discuss this further with Ofcom if helpful.
(d) the proportion of all journalistic content, content of democratic importance and news publisher content actioned upon by you that is reviewed by human moderators and on what basis content is escalated to be reviewed by human moderators
Please see response to question above.
(e) any insights into the costs of moderating journalistic content and content of democratic importance, including set up and ongoing costs in terms of employee time and other material costs.
Response: Since YouTube launched nearly 20 years ago, we've collaborated and supported local, national and international news organisations directly and indirectly, to support innovation and success across the ecosystem. Our support is primarily through our leading revenue-sharing system, training programs, and tools to promote innovation, ease of upload, offer a diversity of format, support entrepreneurism and digital transformation. On YouTube, all publishers who choose to participate on the platform are eligible to pursue monetization through our YouTube Partner Program.

As a result of our commitment to the sustainability, transformation, and success of the news industry, we incur direct and indirect financial cost. This comes in the form of employee salary and time, engineering and operations costs, event execution (like our annual news working group summit and regional fora), and more. At YouTube we offer our creators and users customer support teams; and for eligible managed creators, which include news publishers, we offer more dedicated human support through our strategic partner manager teams. There are also dedicated full time divisions of our Trust and Safety teams who support news publishers' content by applying appropriate EDSA exceptions, alongside other specific support during sensitive events (like elections). This is not exhaustive, but is reflective of the ways in which we have prioritised staff experience, time, and hiring decisions as a result of the value we place on news content.

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

Questions 19 – 21: Complaints and appeal processes for journalistic content, news publisher content and content of democratic importance

For all respondents

Question 19: What complaint, counter-notice or other appeal processes should be in place for users to contest any action taken by service providers regarding journalistic content and content of democratic importance?

In your response to this question, please provide information relating to (a) and (b) where relevant.

Response:

(a) examples of effective redress mechanisms that you consider would be most suited to these content types

Response: We consider that our current redress mechanisms that we have in place are effective for journalistic content, content of democratic importance and news publisher content. We further noted that forms of must carry notifications risk users being exposed to harmful content particularly due to the risk of hostile foreign states and extremist groups exploiting this in order to spread misinformation.

(b) briefings, investigations, transparency reports, media investigations and research papers that provide more evidence

Response: YouTube's Transparency report is available here.

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

Question 20: What initiatives could service providers use to create and increase awareness about the process for users to complain and/or appeal content decisions and to minimise its' misuse?

In your response to this question, please provide information relating to (a) and (b) where relevant.

Response: Service providers should be transparent about the types of misuse or violations that would impact the visibility of content uploaded by a creator.

YouTube makes transparent in our Help Center information about how creators can react to and appeal enforcement actions taken by the platform.

If a creator's content was affected by an enforcement that they disagree with or think is a mistake, they have options to respond. The process for responding to an enforcement depends on what action was taken on the content. For example, the process Creators follow for appealing a Community Guidelines action is different from appealing an age restriction on their video.

There are a range of actions that can be taken on a piece of content - beyond Community Guidelines enforcement - which include legal enforcement, copyright enforcement, made for kids actions, and more. All of which are available publicly in detail.

(a) any known impacts of over-removal or erroneous removal of news publisher content, journalistic content or content of democratic importance

Whilst we appreciate the critical importance to news organisations of speed and virality in breaking stories, we equally have a responsibility to protect users from harm, irrespective of who uploaded that content.

(b) briefings, investigations, transparency reports, media investigations and research papers regarding misuse of such speech protective provisions

Response: Please see further details regarding removals at https://transparencyreport.google.com/youtube-policy/removals?hl=en_GB

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: No

For providers of online services

Question 21: What are the current complaints, counter-notice or other appeal processes for users to contest any action taken by you regarding journalistic content, news publisher content and content of democratic importance on your service?

In your response to this question, please provide information relating to (a) and (b) where relevant.

Response:

Currently, all Creators enjoy the same rights on YouTube regarding complaints and appeals, namely:

Appeal a Community Guidelines strike or video removal:

When we remove any content for a Community Guidelines violation, the creator may be issued a strike. Strikes are issued when content on YouTube is flagged for review, either by members of the YouTube community or our smart detection technology, and our review teams decide that it does not follow our Community Guidelines. If the channel gets a strike, the creator will get an email, notifications on mobile and desktop, and an alert in their channel settings the next time the creator signs in to YouTube.

The creator can appeal for 90 days after the warning or strike was issued.

After submission of an appeal

The creator gets an email from YouTube letting them know the result of their appeal request. One of the following will happen:

1. If we find that the content followed our Community Guidelines, we'll reinstate it and remove the strike from the channel. If the creator has appealed a warning and the appeal is granted, the next offence will be a warning.
2. If we find the content followed our Community Guidelines, but isn't appropriate for all audiences, we'll apply an age-restriction. If it's a video, it won't be visible to users who are signed out, are under 18 years of age, or have Restricted Mode turned on. If it's a custom thumbnail, it will be removed.
3. If we find that the content was in violation of our Community Guidelines, the strike will stay and the video will remain down from the site. There's no penalty for unsuccessful appeals.

The creator may appeal each strike.

(a) any initiatives taken to create and increase awareness about the process for users to complain and/or appeal content removals

Response: When a Creator receives a strike, they receive alerts via email, mobile and computer notifications, and/or in their channel settings. That notification makes clear

what content was removed, which policies it violated, how it affects their channel and what they can do next.

We understand that mistakes happen and the first violation is typically only a warning. To have this warning expire after 90 days, we offer Creators the opportunity to complete policy training. However, if their content violates the same policy within that 90-day window, the warning will not expire and their channel will be given a strike.

Policy training sessions are short in-product educational experiences based on the specific community guidelines policy that they have violated. If they violate a different policy after completing the training, they will receive another warning.

Sometimes a single case of severe abuse will result in channel termination without warning. Repeated violations of our policies – or a single case of severe abuse – may also still result in the termination of the account.

(b) any measures currently in place to prevent individual or systematic misuse of any protections for news publisher content, journalistic content or content of democratic importance.

Response:

The ways in which a platform like YouTube can safeguard misuse is by applying and enforcing policies (including appeals and redress) equally across the globe.

Currently, our guidelines are applied to all creators equally, and our classifiers and systems for reviewing content act agnostically. As previously described throughout, we then have processes in place to review content with a specific lens to the nuance of whether the content is educational, documentary, scientific, or artistic. And while our EDSA framework (see **question 16**) for evaluating the content applies equally, we often see it as a complementary framework by which news content and reporting is evaluated.

Misuse of protections and publisher privileges is certainly a concern for our platform, but equally so we are concerned by broad legislation and frameworks that look to privilege publisher content in a way that too could result in misuse. Broad definitions, lack of clarity, and regulatory overlap create a complicating environment for a platform like YouTube to operate in -- especially given any protections offered within one jurisdiction due to local laws are not universal, such that an obligation to not moderate news publisher content in the UK until an appeals process has been complete does not remove our obligation to users in other markets to apply our Community Guidelines fairly and consistently. In some jurisdictions (e.g. EU) there is recognition that users need to be protected from harm as a matter of priority which is why there are certain safeguards to ensure the provisions are more nuanced and allow platforms to address systemic risks first.

What is more, not all regulatory frameworks are created the same and there are real risks that jurisdictions looking to copy the Ofcom Codes may not be motivated by the same impartial objectives as our UK OSA regime.

Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No

Questions 22 – 24: Other information for journalistic content, news publisher content and content of democratic importance

For providers of online services

Question 22: Do you carry out any internal impact assessments to understand the freedom of expression and privacy implications of existing policies regarding journalistic content, news publisher content and content of democratic importance? In your response to this question, please provide information relating to (a) and (b) where relevant.
Response:
(a) explain which elements of your service design or operation they relate to and which factors they take into account
Response:
(b) provide relevant briefings, investigations, transparency reports, media investigations and research papers.
Response: As referenced in response above, please see transparency report details at https://transparencyreport.google.com/youtube-policy/removals?hl=en_GB
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No

Question 23: What, if any, measures are in place to ensure that protection of content of democratic importance applies in the same way to a wide diversity of political opinion? In your response to this question, please provide information relating to (a) where relevant.
YouTube's systems and signals apply equitably to all content uploaded to the platform and are not influenced by political opinion. As we've shared herein, our platform relies on a set of guidelines, human evaluators and dynamic signals to ensure the safety of the user. As we've stated, our policies determine what isn't allowed on YouTube and apply to all content - regardless of language or political viewpoint. In instances where our systems or raters take action on / enforce against content, they are specific to the individual piece of content, and not at the account level. This practice mitigates the risk of unintended bias by our systems and raters, as they evaluate individual pieces of content as needed, and do not place judgement on the channel or account itself.

Certain types of misleading or deceptive content with serious risk of egregious harm are not allowed on YouTube. This includes certain types of misinformation that can cause real-world harm, like certain types of technically manipulated content, and content interfering with democratic processes.

These policies prohibit certain types of content relating to free and fair democratic elections. We remind users not to post elections-related content on YouTube if it fits any of the descriptions noted below.

- Voter suppression: Content aiming to mislead voters about the time, place, means, or eligibility requirements for voting, or false claims that could materially discourage voting.
- Candidate eligibility: Content that advances false claims related to the technical eligibility requirements for current political candidates and sitting elected government officials to serve in office. Eligibility requirements considered are based on applicable national law, and include age, citizenship, or vital status.
- Incitement to interfere with democratic processes: Content encouraging others to interfere with democratic processes. This includes obstructing or interrupting voting procedures.

Our global team of reviewers combine with machine learning technology to apply these policies at scale, 24/7. Our Intelligence Desk has also been working for months to get ahead of emerging issues and trends that could affect the EU elections and any snap UK General Election, both on and off YouTube. This helps our enforcement teams to address these potential trends before they become larger issues.

(a) whether there are any additional measures/safeguards that are put in place during local or national elections.

Response:

Users may come to YouTube for news and information about elections near and far—seeking information on anything from where to vote, to learn what candidates are saying—which is why we believe we have a responsibility to support an informed citizenry and foster healthy political discourse. We've invested heavily in the policies, processes, and systems necessary to create a robust, multi-layered approach to support elections around the world and ensure that YT is a reliable source for timely news and information.

Central to this approach is our 4Rs responsibility framework (see question 16).

We have a set of misinformation policies specific to election contexts. For example, content that misleads voters on how to vote or encourages interference in the democratic process is not allowed on YT. In addition, policies across other areas may be applied to election-specific content—e.g., our harassment policy prohibits content that threatens poll workers. We work to swiftly detect and remove content that incites violence, encourages hatred, or promotes harmful conspiracy theories. We also

continuously review our policies, and in the context of elections, leverage our Intel Desk to monitor trends and stay ahead of potential threats to democratic processes that may appear on our platform.

Our policies apply to all forms of content, including elections — regardless of the political viewpoints expressed, the language the content is in, or how the content is generated. All users must abide by our content policies, from private citizens to the most visible public figures, and we rigorously enforce these policies.

See also our description of actions we take around EDSA content in response to question 16 above.

Our investments in connecting people to authoritative information on YT are more important than ever to help ensure YT continues to be a place where authoritative election news and information is front and center for voters. Connecting people to trustworthy, high-quality election news and information on YT is central to this effort. Our teams review our systems and processes to ensure we are elevating authoritative information as intended. When voters look for election-related information on YT, our systems prominently raise content and relevant context from authoritative sources in search results, the homepage, and the "watch next" panel. Through these systems, voters are connected to high-quality content from a variety of authoritative news sources, ranging in size, opinion and independence. During key civics and election moments, we make additional efforts to raise information from authoritative sources, which may include information panels that point users to authoritative sources about certain election-related topics.

Regardless of whether a country is in an active election season, YT continuously displays relevant information panels at the top of search results and under certain videos for topics prone to misinformation. We regularly evaluate and add topics where relevant and may roll out panels for topics that are specific to election topics, like methods of voting.

We've also invested heavily to strengthen transparency around AI-generated content. We recognize that bad actors may try to exploit Generative AI tools during an election — which is why we're ensuring that several layers of transparency and protections are in place, including current and upcoming elections:

- We now require creators to disclose whether the video they're uploading contains altered or synthetic content and is realistic.
- We will then add a transparency label to realistic synthetic or altered content to provide this important context to viewers. For election content, the label will appear on the video itself and in the video description.
- In certain cases, we may also add a label even when a creator hasn't disclosed it, especially if the use of altered or synthetic content has the potential to confuse or mislead viewers.

Coordinated influence operations are not allowed on YouTube, regardless of the political viewpoints they support. We work closely with Google's Threat Analysis Group (TAG) to

spot these types of campaigns on YouTube and terminate their channels. This can include attempts to interfere with elections. For example, as TAG shared, before the recent Portuguese elections, we terminated 7 YouTube channels linked to individuals in Argentina as part of our ongoing investigations into disinformation campaigns. Through TAG, we also share threat information with law enforcement agencies.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No

For all respondents

Question 24: What, if any, measures can online service providers put in place to ensure that protection of content of democratic importance applies in the same way to a wide diversity of political opinion? In your response to this question, please provide information relating to (a) where relevant.
Response: See our description of the application of EDSA exceptions in response to question 16 above.
(a) whether there are any additional measures/ safeguards that can be put in place during local or national elections
Response: Please see responses in **Question 23 and 23a** for YouTube's approach to local and national elections, including how we respond to harmful content, new policies to ensure transparency around AI generated content, and combatting disinformation campaigns.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No

Your response – User empowerment duties

Question 25: Detecting and moderating relevant content

For providers of online services

Question 25: What processes do you use to detect relevant content and how do you moderate it? In your response to this request, please provide information relating to (a) – (g) where relevant.
Response:

As the user empowerment duties apply to adult users in respect of content that is lawful it is critical to ensure that services have sufficient flexibility in complying with the duties. This will avoid the over-censorship of lawful content. Further, to the extent that services prohibit all categories of relevant content, they shouldn't be required to implement user empowerment tools.

Our Community Guidelines

By way of example, in respect of suicide, self-harm and eating disorder content, our Community Guidelines state:

> *YouTube users should not be afraid to speak openly about the topics of mental health, suicide, self-harm and eating disorders in a supportive and non-harmful way.*
>
> *However, there are times when content is created that is sensitive and may pose a risk for some users. When you create content that contains suicide, self-harm or eating-disorder-related topics, take into account the possible negative impact of your content on other users, especially minors and users who may be sensitive to this content.*
>
> *To protect and support your viewers and other users, please follow the Community Guidelines below when creating content related to suicide, self-harm or eating disorders. Not following these Community Guidelines may result in a strike, removal of your content or other restrictions to protect users.*

We then provide guidance on what can and cannot be posted - see here.

In respect of abuse and speech that incites hatred, we make clear in our Community Guidelines that such content is not allowed on YouTube - see here.

We have provided detailed guidance on the measures we take to identify and remove content which violates our Community Guidelines as part of our previous response to the Protecting people from illegal harms online consultation.

Measures in place on YouTube

On YouTube we already have a range of measures in place that allow adult users to regulate the content creators or content they see (that is permitted under Community Guidelines, but may be sensitive). In particular:

- We offer an optional setting (Restricted Mode) that helps to screen out potentially material content that some users may prefer not to view. This feature is designed to filter out graphic content that is permitted under Community Guidelines, but which is more appropriate for mature users.
- We offer several options within the platform that allow users to fine tune their recommendations. These include:
    - the ability for signed-in users to like or dislike a video (which may impact on the extent to which more of fewer similar videos of that type are recommended);

- the ability of users to notify us that they're not interested in a specific video which appears in their Watch Next feed;
- the ability for users to notify us 'not to recommend' content from the channel in question;
- the ability to delete individual videos from their search history;
- the ability to 'subscribe' to channels they are particularly interested in.

- There are a variety of tools for creators to allow them to control their interactions with other users. Viewers have the ability to block other viewer's messages on live chat, block specific users from commenting on their content, and as mentioned above, the ability to refine their recommendations to where they do not have to encounter creators whose content they do not wish to see.

- YouTube applies warning interstitials and crises pannels to content that, while not violative of our Community Guidelines, may be triggering for certain viewers or cause emotional distress, such as videos that reference suicide or self harm. We recently updated the warning interstitial for videos that reference suicide or self-harm content to more explicitly call out that these topics are present in the video to help protect users and enable their decisionmaking as to whether to view such content, as well as crisis panels that offer ways to contact local crisis resources in order to get help.

(a) what systems you use for detection

Response:

(b) further to the above, if there are any important features that you take into account to make distinctions between content, e.g. features that might identify a piece of content as promotional suicide material versus content intended to support users at risk of suicide

Response:

See the description of EDSA exceptions in response to question 16 above.

(c) where distinctions are made, the extent to which content is actioned automatically, by human moderation, through user reports, other methods or a combination of methods

Response:

In relation to YouTube, our automated systems use machine learning, which allows them to use data from previous human reviews to identify other potentially violative content. Most of our systems are continuously supplied with millions of data points from human reviews. This means our automated systems can offer a high level of accuracy in detecting violations. Automated systems also provide efficient response times to our users for the high volume of content that YouTube receives. When our systems have a high degree of confidence that content is violative, they may make an automated decision. However, in the majority of cases, our automated systems will simply flag content to a trained human reviewer for evaluation before any action is taken.

When a human reviewer checks potentially violative content, it means a trained human evaluates the content and makes a decision based on the relevant policy or law. If content is found to be violative, our human reviewers may remove content or age-restrict it if it's not appropriate for all audiences. If the content has an educational, documentary, scientific, or artistic purpose, we may allow it to remain on YouTube.

Please additionally see responses to **question 18** in our response form that Google submitted on February 23, 2024 for the "Protecting people from illegal harms online consultation".

(d) any insight into the cost of these processes, including set-up and on-going costs, in terms of employee time and any other material costs

Please see responses to question 17 in our response form that Google submitted on February 23, 2024 for the "Protecting people from illegal harms online consultation".

(e) whether relevant content is allowed or prohibited on your service

On YouTube, we have Community Guidelines that closely mirror and prohibit the types of harmful content considered as 'relevant content' under the Act, to which the user empowerment duties apply.

(f) whether you measure the incidence of users encountering such content, and if yes, whether these systems are different to those measuring other types of content, including illegal content

We measure rates of removals of what could be considered relevant content as part of our transparency reports. Please see further details at https://transparencyreport.google.com/youtube-policy/removals?hl=en_GB

In order to understand the user exposure to relevant content, we also track the percentage of views on YouTube that are generated by this content. We refer to this metric as the Violative View Rate or VVR. VVR is our North Star for measuring our progress on removing content that violates our Community Guidelines as quickly as possible.

In the fourth quarter of 2023, the VVR was 0.11% - 0.12%. In other words, for every 10,000 views on YouTube, only 11-12 were proved to be violative content. For further details of how we enforce our community guidelines, please refer to our responses in Question 25.

(g) if you offer users separate complaints procedures for moderated legal content versus illegal content, how often users report content through these channels, and what proportion of content is removed following a complaint

Response:

We do offer a separate complaints procedure for moderated legal content (community guidelines violations) as well as illegal content. Please see f our transparency reports for

Community Guidelines Removals as well as our Transparency report for Government requests to remove content https://transparencyreport.google.com/youtube-policy/removals?hl=en_GB & https://transparencyreport.google.com/government-removals/overview?hl=en_GB
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 26: Impact of relevant content

For all respondents

Question 26: Can you provide any evidence on whether the impact of relevant content differs between adults and children on user-to-user services? We are interested in particular in briefings, investigations, transparency reports, media investigations and research papers that provide more evidence.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 27 and 28: Experience of specific types of users

For all respondents

Question 27: Can you provide evidence around the types of adult users more likely to encounter relevant content, and the types of adult users more likely to be affected by such content?
Response: Google does not collect and process personal data in order to assess whether particular types of adult users are more likely to encounter relevant content and those which are more likely to be affected by such content. Safety is core to how we develop and operate our services, and we understand our responsibility to keep users safe while protecting their privacy and promoting the free flow of information. However, it should be recognised that it would be rare for a service to have this level of information about their user base demographics or a particular user's characteristics, and nor can this information be reliably inferred from user behaviour. For example, a user searching for information about Judaism or a Jewish festival, is not necessarily Jewish themselves.

Ofcom should allow services to use external publicly available data or insights from external experts and civil society organisations to inform their consideration of vulnerable users.

And while we can not provide evidence around the types of adult users more likely to encounter relevant content, relevant content is against our Community Guidelines and therefore is prohibited on the platform and we work tirelessly to minimise user exposure.

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: N

For all respondents

Question 28: How do you consider the experience of users who have a protected characteristic, or those considered to be vulnerable or likely to be particularly affected by certain types of content?

In your response to this request, please provide information relating to (a) – (c) where relevant.

Response:

We design all of our services to mitigate the risk of harm. For example, and as explained above, our services are designed so that they prioritise what appears to be the most helpful content on a given topic, and not to surface content that violates our content policies. This design approach works alongside our review and removal processes.

As an example, in developing our policies, tools and features around self-harm or suicide queries or content, we consult with both internal and external experts in psychology, mental health, and related areas. These include not only academics and clinicians, but also practitioners who provide direct services to vulnerable populations. We know how important it is to increase awareness around help-seeking behaviours, while decreasing risk-taking and reducing stigma. This issue is complex and requires highly specialised expertise, which is why we have a dedicated Health team, led by Google's Chief Health Officer, with whom we work closely to inform our product design.

All of our user-to-user products are likely to be subject to one or many sets of rules relating to user privacy, such as GDPR. These may include limits on the processing of personal data, particularly sensitive personal data. For example, the ICO has recently published guidance on the use of personal data in content moderation. This may impact both the categories of personal data we collect from end-users, as well as the ways in which we process this data.

On **YouTube:**

- We have specific functionalities to mitigate the risk of harm to children.

- As explained above, we also offer users an optional setting (Restricted Mode) that helps to screen out potentially mature content that some users may prefer not to view. This feature is designed to filter out graphic content that is permitted under Community Guidelines, but which is more appropriate for mature users.

- We anticipate problems before they emerge and adapt. Our Intelligence Desk monitors the news, social media and user reports from around the world to detect new trends, and works with the right teams to investigate and address them before they can become a larger issue.

- For viewers, all viewers are able to report content that goes against our Community Guidelines. For content that that viewers would prefer not to see, they are able to dislike specific videos as well as designate that they are not interested in order to see less of that type of content in their recommendations

- For creators, we have a variety of tools available for all creators to manage their community and comments on YouTube. For comment settings for videos, this allows for creators to pause comments or turn comments off for a specific video. At the channel level, users are able to hide comments from specific commenters across all videos on their channel, allow for approved users, as well as add a list of words and phrases that a user doesn't want to show up in their comments. Comments containing or closely matching these terms are then held for review.

(a) what criteria you use to determine whether a user is vulnerable or likely to be particularly affected by certain types of content, or if you do not categorise users as vulnerable and why

Response:

See response to Question, **25, 27, and subsection (a) above.**

(b) if your service collects any information about users that could be used to identify them as having a protected characteristic, vulnerable or likely to be particularly affected by certain types of content and, if so, what information you collect

Response:

(c) if you conduct any research into the experience of the above users on your service

Response:

YouTube leads ethnographic research with content creators in order to understand their everyday life, and to learn how we can build a better platform for all creators. We do this because creators are a fundamental part of our business, without them our platform would be empty. There'd be no content. So providing a safe space for their creativity is in

our best interest, not just so they can reach the viewership they seek but also because this then attracts advertisers who want their ads served alongside our creators' content.

Additionally, YouTube offers the Creator in Residence (CIR) program which has the goal of empowering diverse groups of creators to share direct feedback into how YouTube works for them (and how we can make things better). The program brings together 10 creators for a 6-month term, and each new class of residents meets weekly with different YouTube engineers, product managers and designers for a behind-the-scenes look at the development of upcoming projects. In return, the creators offer their unique perspectives. They stress test assumptions, question approaches, and weigh in on the user experience. And, as our teams incorporate their input over the months-long development of a project, the Creators in Residence end up playing an ongoing, indispensable role in shaping the direction of YouTube products.

At YouTube, we strive to create a platform where creators of all backgrounds can share their voices and find community. As we continue our commitment to supporting diverse creator communities, we include Black+, Women, and LGBTQIA+ creators in our CiR cohorts. Today, YouTube's engineering, product manager and designer teams think of creators as co-development partners in innovation, not just as end users of our product solutions. And diverse creator points-of-view will continue to be fundamental in the development process from concept to launch, allowing us to deliver the best possible experience to everyone.

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: N

Questions 29 and 30: Features employed to enable greater control over content

For all respondents

Question 29: What features exist to enable adult users to have greater control over the type of content they encounter?

In your response to this request, please provide information relating to (a) – (d) where relevant.

Response:

On **YouTube** we offer users an optional setting (Restricted Mode) that helps to screen out potentially material content that some users may prefer not to view. This feature is designed to filter out graphic content that is permitted under Community Guidelines, but which is more appropriate for mature users.

In addition, we offer several options within the platform that allow users to fine tune their recommendations. This includes:

- the ability for signed-in users to like or dislike a video (which may impact on the extent to which more of fewer similar videos of that type are recommended);

- the ability of users to notify us that they're not interested in a specific video which appears in their Watch Next feed;
- the ability for users to notify us 'not to recommend' content from the channel in question;
- the ability to delete individual videos from their search history;
- the ability to 'subscribe' to channels they are particularly interested in.

(a) features offered to users to reduce the likelihood of them encountering content they do not wish to see

Response: Please see above response to question 29.

(b) features offered to users to alert them to the presence of certain categories of content

Response:

YouTube applies warning interstitials and crises pannels to content that, while not violative of our Community Guidelines, may be triggering for certain viewers or cause emotional distress, such as videos that reference suicide or self harm. We recently updated the warning interstitial for videos that reference suicide or self-harm content to more explicitly call out that these topics are present in the video to help protect users and enable their decisionmaking as to whether to view such content, as well as crisis panels that offer ways to contact local crisis resources in order to get help.

(c) features offered to users to enable them to control their interactions with different types of users (e.g., non-verified)

Response: Users have a few ways to control their interactions with different types of users. As referenced in question 28, Youtube offers a variety of tools to creators to allow them to control their interactions with other users. In addition, viewers have the ability to block other viewer's messages on live chat, block specific users from commenting on their content, and as mentioned above, the ability to refine their recommendations to where they do not have to encounter creators whose content they do not wish to see.

(d) whether certain features are particularly valued or of use to users with protected characteristics, or by users likely to be affected by encountering relevant content

Response:

On Youtube, something that we have learned through our research is that creators like the ability to manage their comments and community. Creators use two main tools to do this, they 'heart' comments that they feel are aligned with what they expect from their audience, and they pin comments to the top of their feed as a way to signal to new commenters the kind of comment they value. And finally, creators want to exclude bad actors and keep their comments a safe space for their viewers. Something we've noticed is that creators prefer over moderation and that they tend to trust our automated systems, which include automatically holding comments for review dependent on creator set parameters.

We're constantly asking the creator community how they expect our features to work. For example, what should happen when blocking users. And creators overwhelmingly prefer blocked users to never be notified of this action. This because notifying a user that they've been blocked or their comment was removed only encourages more hate. Creators worry that if adversarial commenters knew they were blocked then they'd come after the creator or their audience with even more force, and that they'd try to see where the moderation line is between what gets published and what doesn't for each creator.

Creators like features that work as 'set it and forget it'. This means, they are happy to spend time upfront setting their preferences, building a blocked words list, and configuring our moderation tools, because that time investment then pays off every time they have a new upload. That small time investment upfront keeps their viewers safe each time, and that's a worthy investment because absolutely no one wants to spend significant time and effort on hateful and adversarial actors. Creators want to be creative, they want their videos to reach their audience, and they want to develop their creative video craft. Every minute they have to spend moderating comments and removing adversarial actors from their channel is time spent away from the creative mission that brought them here in the first place. We heard from groups who are historically underrepresented that they wanted more protection, so we created increased strictness, which holds more comments for review that are potentially inappropriate. Then Creators can decide whether they want these comments published to their community.

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response:

For providers of online services

Question 30: How do you design features to enable adult users to have greater control over the content they encounter, when are they offered to users, and what are the broader impacts on your system in deploying them? (For the purposes of our evidence base we are interested in features that enable control over a range of content, not solely **relevant content**).

In your response to this request, please provide information relating to (a) – (d xi) where relevant.

Response:

Please see our response to Question 29.

(a) how you measure and what evidence you can provide around the effectiveness of these features in terms of achieving their respective aims to prevent adults from encountering content that they do not want to see

(b) how you measure user engagement with these features, and any evidence you can provide around this
(c) how you ensure that these features are suitable for all adult users and that they're easy to access, including considerations for users with protected characteristics and/or vulnerable users
Response
(d) how you decide when to offer users these features, or how to present the use of these features to users. This includes but is not limited to the following aspects, i) – xi).
Features for recommendations are available to everyone, though users can choose to refine their recommendations or turn them off as mentioned in the response above.
i) how you develop the user need for these features, and the factors considered when determining to develop them
Response:
ii) whether these features are on by default, and in what circumstances
Response:
iii) whether these features are personalised for specific types of users
Response:
iv) when to offer users these features
Response:.
v) whether, when or how often to remind users of these features - this can mean reminding users to make an initial choice, or checking if a user wants to update the initial choice later on (and if so, how frequently)
Response:
vi) where users learn about these features
Response:
vii) how to provide information about these features, including the level of detail and the words used to describe complex or technical concepts
Response:
viii) whether users have choice of controls over specific types of content
Response:
ix) how you decide whether to iterate, replace or keep such features
Response:
x) any other factors not already covered above that you take into account when considering such features
Response:

xi) any insight into the cost of these features, including set-up and on-going costs (in terms of employee time and any other material costs) as well as any intended and unintended impacts on the service more broadly (e.g., the technical feasibility of implementing filter tools, or reducing functionality based on verification status).
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Your response – User identity verification duties

Question 31 and 32: Circumstances where user identity verification is offered and how

For all respondents

Question 31: What kind of user-to-user services currently deploy identity verification and in what circumstances? In your response to this request, please provide information relating to (a) – (c) where relevant.
Whilst we recognise that user verification can help combat illegal content in certain circumstances, we are also aware that mandating intrusive identity verification can have a chilling effect on free expression, particularly for vulnerable individuals. Therefore, we use identity verification when narrowly tailored to a specific purpose or where the only answer to the problem relies on establishing a user's identity. For example, when we suspect an account has been hijacked or we think one user is impersonating another user, we may ask for the user's identification for the limited purpose of determining whether the account is policy violative – that is whether the account is presently under the control of the appropriate user. On **YouTube** we operate a tiered approach to verification whereby users gain access to more advanced features (which could potentially be vectors for harm) through providing additional verification and/or having 'sufficient history' e.g. a sustained period complying with Community Guidelines . ● There are essentially four 'tiers' of access on YouTube, in brief: ○ Signed-Out Users - are only able to view content on YouTube; ○ 'Basic' Signed-In Users - Requires a Google account - Allows VOD upload / comments; ○ 'Intermediate' Signed-In USers - Requires phone verification - Allows access to intermediate features e.g. desktop live-streaming ○ 'Advanced' Signed-In Users - Requires ID/video verification OR sufficient channel history - Allows advanced features e.g. clickable links to 3rd party sites

• Beyond this, the focus should be on the actual behaviour of users. Hate and harassment is already violative of YouTube's policies, and platforms should be able to focus on combating this equally against both verified and unverified users.
(a) the ways in which these identity verification methods are beneficial, both to the user and to the service
Response:
(b) what documentation you understand to be necessary for different types, or levels, of identity verification on user-to-user services
Response: For a platform such as YouTube, which provides a voice to billions and, the impact of intrusive verification on freedom of expression can be significant. This is especially the case for marginalised communities, individuals without access to government documentation or those looking to find answers to difficult personal issues. Therefore, we would welcome platforms maintaining some level of agency in how and when they verify users.
(c) whether you believe there are there any other circumstances where identity verification should be offered on user-to-user services.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For providers of user-to-user services that provide some types of identity verification for individual adult users

Question 32: In respect of the identity verification method(s) used on your service, please share any information explaining:
(a) in what circumstances identity verification is offered on your service and why, and to which category/categories of users
(b) what evidence and steps are taken to verify the identity of a user, e.g., which attributes are checked, what aspects of verified users are known only to the provider and what aspects are made available for other users to see, including whether processes regarding adult users are different to those regarding children
Response:
(c) whether the process is, or can be, tailored to users in different geographical areas, such as the UK
Response:

(d) whether you engage third party providers to provide all or part of this identity verification process and, if so, which providers
Response:
e) once a user has their identity verified, what this allows them to do on your service, and if relevant, what activities this enables on another service
Response: See above
f) how your identity verification policies have been developed, including any research that you can share
Response:
g) any steps you take to ensure that identity verification is available to all adult users, including users who may not be able to access certain types of identity verification
Response:
h) any consideration around users who may be vulnerable participating in the identity verification method
Response: Important consideration. Consider e.g. Arab Spring. Or examples of how LGBT teens find support.
i) how you manage the identity verification of users who have multiple accounts
Response:
j) how you manage different identity verification methods operating simultaneously on your service, such as forms of age verification that require ID to complete the process, monetised schemes and notable user schemes, and how you consider user perceptions of these different methods
See **question 32** for details on verification and features access.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 33: Cost and effectiveness of these methods

For all respondents

Question 33: Please share any information about the costs and the effectiveness of identity verification methods In your response to this request, please provide information relating to: - (a) – (d) where relevant for all respondents, and - f) and g) where relevant for providers of user-to-user services that provide some types of identity verification for individual adult users.
Response:

(a) any insight into the cost of identity verification methods, including set-up and on-going costs, in terms of employee time and any other material costs, as well as any intended and unintended impacts on services more broadly
Response:
(b) how effective these identity verification methods are in verifying the identity of a user for the particular purpose for which verification is carried out
Response:
(c) any other benefits or unintended consequences from these schemes existing
Response:
(d) the safeguards necessary to ensure users' privacy is protected
Response:
For providers of user-to-user services that provide some types of identity verification for individual adult users
(e) any unintended consequences of implementing identity verification, such as the impact this may have on your site's ecosystem
Response:
(f) how you envisage your service operating in the digital identity market, bearing in mind moves towards cross-industry and federated identity schemes
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 34 and 35: User attitudes and demand for identity verification on user-to-user services

For all respondents

Question 34: What are user attitudes and demand for identity verification on user-to-user services? In your response to this request, please provide information relating to (a) – (d) where relevant.
Response:
(a) whether they value verification being offered on a service
Response:
(b) whether verification influences user behaviour, such as whether they perceive identity verification to signify authenticity
Response:

(c) attitudes towards non-verified, anonymous or pseudonymous users and the willingness to engage with them
Response:
(d) who you deem to be 'vulnerable' in terms of verifying their identity online – for example, whether this includes users unable to access or less likely to hold identification documentation, and those who may become vulnerable by displaying their identity to other users.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For providers of user-to-user services that provide some types of identity verification for individual adult users

Question 35: How do you measure engagement with your identity verification methods? In your response to this request, please provide information relating to (a) and (b) where relevant.
Response:
(a) take-up of identity verification by your users
Response:
(b) any insight into whether identity verification has any other effect on user behaviour, such as the content that users post and the amount that they engage with your service.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Your response – Fraudulent advertising

Questions 36 – 42: Overarching considerations

For all respondents

Question 36: Please provide evidence of the following:
(a) The most prevalent kinds of fraudulent advertising activity on user-to-user and search services (e.g. illegal financial promotions, misleading statements, malvertising)
[redacted]

(b) The harms associated with different kinds of fraudulent advertisements, the severity of such harms, and, if relevant, how this varies by user group

[redacted]s as

(c) The key challenges to successfully detecting different types of fraudulent paid-for advertising, and how these challenges can be minimised or resolved

[redacted]

(d) The prioritisation of suspected fraudulent advertising within all categories of harmful advertising queues, e.g. account verification, user reports, appeals

[redacted]

(e) The proportion of fraudulent advertisements that are currently estimated to remain undetected by services' systems.

Response:

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: **Yes**

Question 37: What technological developments aiding the prevention/detection of fraudulent advertisements do you anticipate in the coming years, and how costly and effective do you expect them to be? What are the challenges/barriers to their development?

Generative AI presents an opportunity to improve our enforcement efforts significantly. Our teams are embracing this transformative technology, specifically Large Language Models (LLMs), so that we can better keep people safe online.

