BILETA's Response to Ofcom's Call for Evidence on Online Safety Regulation, Additional Duties for Categorised Services; submitted on 20th May 2024.

This response is led by BILETA (British and Irish Law, Education and Technology Association). BILETA was formed in April 1986 to promote, develop, and communicate high-quality research and knowledge on technology law and policy to organisations, governments, professionals, students, and the public. BILETA also promotes the use of and research into technology at all stages of education. This response has been prepared by Reader Edina Harbinja, Dr Allison Holmes, and Dr Felipe Romero-Moreno on behalf of BILETA and approved by the BILETA Executive Committee.

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: The main features used to determine the clarity and accessibility of privacy policies can be used as a guide here.[1] The main features of clear and accessible terms include: the use of plain and simplified language, devoid of legal jargon and complex sentences. Simple, everyday language that can be easily understood by a broad audiences should be used. If the technical or legal terms are used, a glossary for those should be provided. In terms of structure, it is advisable to use clear, descriptive headings and subheadings to organise content. Bullet points or numbered lists for easier reading are also good practice. In addition, tables, charts, and infographics could be used to illustrate complex information and policies visually. It would also be advisable to consider tailoring the language to suit different levels of understanding, as well as using 'plain and intelligible language', especially where children are a target/likely audience, the terms should be written in a language/form that's intelligible to children. This also includes accommodating individuals with learning disabilities, ensuring the content is accessible to a wider audience. Additionally, also exploring the potential of incorporating graphics or graphic stories. Visual elements can enhance engagement and comprehension, making the material more appealing and easier to understand.

[1] EDPB, Guidelines 01/2022 on data subject rights - Right of access Version 2.0 Adopted on 28 March 2023; ICO, How should we draft our privacy information?, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/be-informed/how-should-we-draft-our-privacy-information/

Further, executive summaries could provide brief summaries at the beginning of each section or a general summary at the start of the document. Key points, obligations, and rights should be highlighted for quick reference. A frequently asked questions (FAQ) section is normally included to address common concerns and queries. A user-friendly search function within the ToS and policy documents would help users find specific information quickly. Readability Scores need to be noted, with an aim for a readability score appropriate for a wide audience, such as aiming for a Flesch-Kincaid grade level of around 8.

Wherever applicable, multiple formats, such as audio, video, and text should be used to cater for a wider range of diverse audience. Providers should also ensure documents are compatible with screen readers for visually impaired users. They should notify users of any updates or changes to the terms of service or policies promptly and maintain a version history so users can see what has changed over time. Larger service providers could offer tutorials or guides to help users understand the terms and how they apply. An accessible customer support for users who have questions or need clarification should be established, if not present already.

Providers should encourage and facilitate user feedback on the clarity and comprehensibility of the terms and use this feedback to continuously improve the documents.

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: To help users understand whether actions taken against content or users are justified under the terms of service by ensuring transparency, clear communication, and providing detailed explanations of the decision.

In their ToS, providers should clearly outline what constitutes a violation of terms of service. Specific examples of prohibited behaviours and content can help users understand the boundaries. When action is taken against content or a user, they should provide a clear and detailed explanation of why the action was taken, referencing the specific ToS violations to help users understand the rationale behind the decision. They should clearly explain how the appeal process works

and what information is needed for the review. As per above, it would also be advisable to consider tailoring the language to suit different levels of understanding, as well as using 'plain and intelligible language', especially where children are a target/likely audience, the terms should be written in a language/form that's intelligible to children. This also includes accommodating individuals with learning disabilities, ensuring the content is accessible to a wider audience. For information and support, they could provide examples or case studies that illustrate what constitutes a violation and what doesn't. In accordance with the Online Safety Act, providers will need to publish regular transparency reports and there, detail of the number and types of actions taken against content and users should be included. This should include statistics on the types of violations and outcomes of appeals.
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:
(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: Multiple access points are good practice. For registered users, providers should provide in-app (in-service) reporting tools, customer support portals, and direct contact options (email, chat). For non-registered users, they could offer public forms on the website and email contact points. For certain cases, providers should ensure that third parties (e.g., parents, guardians, personal representatives) can report issues even if they are not direct users of the service.
(b) how to ensure that reporting and complaints mechanisms are not misused
Response: To prevent misuse, verification and moderation mechanisms are advisable. Providers could implement automated and manual screening processes to filter out frivolous or malicious reports, making sure these are robust and open to human inspection if automated. They could also carefully restrict the number of complaints a single user can file within a certain period to prevent spamming. This should not be set out strictly so to prevent users from lodging complaints.
(c) the key choices and factors involved in designing these mechanisms
Response: User-centric design should be employed. Providers should use simplified interfaces with straightforward and intuitive designs for complaint forms. They should also provide clear instructions and step-by-step guidance on how to file a report, what information is needed, and what to expect during the process.
(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: Ideally, providers should create dedicated channels for different issues (e.g., reporting users, inappropriate content, appealing decisions and bans). They should offer FAQs, guides, and support agents to assist users in navigating the reporting process.
(e) how to ensure they are user-friendly and accessible to all users (e.g., disabled users, children)
Response: We talked about accessibility under a) as well, but in addition to language, inclusive design with accessibility features should be used (e.g. compatibility with screen readers, provide text alternatives for images). For child

users, where appropriate, child-friendly interfaces that are easy for children to use, with parental controls and guidance where necessary should be implemented. It would also be advisable to consider tailoring the language to suit different levels of understanding. This includes accommodating individuals with learning disabilities, ensuring the content is accessible to a wider audience. Additionally, also exploring the potential of incorporating graphics or graphic stories. Visual elements can enhance engagement and comprehension, making the material more appealing and easier to understand.

(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: Providers should ensure confidentiality and anonymous reporting options. They should clearly inform users whether their reports will be anonymous to protect their identity from other users.

(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: Providers should maintain educational resources, including help centres with articles, tutorials, and videos explaining the reporting and complaints process. They should keep support materials current and provide examples of how complaints are handled and what the outcomes might be.

(h) any other information.

Response: As noted above, providers should collect feedback on the reporting process to identify areas for improvement and ensure the system evolves with user needs. Regular transparency reports should include the number and types of complaints received and how they were resolved to build trust and transparency.

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

Response: No.

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: Content moderation can be significantly enhanced by focusing on transparency, human oversight, advanced technology, and user empowerment. Platforms should prioritize clear, concise terms of service with detailed explanations for content removal, coupled with regular transparency reports to foster trust and accountability. Investing in well-trained human moderators can ensure nuanced decisions and reduce errors, while establishing clear appeal processes safeguards user rights. Advanced technological solutions, such as AI-powered moderation with contextual analysis and Natural Language Processing (NLP), can efficiently identify harmful content while minimizing false positives. Empowering users with granular content controls and encouraging reporting and feedback further improves moderation accuracy and responsiveness. Additionally, fostering community moderation can instil a sense of shared responsibility. While these improvements hold promise, they do come with cost considerations. Human moderation and the development of advanced technological tools require significant investment. However, striking the right balance between user safety, human rights such as, freedom of expression, privacy, data protection, non-discrimination, and due process, along with cost-effectiveness is crucial for creating a safer and more inclusive online environment. It necessitates ongoing evaluation and adaptation to meet the evolving challenges of content moderation
(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: Existing content moderation systems demonstrate a mixed track record regarding accuracy, bias, and effectiveness. While automated systems, particularly

those utilizing AI and machine learning, have effectively identified harmful content like child sexual abuse material (CSAM) and terrorist propaganda[2] on platforms such as Facebook, they often struggle with nuanced content requiring contextual understanding, leading to over-removal and false positives.[3] Biases in moderation algorithms have been revealed, disproportionately affecting marginalized groups and minorities, as seen in the case of content from Black and LGBTQ+ users being disproportionately flagged.[4] Platforms like Facebook and YouTube have made strides in removing harmful content through AI and human moderation but face challenges with inconsistencies in enforcement and potential amplification of harmful content through algorithms. Twitter grapples with addressing harassment and hate speech due to reliance on user reporting and limited resources.[5]
(c) any other information.
Response: while existing moderation systems have demonstrated some effectiveness in reducing harmful content, there is significant room for improvement. Addressing issues of accuracy, bias, and scalability requires ongoing research, investment in technology, and a commitment to transparency and accountability from platforms.
Is this response confidential? (if yes, please specify which part(s) are confidential) No.
Response: No.

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:

[2] Viscount Camrose, Parliamentary Under Secretary of State, Department for Science, Innovation & Technology—further supplementary written evidence (LLM0120) House of Lords Communications and Digital Select Committee inquiry: Large language models https://committees.parliament.uk/writtenevidence/127855/html/

[3] Ofcom, ''Use of AI in Online Content Moderation'' (Ofcom, 2019) https://www.ofcom.org.uk/research-and-data/online-research/online-content-moderation

[4] The Organization of American States (OAS) criticized Facebook for not adequately considering marginalized groups in its handling of an Arabic language post intended to reclaim hurtful language used against the LGBTQ+ community. OB, Reclaiming Arabic Words, 2022-003-IG-UA.

[5] Endres, Dorothea, Luisa Hedler, and Kebene Wodajo. "Bias in Social Media Content Management: What Do Human Rights Have to Do with It?" American Journal of International Law 117 (2023): 139-144. DOI: https://doi.org/10.1017/aju.2023.23

(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: The European Court of Human Rights (ECtHR) jurisprudence emphasizes protecting public interest content that contributes to democratic discourse. Journalistic content encompasses various formats produced by journalists and media outlets, including news, opinions, and investigations, while

considering journalistic methods like source protection. Content of democratic importance extends beyond political news to social, cultural, and scientific topics that hold power accountable. Restrictions on such content must be proportionate to legitimate aims, ensuring minimal interference with freedom of expression.[6] These principles guide the formulation of definitions in terms of service to ensure users understand what content falls under these categories.
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.
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

[6] See for instance ECtHR, Council of Europe, 'Factsheet - Protection of journalistic sources' (2022) https://www.echr.coe.int/documents/d/echr/fs_journalistic_sources_eng

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: effective redress for complaints regarding journalistic, news publisher, and democratic content balances freedom of expression with accountability. Internal review processes with dedicated teams or individuals are crucial for initial assessment, investigation, and potential corrections.[7] If unresolved, external mediation by independent organizations or ombudsman services can facilitate communication and resolution.[8] For complex cases or those of significant public interest, independent review boards with diverse expertise can conduct thorough

[7] See for instance YouTube, 'What is FOA internal review'? (2024) https://www.youtube.com/watch?v=eNXqTXkdQMw

[8] See for example the International Ombuds Association site at https://www.ombudsassociation.org/what-is-an-ombuds-

investigations and issue decisions.[9] Self-regulatory bodies can establish industry standards and offer an avenue for complaints resolution.[10] Legal recourse remains a last resort, with legal frameworks prioritizing freedom of expression. These mechanisms, tailored to specific content types, like journalistic content, news publisher content and content of democratic importance create a robust system for addressing complaints while upholding democratic discourse.

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

Response: Multiple resources offer insights on complaints and appeals in the media industry. The Reuters Institute for the Study of Journalism publishes research on media accountability and best practices.[11] The Columbia Journalism Review provides in-depth investigations on journalistic ethics and complaints processes.[12] The OSCE Representative on Freedom of the Media issues reports on media freedom violations and the need for effective complaints mechanisms.[13] The Council of Europe produces reports on media freedom and self-regulatory bodies.[14] Transparency reports from major platforms offer data on content moderation and complaints handling.[15] Academic studies explore the effectiveness of various complaints mechanisms and their impact on public trust.[16] These diverse resources can inform the development and implementation of effective redress mechanisms that promote accountability and protect freedom of expression.

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:

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

[9] See for example Meta's Oversight Board at https://www.oversightboard.com/

[10] El Pais, ''The EU favours self-regulation in new AI law'' (December 2023) https://english.elpais.com/technology/2023-12-05/the-eu-favors-self-regulation-in-new-ai-law.html#

[11] See generally https://reutersinstitute.politics.ox.ac.uk/

[12] See generally https://www.cjr.org/

[13] See generally https://www.osce.org/representative-on-freedom-of-media

[14] See for instance ECtHR, Council of Europe, 'Factsheet - Protection of journalistic sources' (2022) https://www.echr.coe.int/documents/d/echr/fs_journalistic_sources_eng

[15] See for example Google Transparency Report https://transparencyreport.google.com/?hl=en

[16] See for instance Transparency International, ''Complaint mechanisms reference guide for good practice'' (2016) https://knowledgehub.transparency.org/assets/uploads/kproducts/ti_document_-_guide_complaint_mechanisms_final.pdf

Response: Over-removal or mistaken removal of news, journalistic, or democratically important content by service providers has far-reaching consequences. It stifles free speech, limits access to diverse viewpoints, and undermines informed decision-making.[17] This erosion of public trust in platforms as reliable information sources can fuel misinformation. News outlets and journalists suffer financially, threatening independent journalism and potentially concentrating media power. Accusations of bias and manipulation arise, further polarizing society. Legal challenges and liability for platforms may ensue. Marginalized communities, whose content is often disproportionately targeted, are further silenced, amplifying existing inequalities.[18] To counteract these detrimental effects, platforms must adopt transparent content moderation, invest in human review, and ensure fair appeals processes.

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

Response: Various resources shed light on the misuse of speech protective provisions. The Center for Democracy & Technology (CDT) investigates how governments and platforms misuse these provisions to suppress legitimate expression.[19] The Electronic Frontier Foundation (EFF) focuses on the misuse of legal mechanisms to silence critical speech, providing case studies and legal analysis.[20] Article 19 produces reports on the global misuse of speech laws, especially in cases of political censorship.[21] Transparency reports by services offer some insights into their handling of complaints related to speech protective provisions.[22] Relying on media investigations, investigative journalists and news organizations uncover instances where these provisions are weaponized to silence critics.[23] Academic research delves into the theoretical and empirical aspects of misuse.[24] These resources collectively illuminate the potential negative impacts on free expression and democracy.

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

[17] See for instance ECtHR, Council of Europe, ''Factsheet – Access to the Internet and freedom to receive and impart information and ideas'' (Sept 2022) https://www.echr.coe.int/documents/d/echr/FS_Access_Internet_ENG

[18] The Organization of American States (OAS) criticized Facebook for not adequately considering marginalized groups in its handling of an Arabic language post intended to reclaim hurtful language used against the LGBTQ+ community. OB, Reclaiming Arabic Words, 2022-003-IG-UA.

[19] See generally https://cdt.org/

[20] See generally https://www.eff.org/

[21] See generally https://www.article19.org/

[22] See for example Google Transparency Report https://transparencyreport.google.com/?hl=en

[23] See for instance Thomson Reuters Foundation, ''Weaponizing the Law: Attacks on Media Freedom'' (2023) https://www.trust.org/documents/weaponizing-law-attacks-media-freedom-report-2023.pdf

[24] Ibid.

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

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

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:
(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)
Response: Recent figures demonstrate that online communication channels are the way in which fraud and scams are most often experienced, with online advertisements accounting for 11% of scams and fraud experienced.[25] While a precise breakdown of the kinds of fraudulent activity is not available there are several notable trends concerning fraudulent advertising. Impersonation scams featuring celebrities are prevalent across platforms. Images and videos of celebrities are misused, appearing in ads for provides which they do not endorse. A related issue is the rise in deepfake videos of celebrities which are appearing in adverts promoting products. These adverts frequently focus on cryptocurrencies, investment platforms, and health and wellbeing. An increase in fraudulent adverts

[25] *Survey on "Scams and Fraud Experienced by Consumers" Final Report* (European Commission, January 2020) < https://commission.europa.eu/document/download/2667f3c9-d72a-499d-9f13-2f0942699b8d_en?filename=survey_on_scams_and_fraud_experienced_by_consumers_-_final_report.pdf>.

can be seen at particular points of the year where individuals are likely to be engaging in higher than normal retail such as black Friday.

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

Fraudulent advertisements bring with them a range of harms to users of online services. It is worth noting that the data which exists on cyber fraud is not limited to fraudulent advertising but rather captures a wide range of cyber enabled and cyber dependent fraudulent activities. As such, only a general overview of the impacts of cyber harm can be provided.

Financial harms are the most obvious consequence of fraudulent online scams. Data indicates that the median financial loss for victims of cyber fraud was £95.[26] For some victims reported losses can greatly exceed this amount, often going into the thousands of pounds.

Fraudulent advertising can result in harms beyond financial loss. Victims of scams report emotional impacts including a loss of confidence and self blame.[27] Victims also suffer from a loss of time and inconvenience. Victims of cyber fraud may similarly adapt their behavioural patterns, disengaging with particular internet sites or platforms due to the impact of the fraud. More serious impacts can also be felt, including damage to relationships, the need to take time off work, the desire to avoid social situations and potentially the loss of employment.[28]

The impacts of fraud are likely to be under-reported as there is a demonstrated reticence in reporting online fraud to law enforcement due to the potential negative perceptions associated with being a victim of online fraud. The shame and emotional impacts of fraud can also represent an impediment to effective reporting. On average only a fifth of those who experience fraud report it to an official authority.[29] Difficulties in reporting and significant gaps within the data also impact on the ability to determine whether there are significant demographic differences in the victims of fraud. There is no set profile for victims of online fraud.[30] However, research has shown that people with mental health problems are three times more likely to have been the victims of an online scam compared to the wider population.[31] The manner through which fraud is experienced also varies by socio-demographic groups. In the context of online advertisement fraud, the largest discrepancies are within age range with users aged 18-34 the most

[26] Office of National Statistics, 'Crime Survey for England and Wales: year ending Dec 2023' (24 April 2024) <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingdecember2023#fraud>

[27] Mark Button, Chris Lewis, & Jack Tapley, *Fraud typologies and victims of fraud* (National Fraud Authority, 2009) 26.

[28] Office of National Statistics, 'Crime Survey for England and Wales: year ending Dec 2023' (24 April 2024) <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingdecember2023#fraud>

[29] *Survey on "Scams and Fraud Experienced by Consumers" Final Report* (European Commission, January 2020) < https://commission.europa.eu/document/download/2667f3c9-d72a-499d-9f13-2f0942699b8d_en?filename=survey_on_scams_and_fraud_experienced_by_consumers_-_final_report.pdf>.

[30] Gareth Norris and Alexandra Brookes, 'Personality, emotion and individual differences in response to online fraud' (2021) 169 Personality and Individual Differences 2.

[31] Merlyn Holkar & Chris Lees, 'Caught in the Web: Online Scams and Mental Health' (Money and Mental Health Policy Institute, Dec 2020) < https://www.moneyandmentalhealth.org/publications/online-scams/>

likely to encounter online advertisement fraud.[32] Other demographic factors such as gender and education level have a lesser effect on who experiences fraud.
(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:
(c) The key challenges to successfully detecting different types of fraudulent paid-for advertising, and how these challenges can be minimised or resolved
Response: A key difficulty in detecting fraudulent advertising rests with the reporting mechanisms. Complaints about particular fraudulent advertising are often not made by the most vulnerable groups or harmed consumers but rather by those who are less likely to be caught by the scams. Those who report tend to be consumers with relevant information or knowledge, or who are unlikely to be mislead themselves.[33] The lack of consumer awareness concerning fraudulent advertising represents a challenge to detecting fraud in the current user reporting system. Consumers may similarly face difficulties understanding how to report a fraudulent advertisement. Design interfaces within platforms can require users to take multiple steps or access a different website to report the advert. Such mechanisms may dissuade users from reporting content. User interfaces need to be designed in a clear and straightforward manner. Users should not have to click through multiple links to report the fraudulent ad. Some platforms do not let users report the ad directly and instead require that they go to a specific website to report the issue. This creates barriers to reporting. Any attempts to detect of prevent fraudulent advertising can also be limited due to the high levels of adaptability to new technologies and societal developments demonstrated by offenders. Fraudsters will often take cues from recent societal developments to exploit individuals. If relying on automated systems to remove such content, the system needs to similarly adapt and keep up with contemporary developments.
(d) The prioritisation of suspected fraudulent advertising within all categories of harmful advertising queues, e.g. account verification, user reports, appeals
Response:
(e) The proportion of fraudulent advertisements that are currently estimated to remain undetected by services' systems.
Response: It is not possible to determine the proportion of fraudulent advertisements that are not detected.

[32] Ibid note 5.

[33] Silvia Milano et al, 'Epistemic fragmentation poess a threat to the governance of online targeting' (2021) 3 Nature Maching Intelligence 466.

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: Machine learning technologies are already employed in the detection of fraudulent content and their use is only likely to increase in the coming years. There are a number of challenges to the successful creation and deployment of technologies to address fraudulent advertisements. In training machine learning algorithms, which are effective at detecting fraudulent adverts, the technology should be trained on both fraudulent materials and the decision-making process of human operators in determining whether content met the threshold of 'fraudulent'. To obtain a useful data set, the adverts which have been removed must be effectively evaluated and data should be shared across platforms. Fraud can have a significant impact even if the advert is only available for a short period of time. The technologies must be responsive and able to act quickly to remove the content. One of the most significant impediments to effective fraud detection is the impact of individualised customer behaviour patterns. Fraudsters will emulate legitimate businesses. Customers may not be aware of the fraudulent nature of the transactions until a significant amount of time has passed. The more the customer interacts with the fraudulent business, the more the malicious actor can learn and adapt to the customers' behaviour. Delays in reporting and taking down the content will worsen this.
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: Targeted advertising capabilities have increased dramatically over recent years. This means that advertisers are better placed to target specific sites, placements, or use contextual keyword targeting to find the audience most

receptive to their content. In the context of fraudulent advertisers, these technologies strengthen the ability of these ads to reach receptive markets. The way in which advertising is offered on the basis of personal interest means that the ads themselves are not seen by wide groups but rather much narrower collections of individuals who are more likely to be taken in by them. Using keywords such as 'mortgage', 'investment' etc. mean that fraudsters can find individuals who are looking into similar services and may be less likely to question the truth of the ads which are put to them. Online platforms have enabled this micro-targeting of individuals through their expansive data gathering practices. Where micro-targeting and affinity groups are likely to put individuals at increased risk of exploitation platforms should take additional steps. Companies should be restricted from utilising targeting techniques which allow advertisers to appeal to individuals' vulnerabilities. To prevent this, online platforms should be prevented from presenting advertisements based on profiling using special categories of data.[34] Such a limitation would mirror obligations placed on providers under the EU Digital Services Act.

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

Response: No.

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:

[34] Special categories of data as defined Article 9 of the GDPR Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1.

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: Different technologies have been identified to proactively review content. Currently, large platforms employ a range of technologies to identify violations. These include employing rate limits which assess how quickly content is being created with the aim of detecting the usage of bots. Platforms may also employ matching technology which identifies identical or near identical copies of content which was previously determined to violate the platforms policies. Finally artificial intelligence can be employed. These technologies could similarly be

applied to identify fraudulent advertising, however their effectiveness at doing so will likely rely on the quality of the data utilised to train the system as well as ex post review by a human operator to check the appropriateness of the decision.

(b) A brief description of how these technologies are/could be integrated into the service

Response: One such example would be the limitation and prevention of impersonation. In online scams, an account may impersonate a celebrity or well-known expert in a particular field. Images of this individual may be used in the advertisements for a variety of services. For example, the image of Martin Lewis, founder of MoneySavingExpert.com has been used by a variety of fraudulent actors to mislead victims about their products and falsely imply that the product has received his endorsement.[35] Such celebrity impersonations can lead individuals to invest or engage with scams where they might otherwise have not done so. Multiple platforms including Meta and X have already created technologies which can address the issues of impersonation on their platforms. Marketed as 'impersonation defence' systems these are offered to advertisers on paid subscription models. These systems monitor accounts for changes including display names, profile photos, and usernames and these accounts are flagged for further review if the impersonation is detected. As the technology already exists within the service, albeit in a paid model, it could similarly be utilised across the platform.

(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: There are limitations to the ability of existing technology in proactively detecting fraudulent content. A significant impediment to the detection of this content relates to the lack of contextual understanding of the content of the advertisement.

(d) How proactive technologies are maintained and kept up to date

Response: A human should remain in the review process throughout the lifecycle of the technology. Best practice would have a human reviewer to evaluate the content and compare decisions against those from the machine learning technology. The review team should manually label the decisions and ensure that accurate data is fed back into the system, creating an effective feedback loop for

[35] Martin Lewis, Joint Committee on the Draft Online Safety Bill Evidence Session 5 (18 October 2021).

the algorithm. There should also be a clear audit trail of the decisions made by the human operator and of the information that was fed back into the system.
e) Information related to the associated time and/or costs for set-up, operation, and human review
Response: Prior to the implementation of the technology there should be a thorough review to determine the effectiveness and accuracy of the product. This human review should be an ongoing process throughout the life cycle of the product.
f) The cost of integrating such technologies: (a) for the first time; and (b) when updating these technologies over time
Response: The development and implementation of technologies necessarily carries with it financial costs. These costs will be ongoing throughout the deployment of the technology. The technologies need to be subject to continuous review requiring both human and technical resources.
g) Whether there are cost savings associated with these technologies
Response: There is insufficient publicly available information provided on the costs of these technologies to determine the financial benefit. Data provided is not disaggregated relevant to specific technologies.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No.

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: The current way in which advertisers are verified differs across platforms and this is emboldened by the lack of any statutory requirements for verification of advertisers prior to publication of their adverts. The piecemeal approach means that different providers institute different policies for verifying advertisers. Google for example requires advertisers to have a Gmail account to create adverts. Subsequent verification can occur wherein the business is asked to verify their business, name, and location. However, not all advertisers will be required to complete the verification programme. Those which are selected to be verified will be informed and once informed will have 30 days to initiate the verification and, once initiated, a subsequent 30 days to complete the verification. Other large user to user platforms has similarly piecemeal approaches. Where verification does occur within platforms it often is limited to basic information such as associated name and URL. This leads to a lack of rigour in the verification process. When asked to complete additional steps for verification this may involve providing documentation as to the companies registration and in some instances official government identification. However, this is not a mandatory requirement for all potential advertisers.

It is worth noting that the provisions of the Digital Services Act in the EU have instituted requirements that advertisers provide the natural and/or legal person on behalf of whom the advertisement is presented and, if different, paid for.[36] Such provision would be beneficial in the United Kingdom.

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

Response: Currently automated processing may be involved in the initial screening of businesses to determine whether additional verification is required. Information provided by the advertiser relating to their business may then trigger additional requests for information which can be reviewed either by automated or human review processes.

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

Response: Annual reports for the platforms in scope do not disaggregate their data on the basis of costs for set activities. However, it is worth noting that current business models pass some of the cost associated with verification to consumers. Companies such as X and Meta offer subscription models which allow advertisers to be designated as 'verified'.[37] These processes supplement the costs associated with advertiser verification. These subscriptions are optional and are not utilised by

[36] Regulation 2022/2065 of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC [2022] OJ L 277 Article 26(1)(b-c).

[37] See: 'Grow with Meta Verified' (Meta 2024) < https://www.facebook.com/business/tools/meta-verified-for-business>; 'Account eligibility for X Ads' (X, 2024) < https://business.x.com/en/help/ads-policies/campaign-considerations/about-eligibility-for-twitter-ads.html>.

all advertisers. These models may defer some of the costs of verification however there is similarly a lack of transparency around the operation of these accounts.

(d) The percentage of advertiser accounts that are verified

Response: There is no publicly available data on the percentage of advertiser accounts that are verified. Companies should be transparent with this information and make it available for review.

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

Response: The approach to advertiser verification differs across platforms. Under current practice, advertisers may be able to publish advertisements without verification or whilst verification is still ongoing. This creates concerns as fraudulent actors will exploit the windows in which they are able to post the ads without stringent verification. If their account is then suspended, they may simply create a new account and resume their fraudulent activities. Effective verification should be completed prior to allowing any advertiser to publish on a user-to-user or search service as a preventative action.

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: Verification is required for certain regulated industries including gambling and games, healthcare and medicines, and financial products or services. Advertisers who are placing adverts in this area may be subject to additional requirements and identity checks. Where the adverts are relevant to users under the age of 18, additional limitations may be required. Profiling and ad personalisation should be removed for those under the age of 18 and ads which fall within certain restricted categories should be prohibited.

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: Platforms may impose requirements to comply with re-verification procedures at set intervals. This is good practice. Platforms should similarly consider requiring re-verification when there has been a change to the relevant business information, such as information relating to the business registration or an alteration to the funding basis for the advertising. Where ads have been flagged for misleading practices, there should be a presumption that the account will be removed. However, if it has been determined following an appeal process that they will be allowed to maintain their presence, additional verification and/or re-

verification should be undertaken prior to permitting the account to resume any advertising operations.

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

Response: No.

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: The ad review process is primarily automated. Ad verification may be done by the user-to-user service or search service directly or through an independent third party ad verification service. The review process can begin automatically following the creation or editing of an ad. Platforms utilise automated technology to review ads for violations of advertising standards. However, adverts may not be reviewed against all policies before delivering impressions. Companies are clear that ads are subject to re-review at any time, however, it is not apparent how frequently this occurs. Under the provisions of the Digital Services Act VLOP/VLOSE are required to provide transparency reports concerning actions taken against certain content. While this data relates to EU operations it provides a useful overview for the scale of operations. Such transparency reports should be required for UK operations to provide the necessary data to evaluate the review and removal of content.

In the most recent reports, Meta reported 16,071,184 removals for advertising and commerce content. In addition, 2,714,843 advertising and commerce accounts were restricted. These restrictions were based off Meta's own initiative, likely through the use of automated tools and proactive monitoring. In terms of content reported by users, 1,190,353 instances of advertising and commerce content were removed however, 654,222 items were restored after the complaint. This represents an approximate 55% restoration of the content following the review process post user reporting. Google offers similar data relating to actions taken against advertisements on its own initiative or following user complaints.

(b) The role (i) automated processing and (ii) human processing play in the review process and how they interact

Response: Content which violates the policies of the user-to-user service or the search service may be subject to automated processes which detect, restrict and remove content. In evaluating the content, the automated processing may take various information into consideration including both the content of the ad (images, video, keywords) as well as associated ad destination. Account information may also be considered as part of the review process.

The automated process will then initiate any subsequent human review. The technology may similarly prioritise the content which does need a human review. The human review process may overturn the automated decision and records should be kept as to the decisions of both the automated and human review processes. Reported rates of human operators overturning automated reviews are low.[38]

(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: Reviews of content can either be triggered by the platform itself through its review processes or on the basis of a flag submitted by a user. Each platform sets out the content which would fall under deceptive/misleading practices or scams and/or fraud. These lists vary across providers without a single universal definition of the content which meets this threshold. The variances in this approach leave the decision open to interpretation which may undermine the effectiveness of the provisions. Similarly, the requirement of constituting one of the relevant Fraud offences requires understanding of what those offences entail. This requires both legal and contextual understanding and it may not be possible for this understanding to be clearly evident in automated processes.

(d) The timescales for review

Response: Once a flag for review has been received platforms the review process is largely completed within 1 day. In more challenging cases, where a human reviewer becomes involved, the process may take longer.

(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: The approach to an advertisements visibility and reach whilst it is flagged for review differs depending on the provider. In some cases, the content is removed from view whilst the review is ongoing. In others the advert remains live

[38] Less than 0.7% of the fully automated enforcement decisions on ads placed by advertisers in the EU were overturned after subsequently going through human review (*EU Digital Services Act (EU DSA) Biannual VLOSE/VLOP Transparency Report*(Google, 26 April 2024).

whilst the review takes place and is only restricted following a decision. There should be a consistency of approach across platforms.
(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: Following the passage of the DSA, very large online providers and very large online search services are required to work with trusted flaggers and other entities who can report content that they should be removed from the services under the applicable law.[39] However, to date there is only one designated trusted flagger who has been approved by the European Commission. As such most platforms have not yet engaged with a trusted flagger system.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No.

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

[39] Regulation 2022/2065 of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC [2022] OJ L 277 Article 22.

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: User reporting tools for paid-for advertisements are distinct from those used for user-generated content or search results. They offer dedicated reporting options specifically for ads, focusing on holding advertisers accountable for misleading or fraudulent content. These mechanisms are often integrated with ad platforms for faster identification and removal of problematic ads. Additionally, they may collect data on reported ads to analyse trends and improve ad review processes.[40] In contrast, reporting tools for user-generated content and search results have a broader scope, encompassing various violations of community guidelines and not necessarily focusing on advertiser accountability. While they may lead to content removal or user warnings, they are less specialized than tools designed explicitly for addressing fraudulent advertising.[41]
(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: While precise figures remain elusive, a substantial portion of user reports regarding advertisements likely pertain to suspected fraudulent content. This can range from a few percent to a significant double-digit percentage, depending on the platform, the type of advertising, and the effectiveness of pre-screening measures.[42] In terms of processes for taking action, to address these reports, platforms typically employ a multi-step process involving initial review by automated

[40] See for instance Paid Media Reporting Tool - Portermetrics https://portermetrics.com/en/solutions/paid-media-reporting-tool/

[41] See for instance Best UGC Tools for Next-Gen Marketers https://taggbox.com/blog/ugc-tools/

[42] See for instance Ad Fraud Statistics (2023) https://www.businessofapps.com/ads/ad-fraud/research/ad-fraud-statistics/

systems or human moderators, followed by a more thorough investigation if fraud is suspected. This investigation may involve contacting the advertiser, verifying claims, and examining documentation. Regarding action based on the findings, platforms can remove the ad, suspend, or terminate the advertiser's account, offer refunds or compensation to affected users, or even pursue legal action. Advertisers typically have the right to appeal these decisions.[43]

However, platforms face challenges in effectively tackling fraudulent ad reports due to resource limitations, evolving fraud tactics, and the risk of false positives impacting legitimate advertisers.[44] Despite these challenges, platforms are actively working to improve detection and enforcement mechanisms by investing in advanced technologies, collaborating with industry partners, and educating users on identifying and reporting fraudulent ads. By continually refining these processes, platforms strive to maintain a safe and trustworthy advertising ecosystem for both users and advertisers.

(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: While precise figures on user reports of suspected fraudulent advertising are often undisclosed due to commercial sensitivities, available information suggests it's a significant issue. Major platforms like Facebook and Google receive millions of ad-related reports annually, with a substantial portion likely tied to fraud. Facebook's 2019 report of removing 2.2 billion fake accounts, potentially linked to fraudulent advertising, illustrates the scale of the problem.[45] Google's 2023 transparency report indicates over 5.5 billion ads removed for policy violations, including fraud, but lacks specifics on user reports.[46]

Limited public data exists on appeal rates for fraudulent advertising reports, but anecdotal evidence and industry reports suggest appeals are common, especially for complex or borderline cases. Success rates vary depending on platform policies and evidence presented. For instance, organizations like the Interactive Advertising Bureau (IAB)[47] and the Coalition for Better Ads (CBA)[48] publish reports on advertising practices and trends, sometimes touching upon issues of fraud and enforcement. While not always focused on appeals specifically, they provide context on the broader landscape.

Challenges in data transparency arise due to commercial sensitivity, privacy concerns, and the evolving nature of fraudulent advertising. However, resources like the Federal Trade Commission's (FTC) Consumer Sentinel Network offer insights into consumer complaints about deceptive advertising, though not exclusive to online platforms.[49]

---

[43] Ibid.

[44] Ibid.

[45] See for instance https://www.statista.com/statistics/1013474/facebook-fake-account-removal-quarter/#:~:text=A%20record%20figure%20of%20approximately,%2C%20or%20non%2Dhuman%20entity.

[46] See for example https://www.ppchero.com/google-blocked-5-5-billion-ads-in-2023-safety-report/

[47] See generally https://www.iab.com/

[48] See generally https://www.betterads.org/

[49] See generally https://www.ftc.gov/enforcement/consumer-sentinel-network

Despite limited transparency, user reports remain crucial in combating fraudulent advertising. By empowering users to flag suspicious ads, platforms can leverage collective intelligence to identify and address problematic content, as done by the Coalition for Better Ads.[50] This collaborative approach, combined with advancements in detection technology and increased platform accountability, is essential for mitigating the impact of fraudulent advertising in the digital landscape.

(d) The criteria used to classify and prioritise user reports

Response: Platforms employ a multi-faceted approach to classify and prioritize user reports of fraudulent advertising. Primarily, the severity of the alleged violation plays a crucial role, with ads promoting scams, counterfeit products, or dangerous goods receiving higher priority than those with minor inaccuracies. The type of fraud also influences prioritization, as different types, like click fraud or deceptive content, have varying impacts and detection difficulties. The credibility of the report, determined by factors like user history and supporting evidence, is also considered. High volumes of reports for a single ad may trigger prioritization due to increased likelihood of violation.[51]

Additionally, ads with a broad reach or targeting vulnerable groups, as well as those from advertisers with a history of violations, are often prioritized. Legal and regulatory concerns, such as ads for prohibited products or deceptive claims, also elevate priority. Resource availability and automated filtering systems play a role, with platforms focusing on quickly addressable or high-risk reports. User feedback on report outcomes further refines prioritization criteria. While these criteria may vary across platforms and lack transparency, understanding them empowers users to submit more effective reports and advocate for improved protection against fraudulent advertising.[52]

(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: The time it takes platforms to respond to user reports of fraudulent advertising varies widely, influenced by factors like platform size, report complexity, prioritization, and the balance between automation and human review. While precise data is often undisclosed, response times can range from a few hours to several days, or even weeks in complex cases.[53]

To ensure timely and accurate responses, platforms employ several measures. Automated filtering and prioritization help identify urgent reports, while clear reporting guidelines aid users in submitting effective complaints. Dedicated teams of trained moderators specialize in investigating fraudulent advertising, ensuring informed decisions. Feedback mechanisms and transparency reports allow plat-

[50] See generally https://www.betterads.org/research/

[51] See for instance FasterCapital, ‘‘Online advertising fraud: how to detect and prevent online advertising fraud’’ (April 2024) https://fastercapital.com/content/Online-advertising-fraud--How-to-Detect-and-Prevent-Online-Advertising-Fraud.html

[52] Ibid.

[53] Ibid.

forms to track response effectiveness and identify areas for improvement. Collaboration with industry partners and regulatory bodies fosters information sharing and best practices.[54]

(f) Any measures taken to make user reporting tools accessible, easy to use and easy to find for users

Response: Platforms have prioritized making user reporting tools for fraudulent ads accessible and user-friendly. In-ad reporting options, such as "Report Ad" buttons or links, are prominently placed for quick access, while menu options ensure consistent availability. The reporting process is designed to be intuitive, requiring minimal steps and providing clear instructions (eg users are typically guided through a series of questions or options to categorize their complaint and provide additional details). Accessibility is ensured through multi-platform support, multiple languages, and features like screen reader compatibility. Educational resources, such as help centers, FAQs sections and awareness campaigns, further empower users to identify and report fraudulent ads. Platforms actively solicit feedback to continuously improve these tools, keeping them effective and user-friendly in response to evolving fraud trends.[55]

(g) How transparency and communication is maintained with users who have submitted reports

Response: Maintaining open communication with users who report fraudulent advertising is crucial for platforms to build trust and ensure the reporting system's efficacy. Upon submission, users receive immediate confirmation, setting expectations for the review process. Platforms may provide periodic status updates, informing users when their report is under review or if further information is needed. Once a decision is reached, users are notified of the outcome, including details about ad removal, reasoning, and actions taken against the advertiser. If dissatisfied, users can often appeal the decision, receiving further communication about the appeal's outcome.[56]

Some platforms enhance transparency by publishing reports detailing the number of reports received, actions taken, and overall enforcement effectiveness. This allows users to understand the impact of their reports and the platform's dedication to combating fraudulent advertising. Additionally, feedback mechanisms and customer support channels provide avenues for direct communication, allowing users to voice concerns, ask questions, and contribute to improving the reporting process.[57] By prioritizing transparency and communication, platforms foster user trust, encourage continued reporting, and strengthen collective efforts to combat fraudulent advertising, contributing to a safer and more transparent online advertising ecosystem.

[54] Ibid.

[55] See for example Google's My Ad Center Help, 'Control the ads you see when you see them' https://support.google.com/My-Ad-Center-Help/answer/12155764?hl=en

[56] Ibid.

[57] Ibid.

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

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: Third-party involvement in paid-for advertising has been linked to various fraudulent activities, including click fraud, ad injection, and the promotion of fake news and misinformation. These practices not only deceive advertisers and harm their campaigns but also contribute to the spread of harmful content and undermine the credibility of online platforms. The involvement of third parties poses challenges for service providers in maintaining control over advertising content due to limited visibility into their actions and the complexity of tracking systems. Even when fraud is detected, enforcement can be hindered by inadequate contractual agreements and the potential for conflicts of interest between service providers and ad networks. While terms of service and advertiser agreements often prohibit fraudulent advertising, their enforcement remains inconsistent. Legal frameworks exist to address such practices, but they can be complex and time-consuming to navigate. To effectively combat fraudulent advertising, service providers need to strengthen their contractual agreements with third-party ad networks, enhance monitoring and detection capabilities, and collaborate with industry partners and regulators to develop more robust enforcement mechanisms. By doing so, they can protect users from harmful content, maintain the integrity of their platforms, and ensure a fair and transparent advertising ecosystem.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No.

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: While precise figures on the prevalence of deepfake fraudulent advertisements remain elusive, emerging evidence suggests a growing trend. Anecdotal reports, such as the recent Elon Musk cryptocurrency scam highlighted in a Euronews article (April 2024), indicate an increasing use of deepfakes in fraudulent schemes.[58] Moreover, in 2023 a video of MoneySavingExpert Martin Lewis was widely shared on social media, using generative AI to create a realistic-looking image and voice of the journalist promoting a fake Elon Musk investment opportunity in Quantum AI.[59] However, unfortunately the opportunity was a scam, not a legitimate investment. Additionally, a recent $25 million fraud case involving deepfakes impersonating senior management underscores the dangers of AI-generated biometrics.[60] Although no definitive figures exist on the proportion of fraudulent ads utilizing deepfakes, experts anticipate a significant increase as the technology becomes more accessible and sophisticated.[61] The ability of deepfakes to convincingly manipulate audio and video makes them a potent tool for deception, prompting concerns about their widespread use in fraudulent advertising.

This growing trend is further evidenced by discussions on platforms like Reddit, where users report encountering deepfake ads promoting scams and pyramid schemes, like the one featuring Elon Musk mentioned in a Reddit post.[62] Furthermore, legislative actions such as the "DEEP FAKES Accountability Act"[63] and the "NO AI FRAUD Act"[64] introduced in the U.S. House of Representatives indicate growing awareness and concern about the potential misuse of deepfakes in advertising. While concrete data is limited, these examples and expert opinions suggest a concerning upward trajectory in the use of deepfakes for fraudulent

---

[58] Euronews, ''It is a scam! How deepfakes and voice cloning taps into your cash'' (April 20224) https://www.euronews.com/business/2024/04/10/its-a-scam-how-deepfakes-and-voice-cloning-taps-into-your-cash

[59] MoneySavingExpert. 2023. "Warning: beware terrifying new 'deepfake' Martin Lewis video scam promoting a fake 'Elon Musk investment' - it's not real." https://www.moneysavingexpert.com/news/2023/07/beware-terrifying-new--deepfake--martin-lewis-video-scam-promoti/.

[60] Biometric Update. 2024. "Deepfake Videos Looked So Real that an Employee Agreed to Send Them $25 Million." https://www.biometricupdate.com/202402/deepfake-videos-looked-so-real-that-an-employee-agreed-to-send-them-25-million.

[61] Romero-Moreno, F. (2024). Generative AI and deepfakes: a human rights approach to tackling harmful content. International Review of Law, Computers & Technology, [online] 1-13. [doi:10.1080/13600869.2024.2324540] https://www.tandfonline.com/doi/full/10.1080/13600869.2024.2324540

[62] https://www.reddit.com/r/youtube/comments/16hrzmg/are_deepfake_video_ads_legal_and_or_acceptable/

[63] https://www.congress.gov/bill/117th-congress/house-bill/2395/text

[64] https://www.congress.gov/bill/118th-congress/house-bill/6943

advertising, necessitating proactive measures from platforms and regulators to mitigate the associated risks.[65]

(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: Detecting deepfakes and AI-generated content in fraudulent advertising relies on a combination of methodologies, each with varying effectiveness. Visual and audio forensics analyse media for inconsistencies, proving highly effective for simpler deepfakes but struggling with advanced techniques like GANs. Metadata analysis, while helpful for identifying manipulations, can be easily circumvented. Machine learning and deep learning algorithms, particularly those based on convolutional neural networks (CNNs), show promise in detecting realistic deepfakes but require extensive training data and computational resources.

(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: The choice between developing in-house or acquiring third-party deepfake detection technologies is a strategic decision for platforms, influenced by resources, expertise, and specific needs. In-house development offers the advantage of customization to the platform's content and user base, along with full control over development and data ownership. However, it can be time-consuming, costly, and requires specialized AI and machine learning expertise. Alternatively, acquiring third-party solutions allows for faster deployment and lower initial costs, leveraging the expertise of specialized vendors. However, customization options may be limited, and platforms become reliant on the vendor for updates and support, potentially raising data sharing concerns.

Integrating deepfake detection tools, whether developed in-house or acquired, can take varying amounts of time. In-house development can span months or years, while third-party integration typically ranges from weeks to months, depending on system complexity. Facebook's Deepfake Detection Challenge[66] (DFDC) exemplifies in-house development, while Google's acquisition of Jigsaw[67] demonstrates the third-party route. Twitter's partnership with vendors showcases a hybrid approach. Ultimately, the optimal choice depends on each platform's unique circumstances and priorities, with a combination of in-house and third-party solutions potentially offering the best balance of customization, speed, and cost-effectiveness.

(d) The accuracy of detection methods, including true positive and false positive rates

[65] Romero-Moreno, F. (2024). Generative AI and deepfakes: a human rights approach to tackling harmful content. International Review of Law, Computers & Technology, [online] 1-13. [doi:10.1080/13600869.2024.2324540] https://www.tandfonline.com/doi/full/10.1080/13600869.2024.2324540

[66] See Meta, 'Creating a dataset and a challenge for deepfakes' (2019) https://ai.meta.com/blog/deepfake-detection-challenge/

[67] https://jigsaw.google.com/

Response: For example, one example of a well-known system is Sensity, which recognises AI-manipulated media and synthesis techniques such as AI-created faces incorporated into social media profiles, and realistic video face swaps. Sensity is trained on millions of gan-generated images to identify imperfections and small details of AI-created images.[68] Moreover, another popular system is Intel's FakeCatcher, which, using Photoplethysmography, analyses the movement of blood vessels in a video. The colour of veins changes as the heart pumps blood through them. These 'blood flow' signals are extracted from the face and then, FakeCatcher can reliably identify real and fake videos.[69]

It is noteworthy that, while Sensity claims that it can identify realistic full bodies and faces generated using AI models like Dall-E with 98.8% accuracy,[70] Intel asserts that its FakeCatcher technology is the first real-time system, with 96% precision.[71]

In this context, the EU Court of Justice (CJEU) Advocate General (AG) opinion in *Poland v Council and Parliament* stressed that if filtering content would inevitably lead to a significant number of 'false positives,' rendering it ineffective, such measures should be excluded.[72] Furthermore, in *UPC Telekabel Wien* the CJEU held that, to strike a fair balance, it was crucial under Article 16 of the EU Charter, to allow companies to choose the measures they will take, considering their capabilities and resources.[73] However, the difficulty is that *Clarkson v OpenAI* also warned about the overfitting problem, where a torrent of AI-generated child abuse images confused the monitoring system since it was designed only to filter and block familiar images of abuse, but worryingly, not recognise newly created ones.[74]

(e) The costs associated with the development/acquisition and deployment of these detection mechanisms

Response: The costs associated with combating deepfakes and fraudulent advertising are substantial and multifaceted. Meta's investment of $10 million in their Deepfake Detection Challenge (DFDC) exemplifies the financial commitment required for in-house research and development.[75] The acquisition of deepfake detection startup Sensity AI by Microsoft for an undisclosed amount further underscores the market value of such technologies.[76] Third-party solutions like Sentinel

[68] Sensity. 2023. "Deepfake Detection." https://sensity.ai/deepfake-detection/

[69] Intel. 2022. "Intel Introduces Real-Time Deepfake Detector." https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html.

[70] Sensity. 2023. "Deepfake Detection." https://sensity.ai/deepfake-detection/

[71] Intel. 2022. "Intel Introduces Real-Time Deepfake Detector." https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html.

[72] AG opinion in C-401/19 *Poland v Parliament and Council* [2021] ECLI:EU:C:2021:613 [214].

[73] C-314/12 *UPC Telekabel Wien GmbH v Constantin FilmVerleih GmbH and Wega Filmproduktionsgesellschaft GmbH* [2013] EU:C:2014:192 [52].

[74] *PM et al v OpenAI LP* (N.D. Cal. 2023) [226].

[75] See Meta, 'Creating a dataset and a challenge for deepfakes' (2019) https://ai.meta.com/blog/deepfake-detection-challenge/

[76] https://sensity.ai/

offer scalable pricing models, indicating the flexibility in costs depending on platform needs.[77]

Industry reports like the one from Partnership on AI in 2020 estimated that developing and deploying robust deepfake detection tools could cost major platforms millions annually.[78] This is supported by DARPA's significant funding for deepfake detection research, with projects receiving millions in grants.[79] Additionally, the high demand for skilled AI and machine learning professionals continues to drive up salaries, further contributing to development costs.

Overall, the financial burden of combating deepfakes is significant and ongoing, requiring continuous investment to stay ahead of evolving threats. While precise figures are often proprietary, the available evidence paints a picture of substantial costs across research, development, talent acquisition, data collection, and computational resources. Platforms must carefully weigh these costs against their individual needs and resources to determine the most effective and sustainable solutions for protecting users and maintaining trust in their platforms.

(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: Deepfake and AI-generated content pose significant challenges for detecting fraudulent advertising across various levels. Automated processes struggle to identify sophisticated deepfakes created with advanced techniques like GANs, which produce realistic impersonations or manipulate text convincingly (eg impersonations of celebrities or public figures endorsing products or services). Additionally, dynamic content that adapts to user interactions (eg an ad might display different images or text depending on the user's location or browsing history), and textual deepfakes mimicking trusted sources can easily evade automated detection (eg used in phishing scams or to spread misinformation about products or services).[80]

Human moderators, while more adept at nuanced analysis, face difficulties with subtle deepfakes involving minor alterations (eg changing facial expressions or micro-expressions), convincing audio imitations (eg used in voice phishing scams or to create fake audio testimonials, and contextual deepfakes that rely on surrounding information to appear authentic such as a news article or social media pos. These subtle manipulations can be hard to spot even for trained professionals. For instance, a video of MoneySavingExpert Martin Lewis was widely shared on social media, using generative AI to create a realistic-looking image and voice of the journalist promoting a fake Elon Musk investment opportunity in Quantum AI.[81] However, unfortunately the opportunity was a scam, not a legitimate investment.

---

[77] https://thesentinel.ai/

[78] https://partnershiponai.org/

[79] Defense Advanced Research Projects Agency. 2019. “Media Forensics (MediFor) (Archived). https://www.darpa.mil/program/media-forensics

[80] See for instance Wired, ‘‘AI-generated text is the scariest deepfake of all’’ (July, 2020) https://www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/

[81] MoneySavingExpert. 2023. “Warning: beware terryfinying new ‘deepfake’ Martin Lewis video scam promoting a fake ‘Elon Musk investment’ - it’snot real.” https://www.moneysavingexpert.com/news/2023/07/beware-terrifying-new--deepfake--martin-lewis-video-scam-promoti/.

Service users are particularly vulnerable to personalized deepfakes tailored to their interests or demographics, which can be remarkably persuasive. Deepfakes that exploit emotions, such as fear or excitement, can cloud judgment, making users more susceptible to fraudulent claims. For instance, a recent $25 million fraud case involving deepfakes impersonating senior management underscores the dangers of AI-generated biometrics.[82] Furthermore, information overload can overwhelm users, making them less likely to scrutinize content and more likely to fall prey to well-crafted deepfakes. Overall, the evolving nature of deepfakes and AI-generated content demands a multi-faceted approach to detection, combining advanced technological solutions with human expertise and user education.[83]

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

Response: No.

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: In principle, parents are likely to want very detailed information about their deceased child's use of online services to help them understand causes of death or other circumstances surrounding the end-of-life period and the time immediately preceding this. They may be interested in accessing all the data (and metadata) available, including usage logs, content accesses, time spent on service, interaction and communication, friends, followers and contacts, chats, the nature of interactions, posts and comments, activities and shared content, location data, safety incidents, reports made, warning and bans, etc.

However, we would like to warn that not all the requests should be fulfilled by the providers. The main reason is the protection of privacy. Even if one argues that child's postmortem privacy isn't protection and personal data aren't protected after death,[84] the protection of privacy of the child's contacts, friends, followers need to be considered. This is less relevant for the content that is publicly available and not protected with privacy settings, however, sensitive data and communications in particular, need to be protected and an unfiltered access to an account should not

[82] Biometric Update. 2024. "Deepfake Videos Looked So Real that an Employee Agreed to Send Them $25 Million." https://www.biometricupdate.com/202402/deepfake-videos-looked-so-real-that-an-employee-agreed-to-send-them-25-million.

[83] Romero-Moreno, F. (2024). Generative AI and deepfakes: a human rights approach to tackling harmful content. International Review of Law, Computers & Technology, [online] 1-13. [doi:10.1080/13600869.2024.2324540] https://www.tandfonline.com/doi/full/10.1080/13600869.2024.2324540

[84] See Harbinja, E, 'Post-mortem privacy 2.0: Theory, law and technology' 22 Feb 2017, In: International Review of Law, Computers and Technology. 31, 1, p. 26-42; Harbinja, E., *Digital death, digital assets and post-mortem privacy*, 2022, Edinburgh University Press

be an option. An idea would be to interpret the phrase from s. 75 'information about the child's use of the service' restrictively and provide redacted information strictly relevant for a request with a valid, significant reason (e.g. certain number of metadata, activity that excludes legitimate and unproblematic personal communication etc.). This would entail more work for the providers, but it is necessary for the protection of privacy and personal data as noted above.

Additionally, responses should be child rights respecting in accordance with the UN Convention on the Rights of the Child (UNCRC) and in particular with reference to General Comment 25 (2021) on Children's Rights in Relation to the Digital Environment'. General Comment 25 makes a strong case for children's privacy, but also speaks to broader children's rights.

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

Response: No.

Question 51: How long should it take to receive information in response to a request?

Response: The timeframe for receiving information in response to a request can vary depending on the type of request and the type of service. It could be contextual. As a guide, the Subject Access Requests (SARs) under GDPR could be used, and the period of one month of receipt of the request, extended by two further months if the request is complex or numerous. Additionally, s. 10 of the Freedom of Information Act 2000 that stipulates that public authorities must respond to a request for information within 20 working days from the date of receipt of the request could also be used. Some companies may have their own internal policies for responding to requests for information with shorter timeframes stated in their privacy policies or terms of service and these should also be allowed.

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

Response: No.

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: To assist parents in finding out how to obtain information and updates regarding their deceased child's use of a service, larger (category 1) online platforms can provide several mechanisms including dedicated parent portals where parents can access all relevant information, including how to request information. They could include detailed instructions on how to use the portal, request information, and manage parental controls. There should be a

comprehensive FAQ section and a searchable knowledge base in this area. They should promptly acknowledge receipt of the request, even if the final response will take longer and offer clear contact options and multiple ways for parents to contact customer support, including email, live chat, and phone support. They should train support teams specifically to handle inquiries from parents, ensuring they can provide accurate and helpful information quickly.

All concerned providers should ensure the ToS clearly outlines parents' rights to access information about their deceased child's use of the service, including step-by-step instructions for making such requests. They should clearly notify parents when their requests for information have been processed or if additional information is required. They should encourage parents to provide feedback on the information request process and use this feedback to improve services.

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

Response: No.

Question 53: What support or information do parents need to guide them through the process of making a request?

Response: To effectively guide parents through the process of making a request for information about their child's use of an online service, several types of support and information should be provided, including detailed instructions and tutorials, step-by-step instructions on how to make a request, including screenshots or video tutorials for visual guidance, downloadable and printable guides that parents can keep for reference, ensure accessible customer support through multiple channels such as phone, email, and live chat. They should maintain a comprehensive FAQ section and help centre with articles specifically addressing common questions parents might have about making requests.

They should create downloadable templates or forms that parents can fill out to make a request, ensuring these templates include all necessary fields and instructions, create user-friendly online forms that guide parents through the request process, making it easy to submit all required information electronically. Importantly, they should explain the data protection measures in place to secure the requested information and the data already provided by the parents in the request or through the verification and identification process. Providers should inform parents about the status of their request through regular updates and notifications to manage expectations and keep them informed throughout the process. They should also provide clear timelines for when they can expect to receive the requested information.

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

Response: No.

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: We would like to refer back to our response to question 1, as much of it applies here as well. In addition to that, there should be a simple online form for submitting requests. These forms should be straightforward, with clear instructions

on the required information and documentation. User awareness is an issue for most aspects of ToS, including user rights, options for various requests and features. Often, these options are hidden and buried within ToS and not advertised properly to users. To address this, in addition what we've said in our responses to the above questions, users should be prompted regularly about the existence of these option, through emails, notification, pop-ups etc.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response: No.

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: