Call for evidence response form

Please complete this form in full and return to os-cfe@ofcom.org.uk.

Consultation title	Third phase of online safety regulation: Additional duties for categorised services
Full name	Sally Dray
Contact phone number	
Representing (delete as appropriate)	Organisation
Organisation name	MoneySavingExpert
Email address	campaigns@moneysavingexpert.com

Confidentiality

We ask for your contact details along with your response so that we can engage with you on this consultation. For further information about how Ofcom handles your personal information and your corresponding rights, see Ofcom's General Privacy Statement.

Your details: We will keep your contact number and email address confidential. Is there anything else you want to keep confidential? Delete as appropriate.	Nothing
Your response: Please indicate how much of your response you want to keep confidential. Delete as appropriate.	None
For confidential responses, can Ofcom publish a reference to the contents of your response?	

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

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

For providers of online services

Question 3: How do you ensure users understand the provisions in your terms of service about taking down content, restricting access to content, or suspending or banning a user from accessing the service and the actions you might take in response to violations of those terms of service? In your response to this question, please provide information relating to (a) – (d) where relevant.
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:
(b) how to ensure that reporting and complaints mechanisms are not misused
Response:
(c) the key choices and factors involved in designing these mechanisms
Response:

(d) how users can or should be supported to report/complain about specific concerns (e.g., other users, certain types of content or, appeal content takedowns or account bans)
Response:
(e) how to ensure they are user-friendly and accessible to all users (e.g., disabled users, children)
Response:
(f) whether users are informed that their reports are anonymous (e.g., other users will not be informed about who has reported their content or account);
Response:
(g) any user support materials that explain how to use the reporting and complaints process and what will happen when users engage with these systems
Response:
(h) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For providers of online services

Question 7: Can you provide any evidence or information about the best practices for effective reporting and/or complaints mechanisms, and how these processes are designed and maintained? In your response to this question, please provide evidence relating to (a) – (j) where relevant.
Response:
(a) how users report harmful content on your service(s) (including the mechanisms' location and prominence for users, and any screenshots you can provide);
Response:
(b) whether there are separate or different reporting or complaints mechanisms or processes for different types of content and/or for different types of users, including children;
Response:
(c) how users appeal against content takedowns, content restrictions or account suspensions or bans;
Response:
(d) what type of content or conduct users and non-users may make a complaint about / report, including any specific lists or categories;
Response:

(e) whether users need to create accounts to access reporting and complaints mechanisms (if there are multiple mechanisms, please provide information for each mechanism);
Response:
(f) whether reporting and complaints mechanisms are effective, in terms of: **(i) enabling users to easily report content they consider to be potentially the types of content specified in the relevant terms of service, and how to determine effectiveness;**
Response:
(ii) enabling, supporting or improving the accuracy of user reporting in relation to identifying the types of content specified in the relevant terms of service, and how to determine effectiveness;
Response:
(iii) enabling, supporting or improving the provider's ability to detect and take timely enforcement action against content or users as specified in the relevant terms of service, and how to determine effectiveness;
Response:
(g) whether there are any reporting or complaints mechanisms you consider to be less effective in terms of identifying certain types of content and how you determine this;
Response:
(h) the use of trusted flaggers (and if reports from trusted flaggers should be prioritised over reports or complaints from users);
Response:
(i) the cost involved in designing and maintaining reporting and/or complaints mechanisms, including any relevant issues, difficulties or considerations relating to scalability; and
Response:
(j) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 8: What actions do or should services take in response to reports or complaints about content that is potentially prohibited or accounts engaging in potentially prohibited activity? In your response to this question, please include information relating to (a) – (g) where relevant.
Response:
(a) what proportion of reports are reviewed, and what proportion result in action taken including;

(i) any potential variation in the number and actionability (i.e., the proportion that result in a takedown or other action) of reports or complaints in relation to different provisions within your terms of service;
Response:
(ii) any differences for cases involving multiple reports/complaints about a single piece of content or user;
Response:
(iii) the costs associated with reviewing reports;
Response:
(b) whether any reports or complaints are expedited or directed to specialist teams, including: **(i) the criteria for this;**
Response:
(ii) the cost involved in facilitating this;
Response:
(c) the extent to which relevant individuals (content creators, users, and non-registered or logged-out users) are informed about the progress of their report or complaint, including: **(i) if they are not, the reasons why;**
Response:
(ii) if they are, what is included when users are informed about the progress of their report (e.g. receipt of the report, the progress of the report through the service's review process, and/or the outcome of the report);
Response:
(iii) the technical mechanisms/process to inform any relevant individuals about the progress of their report (e.g., whether non-registered users are provided an opportunity to provide an email address);
Response:
(iv) any differences in responses to different types of reports (e.g., reports about content or an account a user believes violates the terms of service, about the provider not operating in line with its terms of service, or about the accessibility, clarity or comprehensibility of those terms of service);
Response:
(v) the costs associated with responding to reports;
Response:
(d) what happens to the content while it is being assessed/processed (e.g., if and how it may still be found or viewed by other users);
Response:

(e) any internal or external timeframes or key performance indicators (KPIs) for reviewing and/or acting on reports or complaints;
Response:
(f) any user support materials that are used or should be used to support users understand the service's responses to reports, or how users can appeal moderation decisions about their content or accounts, or about decisions taken in response to reports they have submitted about other users' content or accounts;
Response:
(g) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Questions 9 – 15: Moderation

For all respondents

Question 9: Could improvements be made to content moderation to deliver more consistent enforcement of terms of service, without unduly restricting user activity? If so, what improvements could be made? **In your response to this question, please provide information relating to (a) –(c) where relevant.**
Response:
(a) improvements in terms of user safety and user rights (e.g., freedom of expression), as well as any relevant considerations around potential costs or cost drivers;
Response:
(b) evidence of the effectiveness of existing moderation systems including any relevant examples of the accuracy, bias and or effectiveness of specific moderation processes;
Response:
(c) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For providers of online services

Question 10: Please describe circumstances where you have taken or would take enforcement action against content or users outside of what is set out publicly in your terms of service and the reasons for taking this action. In your response to this question, please provide information relating to (a) – (e) where relevant.
Response:
(a) the types of action taken, and frequency of these actions (including per type of action);
Response:
(b) how relevant content or users were or would be brought to your attention;
Response:
(c) any policies, approaches or processes you have used or would use to guide moderation decisions in these cases;
Response:
(d) whether new policies are or would be written in response to these cases, and if so: **(i) whether and when these new policies are written before enforcement action is taken or after;**
Response:
(ii) when and how these new policies would be added to or included in your publicly available terms of service;
Response:
(e) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 11: If you are made aware of content or an account that potentially violates your terms of service, please describe any relevant circumstances which might not result in enforcement action, immediately or at all. **In your response to this question, please provide describe (with examples) any relevant circumstances relating to (a) – (e).**
Response:
(a) circumstances that relate to issues or challenges within your content moderation system (e.g. moderator error, language or local knowledge gaps, content is no longer available (e.g. livestream), nuance/context of content means it is found non-violative, further investigation needs to be done before action can be taken);
Response:

(b) circumstances that relate to issues or challenges within your terms of service and/or associated policies (e.g. new iterations of a harm falls outside the scope of internal moderation policies, individual piece of content is only of concern at scale (but itself does not violate policies);
Response:
(c) circumstances that relate to competing priorities (e.g., freedom of expression, public interest concerns);
Response:
(d) circumstances that would be understood by a user who has read the terms of service and why or why not, (e.g., the terms of service sets out exception for not removing violating content (e.g. news content), or transparency is not provided to avoid empowering bad actors);
Response:
(e) any other information.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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

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

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

(i) which languages are covered by your moderation team and how you decide which languages to cover;

Response:

(ii) whether moderators are employed by the service or outsourced, or are volunteers/users and any differences regarding how different provisions within the terms of service are moderated;

Response:

(iii) whether and how moderators are vetted, and any relevant consideration for how moderators are assigned to different roles relating to different provisions within the terms of service;

Response:

(iv) the type of coverage (e.g., weekends or overnight, UK time) moderators provide and any relevant considerations for different provisions within the terms of service;

Response:

(b) the process and costs associated with extending the use of human moderation for new/additional provisions in your terms of service, and any relevant barriers or challenges to adding new/additional provisions in your terms of service in relation to your human moderation resources;

Response:

(c) any other information.

Response:

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

Response:

Question 14: What training and support is or should be provided to moderators, and what are the costs incurred by providing this training and support?

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

Response:

(a) whether certain moderators are specialised in certain harms or subject material relating to different provisions in the terms of service;

Response:

(b) how services can/should/do assess the accuracy and consistency of human moderation teams;

Response:

(c) the impact of mental health or well-being support for moderators on the effectiveness of content moderation (including impacts on turn-over in moderation teams);

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

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

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

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

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

For all respondents

Question 16: What methods should service providers use to identify and define journalistic content and content of democratic importance, particularly at scale? In your response to this question, please provide information relating to (a) where relevant.
Response:
(a) how journalistic content and content of democratic importance can be described in the terms of service so that users can reasonably be expected to understand what content falls into these categories.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For providers of online services

Question 17: What, if any, methods are in place for identifying, defining or categorising content as journalistic content, content of democratic importance or news publisher content on your service? In particular, please provide any evidence regarding the effectiveness of any existing methods.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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

For providers of online services

Question 18: What considerations are taken into account when moderating journalistic content, news publisher content and content of democratic importance? In your response to this question, please provide information relating to (a) – (e) where relevant.

Response:
(a) once identified, how journalistic content, news publisher content and content of democratic importance is actioned and what kind of action is taken; and how that differs from the moderation of other types of content
Response:
(b) the factors that are or should be considered when taking action (e.g.: downranking/removal/suspension/ban or other) regarding this content
Response:
(c) the proportion of all journalistic content, content of democratic importance and news publisher content actioned upon by you that is actioned based on algorithmic decision making
Response:
(d) the proportion of all journalistic content, content of democratic importance and news publisher content actioned upon by you that is reviewed by human moderators and on what basis content is escalated to be reviewed by human moderators
Response:
(e) any insights into the costs of moderating journalistic content and content of democratic importance, including set up and ongoing costs in terms of employee time and other material costs.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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

For all respondents

Question 19: What complaint, counter-notice or other appeal processes should be in place for users to contest any action taken by service providers regarding journalistic content and content of democratic importance? In your response to this question, please provide information relating to (a) and (b) where relevant.
Response:
(a) examples of effective redress mechanisms that you consider would be most suited to these content types
Response:
(b) briefings, investigations, transparency reports, media investigations and research papers that provide more evidence
Response:

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

Question 20: What initiatives could service providers use to create and increase awareness about the process for users to complain and/or appeal content decisions and to minimise its' misuse? In your response to this question, please provide information relating to (a) and (b) where relevant.
Response:
(a) any known impacts of over-removal or erroneous removal of news publisher content, journalistic content or content of democratic importance
Response:
(b) briefings, investigations, transparency reports, media investigations and research papers regarding misuse of such speech protective provisions
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For providers of online services

Question 21: What are the current complaints, counter-notice or other appeal processes for users to contest any action taken by you regarding journalistic content, news publisher content and content of democratic importance on your service? In your response to this question, please provide information relating to (a) and (b) where relevant.
Response:
(a) any initiatives taken to create and increase awareness about the process for users to complain and/or appeal content removals
Response:
(b) any measures currently in place to prevent individual or systematic misuse of any protections for news publisher content, journalistic content or content of democratic importance.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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

For providers of online services

Question 22: Do you carry out any internal impact assessments to understand the freedom of expression and privacy implications of existing policies regarding journalistic content, news publisher content and content of democratic importance? In your response to this question, please provide information relating to (a) and (b) where relevant.
Response:
(a) explain which elements of your service design or operation they relate to and which factors they take into account
Response:
(b) provide relevant briefings, investigations, transparency reports, media investigations and research papers.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 23: What, if any, measures are in place to ensure that protection of content of democratic importance applies in the same way to a wide diversity of political opinion? In your response to this question, please provide information relating to (a) where relevant.
Response:
(a) whether there are any additional measures/safeguards that are put in place during local or national elections.
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For all respondents

Question 24: What, if any, measures can online service providers put in place to ensure that protection of content of democratic importance applies in the same way to a wide diversity of political opinion? In your response to this question, please provide information relating to (a) where relevant.
Response:
(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

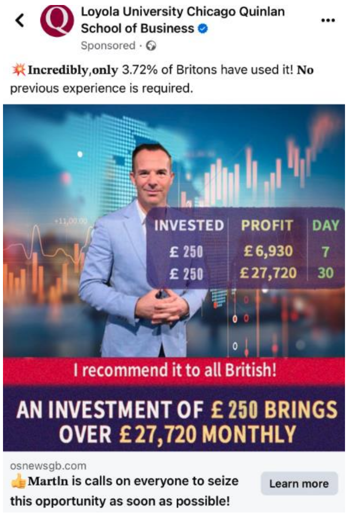
Questions 36 – 42: Overarching considerations

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)
The most prevalent form of fraudulent advertising that MSE encounters on a daily basis is impersonation ads, specifically of MoneySavingExpert (MSE) founder, Martin Lewis. These ads aim to trick people into giving their details, and subsequently money, to fake 'investment opportunities' or crypto-related 'schemes'. Over 2022 and 2023, we received 1,135 reports from our website users who had seen these scams on various platforms. Data provided to MSE by Action Fraud suggested a similar volume of

Martin-fronted scams were reported; it received 1,095 reports across the same time period. These reports were of scam ads featuring just Martin Lewis, but the number of reports increased when a combination of Martin *and* other well-known personalities featured in the scam. Of course, the number reported to either MSE or Action Fraud makes up a tiny fraction of those that that appear on consumers' feeds, so the true scale is likely to be far larger.

Here are just three examples of the type of fraudulent adverts we frequently encounter across platforms:

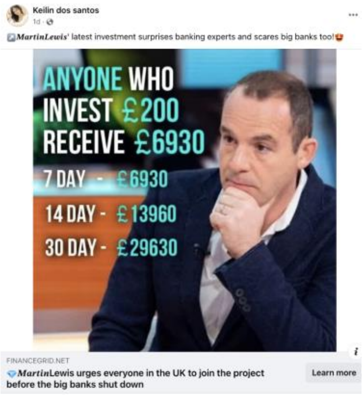


Loyola University Chicago Quinlan School of Business
Sponsored ·

💥**Incredibly,only** 3.72% of Britons have used it! **No** previous experience is required.

INVESTED PROFIT DAY
£ 250 £6,930 7
£ 250 £27,720 30
I recommend it to all British!
AN INVESTMENT OF £ 250 BRINGS OVER £27,720 MONTHLY

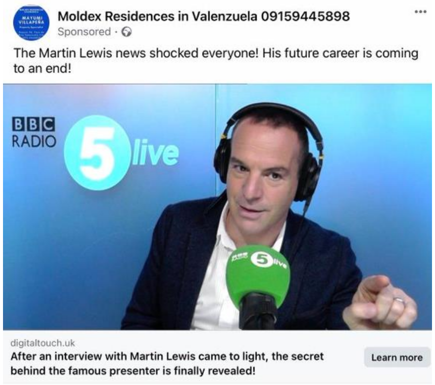
osnewsgb.com
👍**MartIn** is calls on everyone to seize this opportunity as soon as possible!



Keilin dos santos
1d ·

MartinLewis' latest investment surprises banking experts and scares big banks too!😍

FINANCEGRID.NET
💎***Martin***Lewis urges everyone in the UK to join the project before the big banks shut down



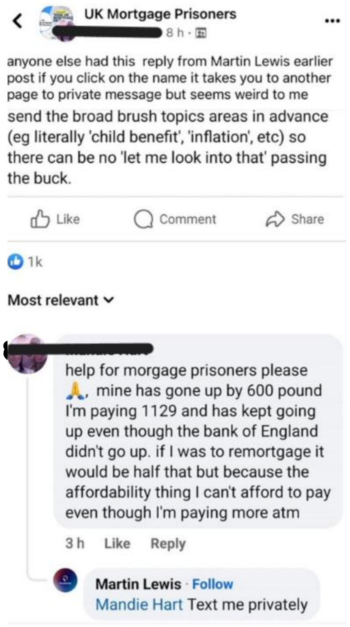
Moldex Residences in Valenzuela 09159445898
Sponsored ·

The Martin Lewis news shocked everyone! His future career is coming to an end!

digitaltouch.uk
After an interview with Martin Lewis came to light, the secret behind the famous presenter is finally revealed!

In addition to these fraudulent adverts, we have seen an increasing number of fake profiles impersonating Martin Lewis and MoneySavingExpert (MSE). These accounts will generally post some legitimate articles from the MSE website in order to appear legitimate, before then posting fraudulent advertisements. Oftentimes, the fake profile will reply to users' comments and ask them to send their phone numbers. The scammer then engages with the user off the platform – over the phone or via WhatsApp.

Below is an example of one such fake profile interacting with a campaigning group that MSE frequently works with:



UK Mortgage Prisoners
[redacted] 8 h ·

anyone else had this reply from Martin Lewis earlier post if you click on the name it takes you to another page to private message but seems weird to me

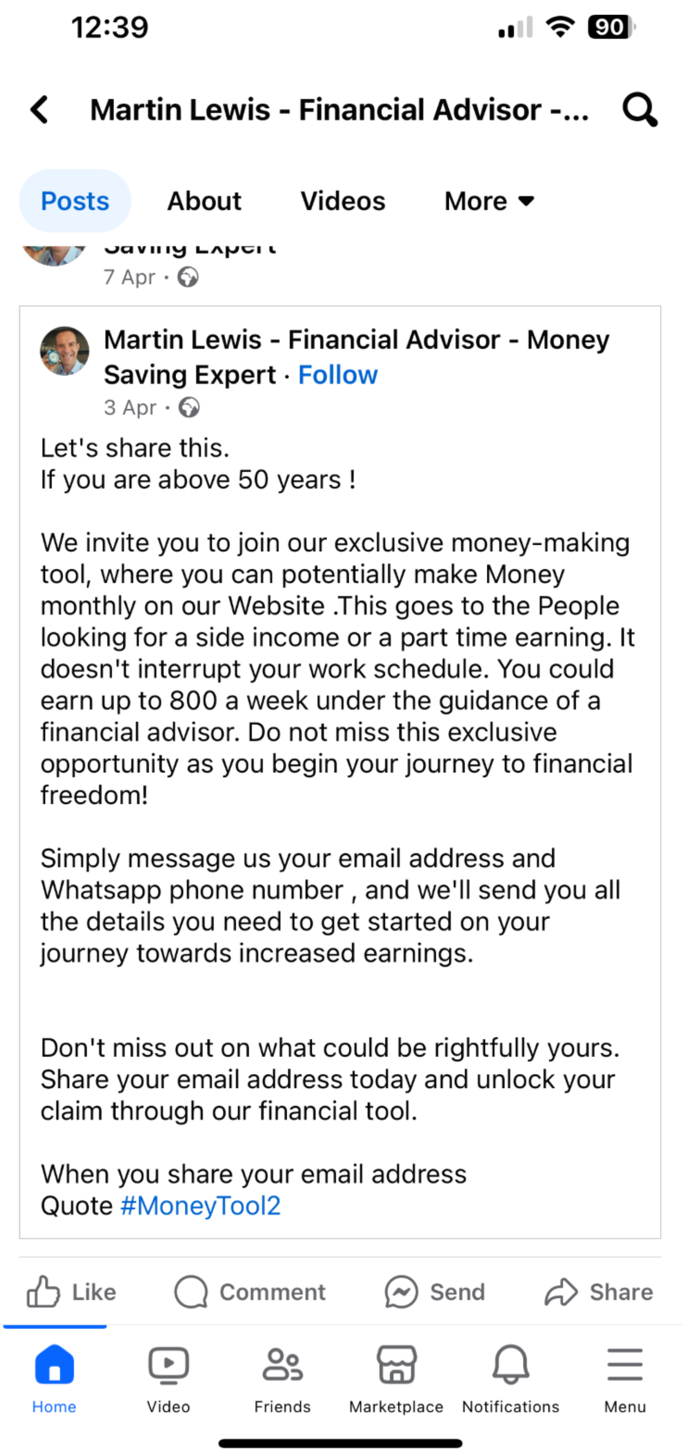
send the broad brush topics areas in advance (eg literally 'child benefit', 'inflation', etc) so there can be no 'let me look into that' passing the buck.

[redacted]
help for morgage prisoners please 🙏, mine has gone up by 600 pound I'm paying 1129 and has kept going up even though the bank of England didn't go up. if I was to remortgage it would be half that but because the affordability thing I can't afford to pay even though I'm paying more atm

3 h Like Reply

Martin Lewis · Follow
Mandie Hart Text me privately

Another example of a fake profile encouraging users to move to contact via WhatsApp:



Martin Lewis - Financial Advisor - Money Saving Expert · Follow
3 Apr ·

Let's share this.
If you are above 50 years !

We invite you to join our exclusive money-making tool, where you can potentially make Money monthly on our Website .This goes to the People looking for a side income or a part time earning. It doesn't interrupt your work schedule. You could earn up to 800 a week under the guidance of a financial advisor. Do not miss this exclusive opportunity as you begin your journey to financial freedom!

Simply message us your email address and Whatsapp phone number , and we'll send you all the details you need to get started on your journey towards increased earnings.

Don't miss out on what could be rightfully yours. Share your email address today and unlock your claim through our financial tool.

When you share your email address
Quote #MoneyTool2

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

MSE continually hears from users who have unfortunately fallen foul of these fraudulent advertisements, and who have suffered significant harm as a result. Users frequently tell us that it's not just their finances that are affected, but their physical and mental health too. Victims often feel their lives are ruined, say they are desperate for help and can't sleep at night. In the most extreme cases, users have reported feeling suicidal.

We have included just a few examples of recent messages received from MSE users who have fallen foul of a Martin-fronted scam, and wanted to share their experience and the impact this has had on their lives.

One user lost over £170,000 in savings and took out an additional £57,000 in loans after seeing a fake ad on Facebook. They told us:

"*I realise […] my case is one of thousands similar, and that I fall into the vulnerable demographic of being a 70 year old man with money to invest. I have been interested in investing in Crypto for sometime but know nothing about it. I saw one of the supposed Martin Lewis recommendations on Facebook earlier this year that recommended about 5 expert companies who could guide one through the process. I contacted two of them, who both responded by phone.*

"To keep a long story short, my advisor, a very intelligent young woman, managed to persuade me, over a period of about three months, to withdraw all my saved investment funds and take out three large loans, moving the money between a number of bank accounts. She persuaded me to move the money about, finally ending up at Revolut before going into Binance from where it disappeared into the ether! … I have lost over £170k in savings and another £57k in loans, with little prospect of recovering any of it, despite following all the advice I can and informing all relevant authorities.

"I have two children and five grandchildren, all of whose lives are being adversely affected. My wife is devastated and my life is in ruins! .… I have no Idea how l was so easily taken in, or how I made such a fool of myself. I had an exemplary credit record up to this, but she was an expert and waved her magic wand to good effect!"

Another user, who lost £38,000 to a Facebook scam, told us:

"*My life is totally ruined and I am now penniless […] I hope you are able to help advise me as I am desperate, distraught and want to help to stop this happening to anyone else. This level of sophisticated scamming is unbelievable and I never thought I would succumb to something like this. My money has been hard earned through nearly 40 years NHS service and instead of looking forward to retirement in a few months, I'm in a state of bankruptcy, as well as emotional trauma.*"

In another case, a user told us:

"I fell for the £250 bitcoin scam from the link on the page which Martin showed Susanna Reid on ITV… I have been conned out of £4,700 and am sick with worry."

While these scams are thought to affect more vulnerable groups, such as older people or those with lower digital literacy, we have heard from people of all ages and socio-economic backgrounds – challenging assumptions that the majority of victims are in some way vulnerable.

As for the financial impact of scams facilitated by these platforms, the scale of the losses is unclear – but according to the aforementioned data shared with MSE by Action Fraud, victims have reported losing over £20m to scams impersonating Martin Lewis in the past two years alone. Over £13m was lost in 2023 compared to £7m in 2022. The largest individual loss as a result of a scam featuring Martin was £500,000.

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

Scammers who seek to steal from users of these platforms have sophisticated methods and tactics to do so. We recognise that one challenge faced by platforms is the speed at which scam ads evolve – for example, the progression from doctored images to convincing AI-generated deepfakes. From our vantage point, it seems that platforms tend to be catching up with new methods, not proactively detecting them.

However, we are concerned that platforms do not seem to be removing and preventing content that has already been flagged to them as fraudulent. MSE has been reporting the same and similar types of fraudulent deepfake/AI videos for over 10 months, showing that detection tools are not currently effective. The videos use roughly the same images – sometimes there are small tweaks, but the bulk of the moving imagery is the same and the AI-generated voiceover tends to use the same script. Even basic moderation tools currently should be able to detect this standard of copy-and-paste content. *An example of this scam ad is provided in question 49 below.*

Nonetheless, throughout our campaigning efforts to tackle fraudulent advertising on platforms, MSE has maintained that if platforms are accepting money from advertisers, those advertisements should be properly vetted before being published and reaching consumers – whether this is through an effective technological solution, or by human efforts. This should limit the challenges in detecting fraudulent content *before* it is live on the platform, significantly reducing the risk of causing devastating harm to countless consumers.

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

N/A

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

It is impossible to estimate how many fraudulent advertisements are currently on platforms, but a few points suggest that the number is significant:

1. Actual reports from users who have come across these fake ads. As we set out above, over two years both MSE and Action Fraud had over 1,000 reports each of fake Martin Lewis ads alone. This is just one type of scam, and just two bodies relying on consumers to actively report them. It is therefore fair to assume that the number of actual scams of any type – reported or not – is drastically higher.
2. We have seen many accounts from our users, and also from our own experience, that fraudulent adverts are not being removed, even when they are reported to platforms. This shows that even fake ads that are undetected by services' systems but are manually flagged STILL remain on platforms – more evidence of this is given in the questions on reporting below.
3. Platforms are paid by scammers to publish their fraudulent advertisements. Yet there still remains no enforceable, practical regulation to financially disincentivise platforms from allowing their users to be harmed in this way.

We recognise that it is impossible for platforms to curb every single scam ad. However, the sheer torrent of scam ads seen by the public shows that platforms are simply not doing enough to detect and remove fraudulent ads from their site.

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

Response:

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?
It is not within MSE's expertise as a consumer group to answer this question specifically, but we would like to reiterate that the solution to preventing scam ads from appearing on platforms does not have to be technical. If the technology does not exist or if there are reasons why it would not be effective or practicable to use, then platforms should not be able to use this as an excuse for not acting. Indeed, if the only way for platforms to prevent fraudulent adverts from appearing on their site is for a human to vet them beforehand, then that is what should happen. Platforms take a huge amount of revenue from scammers who pay them to publish this content, and therefore should be able to put in place a system – in whatever form that they can – to significantly reduce instances of criminals slipping through the net. We are now in a situation where the platforms and the criminals themselves are financially benefiting from fraudulent content, while consumers and other industries pick up the bill.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 38: If you have information/evidence/suggested mitigations to share which may be useful in the preparation of codes of practice, which is not covered by the questions above, please include these under 'Overarching considerations'.
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

For providers of online services

Question 39: What proportion of all paid-for advertising on your service is identified as fraudulent advertising?
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 40: Does your service take any steps to warn users of the risk of encountering fraudulent advertising or to educate them about how to identify potentially fraudulent advertising?
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 41: Please provide information regarding the proportion of successfully identified fraudulent advertisements that are identified via:
(a) automated systems
Response:
(b) human processes
Response:
(c) user reports
Response:
(d) other (please provide further detail).
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 42: What is the average and/or median time taken between the identification of a fraudulent advertisement and its removal/other actions taken? (If other actions taken, please specify what they are).
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 43: Proactive technology

For all respondents

Question 43: Please provide any evidence you have regarding proactive technologies which could be used to identify fraudulent advertising activity. In particular, we are interested in information related to the following points:
(a) The kinds of proactive technology which are/could be applied to identify or prevent fraudulent advertising
Again, MSE does not have the expertise to suggest technological solutions – but we reiterate that if no such technology can be found, a manual solution must be put in place by platforms.
(b) A brief description of how these technologies are/could be integrated into the service
Response:
(c) The effectiveness, accuracy and lack of bias of such technology (including compared to alternative proactive and non-proactive methods) in relation to detecting fraudulent advertising and accounts which post fraudulent advertising material

Response:
(d) How proactive technologies are maintained and kept up to date
Response:
e) Information related to the associated time and/or costs for set-up, operation, and human review
Response:
f) The cost of integrating such technologies: (a) for the first time; and (b) when updating these technologies over time
Response:
g) Whether there are cost savings associated with these technologies
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 44: Advertising onboarding and verification

For all respondents

Question 44: Please provide any evidence you have regarding the processes for advertiser onboarding and verification related to protections against fraudulent advertising. In your response, please indicate whether these processes are currently implemented in respect of services which are in scope of the Act or whether they stem from another sector In particular, we are interested in information related to the following points:
(a) The criteria which advertisers are verified against, including documentation/evidence used to support verification, and what advertisers are required to declare
It is our understanding that platforms have widely different criteria for verifying advertisers, and it's no surprise that we have seen the most issues with fraudulent advertisements on platforms with the weakest verification criteria. To the best of our knowledge: • Facebook appears to allow almost anyone with a bank card to sign up and publish adverts on its platform. Some of the sign-up options seemingly have little to no advertiser verification processes before they can begin to use advertising tools. • X (formerly Twitter) requires you to pay for a premium account. • TikTok asks advertisers to verify their businesses information with official documents before they can start advertising. In the recent past, MSE has encountered actual and reported scam ads on Facebook far more frequently than on the other two platforms mentioned. We would like to see a uniform policy across platforms, and ideally the sharing of information between services, so that would-be advertisers who are identified as bad actors and blocked on one platform are also 'blacklisted' on others. This would have the additional benefit of reducing

the amount of time each platform would have to take vetting, and ultimately reduce the risk that fake adverts would be seen by consumers.
(b) The role of (a) automated processing and (b) human processing in the verification process, and how they interact
Response:
(c) The costs associated with advertiser verification and how those costs vary as scale increases
Response:
(d) The percentage of advertiser accounts that are verified
Response:
e) Whether advertisers are permitted to publish advertisements on the service while the verification process is ongoing
Response:
f) Whether there are additional/specific verification checks for advertisers placing adverts of certain kinds or targeting certain audiences, such as about specific products or services, or targeting users under the age of 18
Response:
g) Whether the verification of an advertiser account expires after a certain amount of time or certain activity, such as when advertisers make changes to their account or profile
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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

For all respondents

Question 45: Please provide any evidence you have regarding the processes that services in scope of the Act have in place to review submitted paid-for advertisements and identify fraudulent advertising material. In particular, we are interested in information related to the following points:
(a) The percentage of submitted advertisements which are reviewed both (i) prior to and (ii) after publication
Response:
(b) The role (i) automated processing and (ii) human processing play in the review process and how they interact
Response:
(c) The red flags which trigger advertisement review processes both (i) prior to and (ii) after publication and the basis on which those red flags are selected

Response:
(d) The timescales for review
Response:
(e) What happens to the advertisement’s visibility and reach, if it is flagged as suspected as being fraudulent (either by a user or automated system)
Response:
(f) The costs associated with the review of submitted paid-for advertisements
Response:
(g) Whether trusted flagger reporting is employed to inform services’ review processes. If it is, how is it applied, what guidelines / criteria does it follow, and who are those trusted flaggers?
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 46: Advertiser appeals of verification/review decisions

For all respondents

Question 46: Please provide any evidence you have regarding advertiser appeals of verification/review decisions relating to fraudulent advertising on services in scope of the Act. In particular, we are interested in information related to the following points:
(a) The role of (i) automated processing and (ii) human processing in the appeals process, and how they interact;
Response:
(b) The level of proof required for an appeal to be accepted;
Response:
(c) The most frequent bases for appeals against sanctions decisions on fraudulent advertising content
Response:
(d) The ratio of decisions that are appealed against
Response:
(e) The costs associated with appeals
Response:
(f) The proportion of appealed decisions which are upheld and overturned
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)

Response:

Question 47: User reporting mechanisms

For all respondents

Question 47: Please provide any evidence you have regarding user reporting mechanisms for fraudulent advertising on services in scope of the Act.

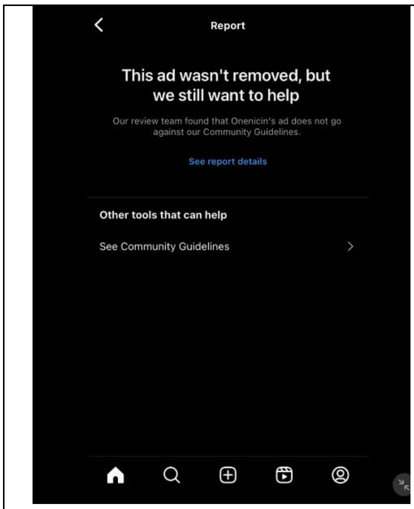
In particular, we are interested in information related to the following points:

(a) What user reporting tools there are for paid-for advertisements, and how these tools differ from those for user-generated content and/or search results and other search functionalities that are not paid-for advertising

As a result of a settlement, following a defamation lawsuit brought against Facebook by Martin Lewis, Facebook agreed to launch a scam ads reporting button, unique to the UK, in 2019. The intention behind the button is for users to have an easy way to flag fraudulent content when they come across it, and for Facebook to see when many users are flagging the same content – suggesting it should be removed urgently. Prior to this, users were expected to find an ad reporting form in the platform's help pages, and then to manually provide unreasonable levels of information (including ad click strings that may no longer be visible).

We would like to see a uniform method of reporting across all platforms, where with one click, users can report fraudulent content which is then swiftly removed. Ideally, this process of removal is automatic, and if necessary, a review of the ad can take place after the content has been taken down. The report button should be easily identifiable as such (universally recognised if possible), and users should have the same experience reporting and having their report reviewed across all platforms, so that fraud cannot slip through the net on one platform more easily than another.

However, the reporting button only works well if platforms respond effectively to users flagging suspected fraudulent content. Users have told us that that they have flagged a post as fraudulent on multiple occasions but no action has been taken; reasons given have ranged from the platform not being able to review reports due to higher volumes, to the advert being deemed as not going against the platforms' advertising standards/guidelines. In one case, an MSE user told us they reported almost a dozen fake ads featuring Martin Lewis in one day, and every single report was marked as 'not going against the community guidelines', resulting in no action being taken against the advert or the advertiser. When they requested a review of this decision, they were met with an error message each time. *Examples of the messages received by this person are shown below:*



Report

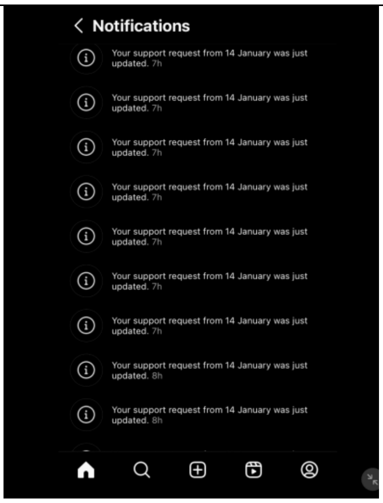
This ad wasn't removed, but we still want to help

Our review team found that Onenicin's ad does not go against our Community Guidelines.

See report details

Other tools that can help

See Community Guidelines



Notifications

Your support request from 14 January was just updated. 7h

Your support request from 14 January was just updated. 7h

Your support request from 14 January was just updated. 7h

Your support request from 14 January was just updated. 7h

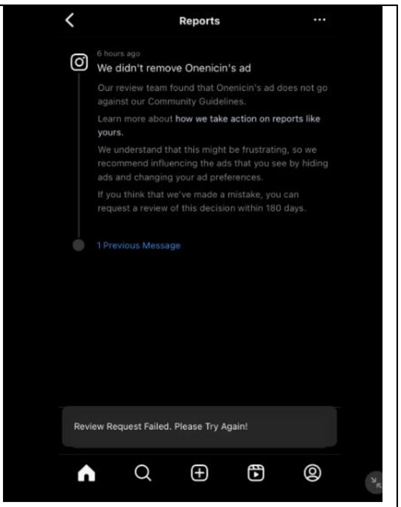
Your support request from 14 January was just updated. 7h

Your support request from 14 January was just updated. 7h

Your support request from 14 January was just updated. 7h

Your support request from 14 January was just updated. 8h

Your support request from 14 January was just updated. 8h



Reports

6 hours ago

We didn't remove Onenicin's ad

Our review team found that Onenicin's ad does not go against our Community Guidelines.

Learn more about **how we take action on reports like yours.**

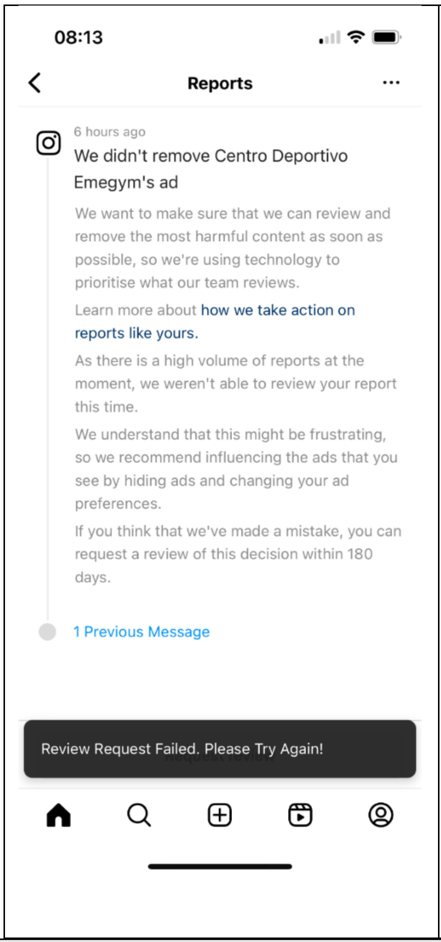
We understand that this might be frustrating, so we recommend influencing the ads that you see by hiding ads and changing your ad preferences.

If you think that we've made a mistake, you can request a review of this decision within 180 days.

1 Previous Message

Review Request Failed. Please Try Again!

Further examples of the responses sent to users after they reported a suspected fraudulent advertisement:



08:13

Reports

6 hours ago

We didn't remove Centro Deportivo Emegym's ad

We want to make sure that we can review and remove the most harmful content as soon as possible, so we're using technology to prioritise what our team reviews.

Learn more about how we take action on reports like yours.

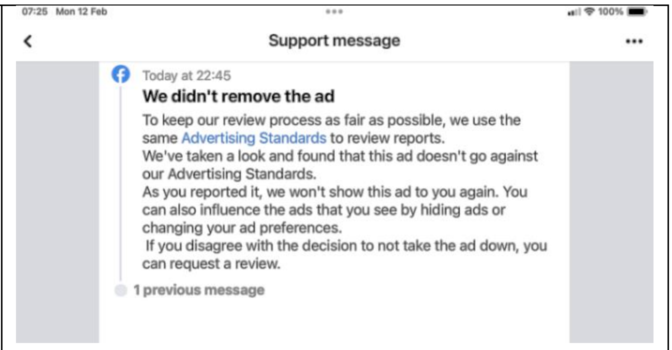
As there is a high volume of reports at the moment, we weren't able to review your report this time.

We understand that this might be frustrating, so we recommend influencing the ads that you see by hiding ads and changing your ad preferences.

If you think that we've made a mistake, you can request a review of this decision within 180 days.

1 Previous Message

Review Request Failed. Please Try Again!



07:25 Mon 12 Feb

Support message

Today at 22:45

We didn't remove the ad

To keep our review process as fair as possible, we use the same Advertising Standards to review reports.
We've taken a look and found that this ad doesn't go against our Advertising Standards.
As you reported it, we won't show this ad to you again. You can also influence the ads that you see by hiding ads or changing your ad preferences.
If you disagree with the decision to not take the ad down, you can request a review.

1 previous message

(b) What percentage of user reports of advertisements relate to suspected fraudulent content, and the processes for taking action in relation to such reports

Response:

(c) Any statistics you can share on (i) the number of user reports of suspected fraudulent advertising received and resolved over a specific period and (b) the number of initial decisions appealed by users who made the report

Response:
(d) The criteria used to classify and prioritise user reports
Response:
(e) The median and/or average time it takes to respond to a user report, and any measures that are in place to ensure timely and accurate responses to user reports
Response:
(f) Any measures taken to make user reporting tools accessible, easy to use and easy to find for users
Response:
(g) How transparency and communication is maintained with users who have submitted reports
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 48: Use/involvement of third parties

For all respondents

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

Question 49: Generative AI and deepfakes

For all respondents

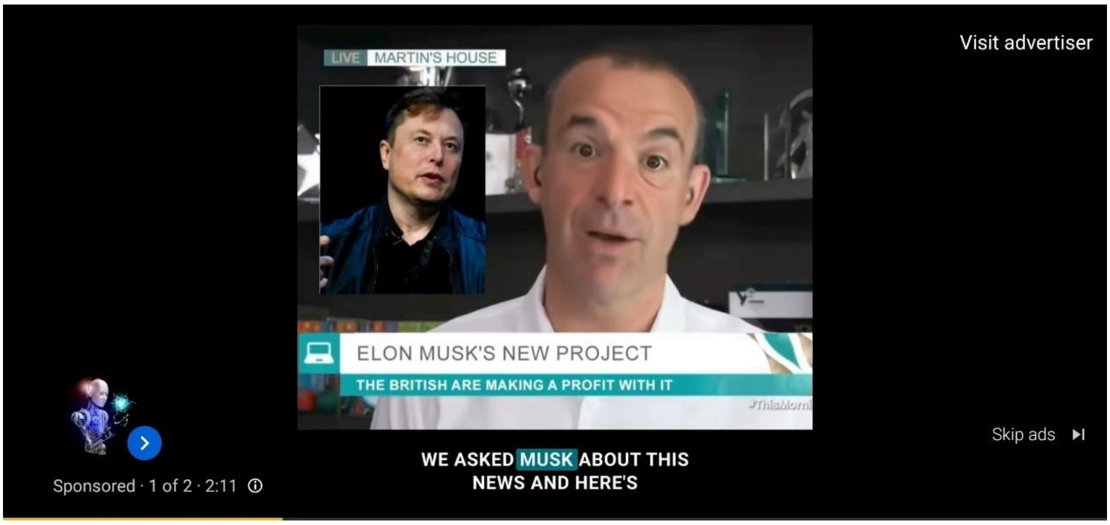
Question 49: Please provide any evidence you have regarding the impact of generative AI developments and deepfakes on the incidence and detection of fraudulent advertisements on services in scope of the Act. In particular, we are interested in information related to the following points:

(a) The frequency of deepfake fraudulent advertisements’ occurrence, in absolute terms and/or as a proportion of all fraudulent advertisements, and how you expect this to evolve in the future

From MSE’s internal monitoring, we started seeing deepfake fraudulent adverts around July 2023.

As soon as we were made aware of the scam, by someone reporting it to Martin Lewis on Twitter (*now X*), MSE published a news story warning our users of the scam and reiterating that Martin does not (and never has done) adverts.

The most prominent deepfake scam we’ve seen is of Martin Lewis appearing to endorse a new investment scheme, supported by Elon Musk. At the start of the video, a deepfake of Martin Lewis, ‘live’ from his home-office, gives an introduction to the ‘scheme’. The deepfake video and voiceover purporting to be Martin ‘says’: “Musk’s new project opens up great investment opportunities for British citizens…. We think it’s safe to say that the experience is legit”. The video then moves on to a deepfake video and voiceover of Elon Musk, who ‘says’ people can make “$500, $1,000, $3,000 in the first day. He goes on to ‘say’ that celebrities such as Lewis Hamilton and Cristiano Ronaldo have already invested. An example of one of those videos seen by a user on Facebook can be found on X (formerly Twitter). Here is an example of a screengrab from a different occurrence of the same deepfake advertisement:



This type of deepfake video has been circulated and recreated many times since July 2023. To give an idea of the scale: before we saw the first deepfake scam ad (pre-July 2023) MSE received 20 reports of fake ads that featured Elon Musk in the last 7 years (2017 to July 2023). In less than a year since July 2023, MSE has received 93 reports of Martin Lewis/Elon Musk scams.

While we cannot predict with certainty what will happen in the future, our experience of fraudulent advertising trends in the past suggests that these scams will only continue to grow in sophistication and reach, likely using technology that isn’t widely used and of which we may not even be aware.

In particular, there is a shared concern among many that fraud may be ‘personalised’ using AI in the very near future, allowing criminals to massively scale up their fraud operations and reduce the need to recruit humans to interact with victims. It’s important that platforms do not enable this – with the current lack of enforcement, a single advertising campaign could lead to a great deal more victims.

(b) What methodologies/technologies are currently employed to detect fraudulent advertisements which include deepfake or otherwise AI-generated content, and the effectiveness of these tools
Response:
(c) Whether detection technologies are developed in-house or acquired from a third-party, and how long it takes to develop and/or integrate those tools into wider systems
Response:
(d) The accuracy of detection methods, including true positive and false positive rates
Response:
(e) The costs associated with the development/acquisition and deployment of these detection mechanisms
Response:
(f) The types of deepfake or AI-generated content (in terms of either media type or subject) in fraudulent advertisements that are most difficult to detect i) via automated processes, ii) by human moderators, iii) by service users
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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

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

For all respondents

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

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

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

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

For providers of online services

Question 54: What kinds of information do you provide and how do you provide this information? In your response to this request, please provide information relating to (a) where relevant.
Response:
a) **If there are certain types of information you cannot provide, please explain why, for example whether there are technological, cost or privacy factors that mean certain kinds of information may not be feasible to provide**
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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

Questions 56 and 57: Complaints systems

For all respondents

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

For providers of online services

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

Question 58: Evidence

For providers of online services

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