LLMs are able to rapidly review and interpret content at a high volume, while also capturing important nuances within that content. These advanced reasoning

capabilities have already resulted in larger-scale and more precise enforcement decisions on some of our more complex policies. Take, for example, our policy against Unreliable Financial Claims which includes ads promoting get-rich-quick schemes. The bad actors behind these types of ads have grown more sophisticated. They adjust their tactics and tailor ads around new financial services or products, such as investment advice or digital currencies, to scam users. To be sure, traditional machine learning models are trained to detect these policy violations. Yet, the fast-paced and ever-changing nature of financial trends make it, at times, harder to differentiate between legitimate and fake services and quickly scale our automated enforcement systems to combat scams. LLMs are more capable of quickly recognizing new trends in financial services, identifying the patterns of bad actors who are abusing those trends and distinguishing a legitimate business from a get-rich-quick scam. This has helped our teams become even more nimble in confronting emerging threats of all kinds.

We continue to dedicate extensive resources and make significant investments in detection technology. The use of LLMs will require significant financial investment in ensuring that our platforms remain safe.

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response:

Question 38: If you have information/evidence/suggested mitigations to share which may be useful in the preparation of codes of practice, which is not covered by the questions above, please include these under 'Overarching considerations'.

Response:

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response:

For providers of online services

Question 39: What proportion of all paid-for advertising on your service is identified as fraudulent advertising?

Response:

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response: □ partly

Question 40: Does your service take any steps to warn users of the risk of encountering fraudulent advertising or to educate them about how to identify potentially fraudulent advertising?

Response:

User confidence in Google's products and services is essential. We want users to be empowered to make informed decisions about the ads they see online. Trust in advertisers on our platforms helps us deliver a smart and useful web experience for everyone. This means providing greater transparency about who our advertisers are, where they are located and which creatives they have served through Google.

Ads transparency comes in at two key points along the user journey.

- **As an ad serves**. We take measures to provide an "About this Ad" disclosure with meaningful information about that ad, including the verified legal name of who is behind the ad and what country they are from.
- **After an ad serves**. We enable advertiser accountability by allowing users to interact with our ads even after they have been served. This means having ad repositories where users can search for ads.

In 2022, we launched My Ads Center, which gives people more control over their ad experience on Google's sites and apps. Within My Ads Center, people can block sensitive ads and learn more about the information used to personalise their ad experience.

Last year we announced the Ads Transparency Center, a searchable hub of verified advertisers where users can view information about the advertiser and see the other ads they are running on our platforms. With the Ads Transparency Center, users will be able to understand:

- The ads an advertiser has run
- Which ads were shown in a certain region
- The last date an ad ran, and the format of the ad

More broadly, Google Safe Browsing warns users if it looks like a website is dangerous and is attempting to phish their credentials. Upon receiving this warning, people can simply click on the "Go back to safety" option to avoid going to a malicious site or downloading a malicious file. We've also updated our machine learning models to specifically identify pages that look like common log-in pages and messages that contain spear-phishing signals. Google makes this technology freely available to other browsers and internet companies and it is deployed in multiple, competing browsers in addition to Chrome (e.g. Firefox, Safari) and across many different platforms, including iOS and Android.

Google is a member of the Online Fraud Steering Group, which is convened by the Home Office , co-chaired by TechUK, UK Finance and the National Economic Crime Centre and attended by the tech platforms and many of the banks. Among other actions such as verification of financial services ads, together the tech members of the OFSG donated $1m in ad credits to UK Finance's "Take 5" campaign, their scam awareness campaign.

We regularly receive and action alerts via the UK Advertising Standards Authority (ASA)'s scam ads alert scheme and have given them ad credits so that they can run ad campaigns to boost awareness of this program amongst consumers.

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response:

Question 41: Please provide information regarding the proportion of successfully identified fraudulent advertisements that are identified via:

In 2023, we blocked or removed over 5.5 billion ads, slightly up from the prior year, and suspended 12.7 million advertiser accounts, nearly double from the previous year. Similarly, we work to protect advertisers and people by removing our ads from publisher pages and sites that violate our policies, such as sexually explicit content or dangerous products. In 2023, we blocked or restricted ads from serving on more than 2.1 billion publisher pages, up slightly from 2022. We are also getting better at tackling pervasive or egregious violations. We took broader site-level enforcement action on more than 395,000 publisher sites, up markedly from 2022.

In 2023 overall, we blocked or removed 206.5 million advertisements for violating our misrepresentation policy, which includes many scam tactics and 273.4 million advertisements for violating our financial services policy. We also blocked or removed over 1 billion advertisements for violating our policy against abusing the ad network, which includes promoting malware.

(a) automated systems

Response:

(b) human processes

Response:

(c) user reports
Response:
(d) other (please provide further detail).
Response: We do not publicly share specifics of our enforcement signals and systems, as bad actors would utilise this information to undermine our efforts and evade enforcement. We use a combination of human reviews and automated systems to enforce our policies and we are constantly monitoring our network for abuse. This combination of efforts has allowed us to match the scale of our adversaries and more efficiently remove multiple accounts associated with a single bad actor at once. As a result, between 2020 and 2022, we tripled the number of account-level suspensions for advertisers.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 42: What is the average and/or median time taken between the identification of a fraudulent advertisement and its removal/other actions taken? (If other actions taken, please specify what they are).
Response: • The great majority of ad creatives removed globally were removed pre-impression, meaning before they were seen by any of our users • For human-reported ads, we aim to review within 48 hours of them being reported, provided all necessary information is available (such as clickstring)
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 43: Proactive technology

For all respondents

Question 43: Please provide any evidence you have regarding proactive technologies which could be used to identify fraudulent advertising activity. In particular, we are interested in information related to the following points:
(a) The kinds of proactive technology which are/could be applied to identify or prevent fraudulent advertising
Response:
□

(b) A brief description of how these technologies are/could be integrated into the service
Response:
(c) The effectiveness, accuracy and lack of bias of such technology (including compared to alternative proactive and non-proactive methods) in relation to detecting fraudulent advertising and accounts which post fraudulent advertising material
Response:
(d) How proactive technologies are maintained and kept up to date
Response: We learn and refine our practices on an ongoing basis: after taking action on certain content, that content is then used to train our models to detect similar policy violations, and to inform new or evolved policies or product features.
e) Information related to the associated time and/or costs for set-up, operation, and human review
Response:
f) The cost of integrating such technologies: (a) for the first time; and (b) when updating these technologies over time
Response:
g) Whether there are cost savings associated with these technologies
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: □ partly

Question 44: Advertising onboarding and verification

For all respondents

Question 44: Please provide any evidence you have regarding the processes for advertiser onboarding and verification related to protections against fraudulent advertising. In your response, please indicate whether these processes are currently implemented in respect of services which are in scope of the Act or whether they stem from another sector In particular, we are interested in information related to the following points:
(a) The criteria which advertisers are verified against, including documentation/evidence used to support verification, and what advertisers are required to declare
Response: **Overview:** In 2020, we introduced an advertiser identity verification program that requires Google advertisers to verify information about their businesses, where they operate from and

what they're selling or promoting. Providing this transparency helps users learn more about the company behind a specific ad and also helps differentiate credible advertisers in the ecosystem while limiting the ability of bad actors to misrepresent themselves.

- Today, we are actively verifying advertisers in over 240 countries and regions. And if an advertiser fails to complete our verification program when prompted, the account is automatically restricted.
- Verification brings 3 key benefits:
  - First, we **better inform customers and enable them to make better decisions**
  - Second, we **increase trust** with greater transparency
  - And, third, the additional verification requirements help us **better detect bad actors** and limit their attempts to misrepresent themselves, preventing abuse before it happens

Advertiser verification:
To provide a safe and trustworthy ad ecosystem for users, Google requires advertisers to complete one or more verification programs. The advertiser verification program comprises multiple steps:

- **About your business:** In the first step of the Advertiser verification program, Google will ask advertisers a few basic questions related to their Google Ads account and business under the 'About your business' section. These questions will help us get a better understanding of your business. For example, we may ask whether your business is an advertising agency, who pays for your ads, whether you advertise your own products/services or some else's, and your specific industry (or industries).
- **Verify your identity:** Upon completion of the 'About your business' questions, you may be asked to verify your legal name via the advertiser identity verification or business operations verification process. This verification must be completed by an authorised representative, who is an admin of the Google Ads account and/or the payments profile paying for the ads. The steps and documents required for verification vary by country and are outlined in Google's Advertising Policies Help Center.
- **Verify your business operations:** Based on your responses in the About your business questions, we may ask you to verify details about your business operations (if applicable) along with supporting documentation, such as your business model, business registration information, types of services you offer, business practices, and relationships with advertised brands, products or third parties, if applicable.

In certain circumstances, such as if we suspect the advertiser's advertising or business practices may cause harm to users, we may pause Google Ads accounts immediately when the Advertiser verification program is initiated. This means that the

advertiser's ads will not be served until they are able to complete the program successfully. Once Google verifies the advertiser's information, we will show the name, location, and the ads they served over a certain time period on Google platforms (including Google Search and YouTube) in the Ads Transparency Center and in ad disclosures.

In order to advertise financial services in certain countries, including the UK, advertisers must undergo an additional layer of verification. For most advertisers, this process will entail demonstrating that they are authorised by their local regulator to promote their products and services through ads.

We believe that there is an important role for verification of advertisers' identity to protect consumers, especially where there is a higher risk of serious harm, and we prioritise the verification of accounts in abuse prone veriticals. In some cases, the advertiser verification process will help identify bad actors or otherwise block or deter would-be bad advertisers from reaching an audience.

We combine the advertiser verification process with technology to identify coordinated activity across accounts using signals in our network - like IP addresses, billing information and traffic patterns. This allows us to remove multiple accounts associated with a single bad actor at once. As a result of this combined approach, we suspended three times the number of accounts for violating our policies in 2022, than the year before.

Advertiser verification also provides a valuable signal that contributes to other safety-by-design product features. For example, in November, we launched Limited Ads Serving, which is designed to protect users by limiting the reach of advertisers with whom we are less familiar. The Limited Ads Serving policy introduces what could be described as a "getting to know you period" for advertisers in abuse prone sectors, which involves limiting visibility of certain ads based on trust scores. We now also use such a trust score based system from sellers on Google Shopping. This policy is specific to a certain set of ad-serving scenarios, and in these instances, only qualified advertisers will be able to serve ads without impression limits.

Verification and financial services ads in the UK

In order to show financial services ads of any kind in the UK - including showing ads to UK users who appear to be seeking financial services - advertisers need to be verified by Google. As part of the verification process, advertisers must demonstrate that they are authorised by the UK Financial Conduct Authority (UK FCA) or qualify for one of the exemptions. Advertisers of financial services in the UK must include in their Google Ads account a contact with the same email domain as the relevant FCA registered firm, and they must provide the following information using the form available in the Advertising Help Center.

- Google Ads account Customer ID
- Name of the Authorised Representative applying for verification
- Business details (name, address, email address)
- Your domain(s) or website(s) included in the UK Financial Conduct Authority registry and any of your other domain(s) or website(s) used for advertising on Google Ads which are not included in the UK FCA registry
- UK FCA registration number (FRN)
- Warranty that you will comply with any and all obligations that relate to the communication of, approval of, and restrictions on, financial promotions pursuant to applicable legal and regulatory requirements

(b) The role of (a) automated processing and (b) human processing in the verification process, and how they interact

Response:

(c) The costs associated with advertiser verification and how those costs vary as scale increases

Response:

(d) The percentage of advertiser accounts that are verified

Response:

e) Whether advertisers are permitted to publish advertisements on the service while the verification process is ongoing

Response:

Yes, in most cases, advertisers are permitted to publish advertisements on Google while the verification process is ongoing. They can continue running ads until the verification process is complete, even if they haven't initiated or finished it.

However, there are some exceptions:

- **False Information:** If an advertiser submits false information during the verification process or fails to meet Google's age requirements, their account will be suspended, and their ads will not serve.

Specific Circumstances: In certain cases, Google may immediately pause an advertiser's account when the verification program starts. This means their ads won't run until they successfully complete the program. These cases typically involve situations where there's a high risk of fraud or policy violation.

f) Whether there are additional/specific verification checks for advertisers placing adverts of certain kinds or targeting certain audiences, such as about specific products or services, or targeting users under the age of 18

Response:

Certification requirements which include the requirement to prove that local licensing requirements are included in a number of our policies, including: Gambling and games, Healthcare and medicines, and Financial products or services.

Additionally, we provide tools which enable advertisers in sensitive categories to target their ads away from vulnerable audiences and do not allow the advertising of sensitive products (e.g. gambling, alcohol) to minors.

We continually review our ad policies in light of industry changes and abuse trends, and may introduce new verification requirements for certain content.

g) Whether the verification of an advertiser account expires after a certain amount of time or certain activity, such as when advertisers make changes to their account or profile

Response:

Google's advertiser verification does not typically expire after a certain amount of time or specific account changes. Once an advertiser successfully completes the verification process, their status usually remains valid unless there are significant changes to their business operations or they engage in activities that violate Google's policies.

However, Google may re-evaluate an advertiser's verification status in certain situations:

- **Significant Business Changes:** If an advertiser undergoes major changes to their business, such as a change in legal name, address, or ownership, Google may require them to re-verify their information to ensure it's still accurate and up-to-date.
- **Policy Violations:** If an advertiser violates Google's advertising policies or engages in fraudulent or deceptive practices, their verification status may be revoked, and they may need to re-verify their account to regain access to advertising on Google's platforms.
- **Random Checks:** We may also conduct random checks on advertisers to ensure ongoing compliance with our policies and to maintain the integrity of their advertising ecosystem.

Is this response confidential? (if yes, please specify which part(s) are confidential)

Response:

Question 45: Service review of submitted advertisements/sponsored search results

For all respondents

Question 45: Please provide any evidence you have regarding the processes that services in scope of the Act have in place to review submitted paid-for advertisements and identify fraudulent advertising material. In particular, we are interested in information related to the following points:
(a) The percentage of submitted advertisements which are reviewed both (i) prior to and (ii) after publication
Response:
(b) The role (i) automated processing and (ii) human processing play in the review process and how they interact
Response:
(c) The red flags which trigger advertisement review processes both (i) prior to and (ii) after publication and the basis on which those red flags are selected
Response:
(d) The timescales for review
Response:
(e) What happens to the advertisement's visibility and reach, if it is flagged as suspected as being fraudulent (either by a user or automated system)
Response:
(f) The costs associated with the review of submitted paid-for advertisements
Response:
(g) Whether trusted flagger reporting is employed to inform services' review processes. If it is, how is it applied, what guidelines / criteria does it follow, and who are those trusted flaggers?
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 46: Advertiser appeals of verification/review decisions

For all respondents

Question 46: Please provide any evidence you have regarding advertiser appeals of verification/review decisions relating to fraudulent advertising on services in scope of the Act. In particular, we are interested in information related to the following points:

(a) The role of (i) automated processing and (ii) human processing in the appeals process, and how they interact;
Response:
(b) The level of proof required for an appeal to be accepted;
Response:
(c) The most frequent bases for appeals against sanctions decisions on fraudulent advertising content
Response:
(d) The ratio of decisions that are appealed against
Response:
(e) The costs associated with appeals
Response:
(f) The proportion of appealed decisions which are upheld and overturned
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 47: User reporting mechanisms

For all respondents

Question 47: Please provide any evidence you have regarding user reporting mechanisms for fraudulent advertising on services in scope of the Act. In particular, we are interested in information related to the following points:
(a) What user reporting tools there are for paid-for advertisements, and how these tools differ from those for user-generated content and/or search results and other search functionalities that are not paid-for advertising
Response: **Public reports** We make available four methods by which any user can report an ad that violates our policies: 1. Reports of policy violations can be made from the ad itself by clicking the AdChoices logo (on YouTube, ads can be reported using the (i) button displayed on the ad itself (example)) 2. Reports of policy violations can be made by completing this form. 3. Reports of policy violations can also be made within the Ads Transparency Center

(and see below for detail on the Ads Transparency Center)

4. Reports of legal issues (e.g. copyright, privacy, court orders, local law violations) can be made by completing this form (g.co/legal). Some background from the Legal Removals Team can be found here

Government and authorised user reports of legal issues

In addition to the above, we make available a webform to government entities and other authorised agencies by which they can report legal issues with ads that have come to their attention. Using this form for all government entities enables Google to quickly route the report to the appropriate team and ensures that we have all the information necessary to take appropriate action.

Note that this webform is strictly a private channel for government and authorised entities and it should not be shared outside of those, or to private individuals.

When completing the form, it's important that authorised users:

- Use their official email address
- Provide specific URLs and other information needed to identify illegal content (note that it's possible to submit multiple URLs in a single request and include attachments if necessary)
- Clearly identify the specific laws allegedly infringed upon by the content

We respond to completed forms with an email auto-response which contain a case number and the phrase "Your Request to Google" in the subject line.

(b) What percentage of user reports of advertisements relate to suspected fraudulent content, and the processes for taking action in relation to such reports
Response:
(c) Any statistics you can share on (i) the number of user reports of suspected fraudulent advertising received and resolved over a specific period and (b) the number of initial decisions appealed by users who made the report
Response:
(d) The criteria used to classify and prioritise user reports
Response:
(e) The median and/or average time it takes to respond to a user report, and any measures that are in place to ensure timely and accurate responses to user reports
Response:
(f) Any measures taken to make user reporting tools accessible, easy to use and easy to find for users
Response:

(g) How transparency and communication is maintained with users who have submitted reports
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 48: Use/involvement of third parties

For all respondents

Question 48: Please provide any evidence relevant to fraudulent advertising that you have, regarding the involvement and role of third parties in the provision of paid-for advertisements on services in scope of the Act. In line with the proportionality criteria under sections 38(5) and 39(5) of the Act, we welcome information related to how the involvement of third parties impacts the degree of control that services have over fraudulent advertising content. We also welcome information regarding contractual arrangements and how those arrangements are enforced.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 49: Generative AI and deepfakes

For all respondents

Question 49: Please provide any evidence you have regarding the impact of generative AI developments and deepfakes on the incidence and detection of fraudulent advertisements on services in scope of the Act. In particular, we are interested in information related to the following points:
(a) The frequency of deepfake fraudulent advertisements' occurrence, in absolute terms and/or as a proportion of all fraudulent advertisements, and how you expect this to evolve in the future
Response:
(b) What methodologies/technologies are currently employed to detect fraudulent advertisements which include deepfake or otherwise AI-generated content, and the effectiveness of these tools
Response:
(c) Whether detection technologies are developed in-house or acquired from a third-party, and how long it takes to develop and/or integrate those tools into wider systems

Response:
(d) The accuracy of detection methods, including true positive and false positive rates
Response:
(e) The costs associated with the development/acquisition and deployment of these detection mechanisms
Response:
(f) The types of deepfake or AI-generated content (in terms of either media type or subject) in fraudulent advertisements that are most difficult to detect i) via automated processes, ii) by human moderators, iii) by service users
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Your response – Access to information about a deceased child's use of a service

Questions 50 – 55: Processes for requesting information about a deceased child's use of a service

For all respondents

Question 50: What kinds of information might parents want to see about their child's use of the service?
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 51: How long should it take to receive information in response to a request?
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 52: What mechanisms could, or should services provide for parents to find out what they need to do to obtain information and updates in these circumstances?
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 53: What support or information do parents need to guide them through the process of making a request?
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For providers of online services

Question 54: What kinds of information do you provide and how do you provide this information? In your response to this request, please provide information relating to (a) where relevant.

Response:

Google takes children's privacy and safety seriously, and we offer a suite of controls for parents to monitor their children's online presence. We also understand the importance of digital legacies, and allow all users, whether children or adults, to name trusted contacts with whom their accounts could be shared in case they are no longer able to access their account.

Google has policies in place allowing immediate family members and representatives to request the closure of a deceased user's account, or a copy of the underlying content or data from the deceased user's account. We can work with immediate family members and representatives to close the account of a deceased person, where appropriate. In certain circumstances, we may be able to provide content from a deceased user's account. In all of these cases, our primary responsibility is to keep people's information secure, safe and private.

Family members and representatives can contact us using webforms regarding a deceased user's account. To produce stored content, we require:

- a scan of a government-issued ID or driver's licence in order to identify the requesting person,
- information about the relationship to the deceased person. We accept requests from Immediate family (spouse, sibling, child, parent) and/or Legal representative or executor and may ask for supporting documents.
- A scan of the death certificate.
- A US court order finding that the disclosure of stored content would be appropriate under applicable US laws. This requirement exists because US law generally prohibits service providers from producing stored content in accounts.

We review the information carefully to make sure that we identify the correct account. If we have any doubts, we may require further information or supporting documents from the requesting parties, e.g. prior email exchange with the account in question, including full headers of sender and recipient.

When a requester reaches the stage that they are eligible for a court order, Google provides a template court order and also offers a direct phone call with the requester to answer any questions they may have. Once a US Court Order is obtained and shared with Google, we expeditiously process the Order.

a) **If there are certain types of information you cannot provide, please explain why, for example whether there are technological, cost or privacy factors that mean certain kinds of information may not be feasible to provide**

Response:

When Google LLC receives data requests from government authorities outside of the US, we may provide user information if doing so is consistent with (1) US law, such as the Electronic Communications Privacy Act (ECPA) and Stored Communications Act (SCA); (2) laws of the requesting country; (3) international norms such as the Global Network Initiative's Principles on Freedom of Expression and Privacy and its associated

implementation guidelines; (4) and our own policies which include any applicable terms of service and privacy policies.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 55: How long does it typically take you to provide information in response to a request? In your response to this request, please provide information relating to (a) where relevant.
Response:
a) How long should it reasonably take services to provide information in these circumstances?
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Questions 56 and 57: Complaints systems

For all respondents

Question 56: What can providers of online services do to ensure the transparency, accessibility, ease of use and users' awareness of complaints mechanisms in relation to deceased user information request processes?
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For providers of online services

Question 57: Can you provide any evidence or information about the best practices for effective complaints mechanisms which could inform an approach to complaints about information request processes pertaining to a deceased user?
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 58: Evidence

For providers of online services

Question 58: What kinds of evidence do you require about the identity of the person making the request and their relationship to the deceased user? In your response to this request, please provide information relating to (a) and (b) where relevant.
Response:
(a) Do you, or would you, require different kinds of evidence in the event that the deceased user is a child?
Response:
(b) What evidence do, or would, you require that a user is deceased?
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: