Call for evidence response form

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: In this context it's useful to think of internet users as learners with different users preferring different learning styles. To be able to fully understand and access the T&C's, privacy policy and other public policy statements these need to be provided in an accessible format and in different mediums to allow users to access in their preferred way. In pilot studies with users accessing our services at the helpline, the variety in differing options to access information (text, audio, video) was highlighted. To start with, we would recommend that at the very least, these policies should be in a text format that is compliant with the reading age of the lower age limit for the platform concerned, (e.g. a reading age of 13 for 13+ platforms) an easy read format provided, an audio descriptor option and a video explainer.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No

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: As a charity that regularly receives cases from service users who are frustrated at the lack of explanation about reports they have submitted to platforms where no action has been taken, we feel there needs to be a higher level of transparency between service and reporter whatever the outcome of a report. This could easily still be automated but should include the area of policy that has been reported (e.g. this content has been removed for violating 'x' policy) and

that the person whose content is affected by a decision is also notified why. We would recommend that the direct subcategory of the policy concerned is linked to within the communication along with signposts to further support for the specific area of concern (e.g. if Fraud/ Scams – a signpost to the Cyber Helpline, if CSAM - a signpost to IWF/ StopItNow etc). This is another opportunity to provide holistic support to a user reporting harm online and with the plethora of services available in the UK to support with various types of online harm, this is an easy fine tuning of reporting mechanisms that helps platforms in scope to uphold their duty of care to safeguard users in the best way they can.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No

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

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.
Response:
(a) how you measure user engagement with/comprehension of your terms of service and the metrics you collect;

Response:
(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:
(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:

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.
Response: In our experience, visibility and ease of use are the most fundamental features that make a difference to the use of reporting and complaints mechanisms. Make them front and centre of the user experience, easy to use and real time prompts to use them, and this would see a significant uptake in reporting and understanding of when to use the mechanisms. Time and time again, when asking internet users why they do not report, the lack of engagement with online reporting on platforms in scope is almost always because a user 'didn't know how to report' or 'couldn't find how to report' Close behind are those users who do not have accounts with the service concerned where the reason is 'I had to have an account to report and I didn't want to create one' or 'reports by users who don't use the platform aren't looked at so what's the point'. Platforms in scope need to work harder to counter this narrative.
(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: We believe that any user should have the ability to report regardless of whether they have an account with the service. They shouldn't have to sign up to report. Many of the service users who access our helpline are the victims of nuanced and lengthy harassment/ stalking campaigns who choose not to have accounts on many of the platforms concerned as a way of managing their own wellbeing, this shouldn't prevent them from having the same access to recourse as those users who do engage with the platform with an account.
(b) how to ensure that reporting and complaints mechanisms are not misused
Response: **[CONFIDENTIAL]**
(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)
Response: We would recommend the use of real time prompts when harm is detected in the creation of content and when users are viewing. To provide this, platforms need to significantly upscale their resource investing much more in Trust & Safety and Moderation
(e) how to ensure they are user-friendly and accessible to all users (e.g., disabled users, children)
Response: As Question 1
(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: This is an important part of transparency and trust in platforms and should be a requirement that is upheld
(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: Again, this is a tricky one to balance with platforms understandably not wanting to give a guidebook to all their reporting process flows but simple flow charts for each reporting type to show the chain of response and the likely time would be a useful reference for users
(h) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: 6b

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: We would recommend that Ofcom look to MindGeek for their approach as a best practice example of handling this. Remove on report, investigate and reinstate if content found not to be in violation. This would prevent delays in time-sensitive situations where waiting for a response to a report that could take days may have devastating real world impact such as misinformation inciting wide-scale violence, hate speech threatening individuals and domestic abuse threats leading to physical assault. If this content was removed pending investigation, it significantly reduces the risk of harm. Currently there are no public KPI's that platforms are held to account regarding effective response times, this is one of the only areas in civil society where this is the case and we believe there should be a timeframe within which platforms in scope should have to respond to reports made with sanctions upheld where there are consistent failings to adhere to these, just as there would be in any other part of civil society.
(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: please see response under main question for our recommendation about what should happen.
(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: No

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

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

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:
(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:
(b) the factors that are or should be considered when taking action (e.g.: downranking/removal/suspension/ban or other) regarding this content
Response:
(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
Response:
(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
Response:
(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:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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:
(b) 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 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:
(a) any known impacts of over-removal or erroneous removal of news publisher content, journalistic content or content of democratic importance
Response:
(b) briefings, investigations, transparency reports, media investigations and research papers regarding misuse of such speech protective provisions
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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:
(a) any initiatives taken to create and increase awareness about the process for users to complain and/or appeal content removals
Response:
(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:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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

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

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

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: This is not a specific area of expertise for The Cyber Helpline however, based on the experience of service users accessing the helpline, we would emphasize the importance of labelling content here in the same way that misinformation has been labelled since the Covid 19 Pandemic.
(a) whether there are any additional measures/ safeguards that can be put in place during local or national elections
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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:
(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:
(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:

(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
Response:
(e) whether relevant content is allowed or prohibited on your service
Response:
(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
Response:
(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:
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: The Cyber Helpline is aware of the likelihood of adult users who have been defrauded being more susceptible to secondary attacks which can involve the algorithmic and in person recommendations of content that could be malicious in nature.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No

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: As the main entry to our service provides anonymity, any demographic information collected is very light tough and self-declared. Before engaging with our service users are prompted to accept our terms which explain the use of their data whilst engaging with our services.
(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: By the very nature of running a helpline, most, if not all users are vulnerable to online harm at the point of entry to our service. The Cyber Helpline deploys a screening set of questions to deem whether they are at any increased risk of harm or whether there are any further vulnerabilities they are reporting at the point of entry to the service. If there are any of these screening questions that give us cause for concern, safeguarding process is followed to ensure duty of care which in some instance may mean not supporting a particular service user where there is an imminent risk of harm that needs immediate attention before we can safely support a user.
(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: Gender, age and country of residence
(c) if you conduct any research into the experience of the above users on your service
Response: **[CONFIDENTIAL]**
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: 28c

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

(b) features offered to users to alert them to the presence of certain categories of content
Response:
(c) features offered to users to enable them to control their interactions with different types of users (e.g., non-verified)
Response:
(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:
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:
(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
Response:
(b) how you measure user engagement with these features, and any evidence you can provide around this
Response:
(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).
Response:
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.
Response:
(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:
(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
Response:
(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:
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:
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
Response:
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: The response from the public is varied with a division between privacy and safety. It is widely agreed that user to user services designed for users over a certain age should be able to properly verify a user's age upon sign-up but where this is 13 and a large portion of UK residents aged 13 are unlikely to have a form of government ID, this is tricky. Age estimation provides just that, an estimate of the age and where we are talking about the safety of younger users an estimation alone does not fulfil a platforms obligation to prevent harm to younger users. The issue of identity verification for those who identify in a different way to their official documentation risks causing further harm so, these processes would need to come with discretion and ways of verifying users where there have been name changes (e.g. people who identify as trans but have not had their name/ gender officially changed/ reassigned).
(a) whether they value verification being offered on a service
Response: With the trust in many platforms treatment of data having been breached several times over recent years and media reporting continuing to paint platforms in a negative light for the way they treat user data, it's not surprising that many users are reluctant to engage with verification processes.
(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: see first part of the response to this question
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)
Response: The most common type of defrauding we see on platforms in scope is where users who are friends have had their accounts hacked and fraudulent advertising is promoted through their account. Because of the trust they hold in their friend to post true and accurate information they are more likely to be susceptible to scams that are then advertised in this way and then followers of the account are likely to engage with the scam and be defrauded out of large amounts of money. This tends to take the form of investment and bitcoin scams. Hacked social media accounts are one of The Cyber Helpline's top attack types and, of those, over 50% involve scams of this nature. This has also been reflected in Action Fraud and NFIB data reflecting reports made to the police in this area over recent months. We would welcome a

further conversation with Ofcom where we could outline the sorts of cases we receive of this nature.
(b) The harms associated with different kinds of fraudulent advertisements, the severity of such harms, and, if relevant, how this varies by user group
Response: The Cyber Helpline runs impact surveys to assess the key areas of impact for the service users who have reported this type of harm and, despite mainstream media and government reporting focussing on financial losses, mental health and the impact of day-to-day life are often reported as being the most impacted areas in life for those who have experienced defrauding in this way.
(c) The key challenges to successfully detecting different types of fraudulent paid-for advertising, and how these challenges can be minimised or resolved
Response: The main issue here is about account recovery. By the time a service user gets in touch with us, they will have tried following the platform-concerned account recovery procedure to no avail. Most platforms in scope have account recovery processes, resulting in an end point where a user's email or phone number associated with the account has been changed. In these cases, they become stuck and are unable to take any further action, just having to watch as the hacker continues to post misinformation and scams posing as them. Very often, the only action they can take is to forewarn their friends and followers by other means of the perpetrator behaviour and encourage them to report. And then, once reports are made, this could result in account suspension meaning the original user then loses their entire account and content associated before ethe hacking. There needs to be a better way of recovering accounts that are flagged for this type of activity and teams at the platforms concerned dedicated to supporting users who have been defrauded because of engaging with the content.
(d) The prioritisation of suspected fraudulent advertising within all categories of harmful advertising queues, e.g. account verification, user reports, appeals
Response: To recommend best practice here we need to understand the current prioritisation of harm reporting within platforms in scope. We would recommend a working group involving industry and relevant members of civic society be set up to explore this and make recommendations about the best approach. This is something The Cyber Helpline, as a charity that supports those who encounter this type of cyber-enabled crime, would be keen to be a part of.
(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: No

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?
Response: There is no doubt that the advancement of AI will play a role here but, due to the ever-evolving human nature of this harm, there will always need to be a human element to the

response. The main challenge will be the same as with other common harms online, the perpetrator will no doubt always be one step ahead of enforcement and looking at the newest ways to exploit human nature for financial gain. To address this more research needs to be done and with quicker turnaround times to understand what the nature of this harm is on the ground now to better predict future evolution of these and pre-empt our responses to them.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No

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:

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:
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:
(a) automated systems
Response:
(b) human processes
Response:
(c) user reports
Response:

(d) other (please provide further detail).
Response:
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:
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:
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:

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:
(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:
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:
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:
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:
(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:
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:
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: