Your response: Please indicate how much of your response you want to keep confidential. Delete as appropriate.	
For confidential responses, can Ofcom publish a reference to the contents of your response?	N/A

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:

1. Write them Better

The terms of Service and policy statements are still often a) too long, b) use too much jargon and overly complex terminology. We want them to use simpler more straightforward language that the average user can understand. Instead of presenting legal language in isolation, provide context by explaining the purpose or rationale behind certain clauses or terms. This helps users understand why certain provisions are necessary. Platforms should provide a summary or TL;DR: Include a "Too Long; Didn't Read" summary at the beginning of the ToS, outlining the most important points in a concise format for users who may not have the time or inclination to read the entire document.

2. Make smarter use of visuals

We think that in general the majority of terms of service could make far better use of visual aids, illustrations and infographics to supplement textual content and reinforce important concepts. This is particularly important given that as platforms evolve so many of the infractions that are being described relate to visual media – memes, inappropriate use of photography – its far easier and comprehensible to communicate similarly.

3. Provide examples and case studies (ideally illustrate) that exemplify both common examples that do and do not breach your terms of service

Many Terms of Service feel highly theoretical and it can be difficult for the average complainant to parse how this would apply in their specific set of circumstances. Platforms should use real-life examples to illustrate how certain terms or clauses apply in different situations. This can help users grasp abstract concepts more easily.

Platforms could be far more transparent about their complaints and could for example provide their 'top 5 rejected' complaints/reports – on the basis that if high volumes of type X complaint are being made there are common misunderstandings of what and what is not considered acceptable.

4. Invest in better synthesis and analytical tools for users to help them align complaints and reduce the numbers of 'mis-complaints' and misreporting.

Social media platforms often have a plethora of different public policy statements all of which inter-relate with each other. If we take X as an example over time it has had separate policies for a hateful conduct, abusive , harassment, hateful imagery and symbols, violent threats and dehumanisation to name but a few. But this creates a

number of challenges for ordinary complainants seeking to make a report. We hear that players and their teams may often have a complaint rejected when submitted under one policy but later accepted when re-submitted under another policy. This is highly demoralising and likely has a cooling effect on the number of valid infractions that are reported. We think platforms should invest in categorisation tools that can take the report, ask a range of triage questions, enable examples of documented abuse to be uploaded and then analysed automatically against all the policies to do more of this legwork on behalf of the complainant. Whilst AI is now commonly used platform-side in the assessment of people's reports we would suggest it could also be used to support and improve the initial submission and make it easier of the user to navigate what is at the moment a labyrinth of different policies. Our experience has shown ambiguity is problematic for everyday people. It is manifestly wrong that reports are often rejected not because they don't reach a platform threshold but because the complainant fails to identify the best infraction. We have examples of reports rejected under one policy and then acknowledged as an infraction of the rules under another.

5. Provide better version histories - Clearly indicate when the ToS was last updated and provide access to previous versions for reference with simple summaries of the core changes that were made and the rational for those changes. We often see that platforms are publicising updates on arcane hidden blogs and expecting users to track them down. Think for example about the significant changes made by X and Instagram this April. Huge changes which no average user has any idea about.
6. Proactively alert to changes to Service – We believe that platforms should be required to re-obtain consent after a material change in behaviour on the platform. This should be done in an easily accessible format which is brought immediately to the users attention by mechanisms including for example pop ups, prompts or other easily navigable interfaces from the 'customer support' toolkit.

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.

1. **Worked Case Studies of successful reported infringements linked to associated Sanctions** This is an expansion of point three above which is about the lack of worked examples and case studies provided by platforms to support users and potential complainants. At the moment platforms aren't transparent about the

prior actions they've taken and the real world examples that help to show users in context what issues actually fall foul of TOS or platforms policies. Showing a range of examples and case studies that make hierarchies of infringement visible and link those case studies to levels and severities of sanction would really help users understand what they can expect.

2. **Platforms should include tools that enable users to call up a range of examples of upheld complaints against any policy in which they are interested**. This functionality is absent on most platforms. The question is not just about level of detail but how easy it is to search and find pertinent details amongst the mass of irrelevant information that might not relate to a user's complaint. The issue is about whether a user has smart (probably AI supported) tools that quickly enable them to judge whether the issue that they have come across is above or below the threshold.

3. **Clarity about the escalation path for each policy**

   Platforms do not hold all infringements equally. The same level of sanction is not meted out for each breach. Platforms need to make these hierarchies clearer. It should be simple for users to understand the journey that their complaint could take for a certain type of complaint.

4. **Train the Trainer** – Not all training and support for users should or could be delivered by platforms - they are not independent enough and they cannot form supportive relationships with every vulnerable or exposed user group. For footballers the right support mechanism might be their club or union. But where platforms should be active is in providing adequate briefings and training that can be adapted by these trusted agencies (clubs/unions) to keep players safe and informed of what they can expect when using social media platforms which are so essential to their careers and how they reach out and communicate with their fans and audiences

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 nonenforcement 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:

To increase awareness of reporting and complaints mechanisms providers could:

Support the growth of a healthy third-party training ecosystem –This would support the upskilling of important sub-groups of users about a) Proactive safety tools b) changes to terms of service/policy and therefore what counts and what does not count as infringements c) overviews of the reporting process

Platforms could also support and fund the development of third-party tools for countering abuse on the basis that civil society organisations with diverse membership groups (Older people, those from minoritised ethnic groups and women) might be best incubating their own tools. These could then be considered by platforms postdevelopment as the source of future innovation.

To make user reporting and complaints systems effective you need to have a plethora of things that are not in place on most platforms today.

1. Make the process clear - most platforms do not share the end to end reporting process

   Our experience is not the typical user experience. We are professionals working in anti-discrimination and often we are involved routinely in requesting the takedown of abusive material. That professional focus gives us and those we consulted a very different level of a) familiarity and b) confidence navigating reporting processes. This is not the case for the average individual facing abuse. Our view is that most reporting processes on the main platforms of interest (X, Instagram, YouTube) are really unclear, frustrating, poorly signposted, and difficult to navigate. In our view platforms could get much better at visualising, describing and providing tools to make the user both comfortable and unphased by each step of the process. How might this be achieved? We think there are some relatively simple fixes. a) Set out each step of the process and what is possible at each step

b) make it possible to call up for assistance if aspects of this process are unclear – have clearer and enhanced messaging facilities to do this. The Airbnb complaint escalation process for hosts is a good example of how this can be done well with a blend of automation and human interaction. C) Provide some form of visual tracker so that people are always clear where they are in the process and what comes next. This could be a simple 'Progress Bar' showing the proportion of the process that has been completed and what is left to go.

2. Make documentation, archiving and retrieval of abuse simple in-platform. One of the basic issues for complainants is the difficulty of proving that the abuse has happened/is happening. To effectively provide proof you need to capture the evidence so it can be shared with the platform but potentially also with prosecuting services, lawyers or employers. At the moment we think that the ability to do this within platforms is pretty tricky if not impossible. In practice this means people are using screen shots – saving images of abusive messages into apps like 'notes'.

   We think complainants deserve to have storage and capture mechanisms that enable them to collect and save details of the abusive content/abusers account – with key information whether it be time, volume, type of abuse. The resulting material should be easy to review and retrieve and share with whomever the complainant needs to share it with.

   In our experience working with the UK Football Policing Unit (UKFPU) or police forces like the MET it has become clear that the ID/reference numbers giving law enforcement access to the raw data is not sufficient. The police are visual and they insist on having actual screenshots to see the abuse in context. As such it is invaluable to have an image of what was actually sent to a player. This is the quality of material needed for an evidence pack and it is also the way in which prosecutors can show the court something and ultimately show the judge the impact of what's being done.

3. Enhanced delegation tools to allow trusted supporters (clubs, business managers, unions, family) to have access to support documentation and reporting in a way that doesn't give whole account access and reduced re-traumatisation. For example Signify AI discovered a work around to the current systems on X and has a product that allows protection from and filtering of DM's sent to players from unsolicited recipients. This is exactly the type of third party tool that platforms should be utilising to enhance in platform safety.
4. Make documentation tools inter-operable with reporting tools
5. Provide preview and editing functionality so that users can review their complaints before final submission so that they can add in additional missed context or can upload additional evidence. We have heard of cases where in the absence of effective guidance and sign posting complainants have had their reports submitted when they were not ready. This should not happen. Include 'back' button

functionality so that if this happens by accident people can append additional information without having to start the whole process again.

6. Once a report is submitted the complainant should have the ability to view it and have an understanding of what stage of processing that report is in. We have seen examples where the only way in which a complainant has known about the decision of the platform was by checking the status of their harassers account and seeing if a) the content is still live b) whether the account has or has not been suspended. Platforms can and must do better.
7. Have digital '999' services that provide special services when a user is experiencing an abuse 'storm' or 'campaign' i.e. high volume of abusive behaviour over a short period of time a) access to a human support team b) mechanisms for expedited processes to complement longer investigation process.
8. Introduce case-linking functionality so that if abuse evolves it's possible to have separate instances considered vis a vis their cumulative impact. We don't think that platforms give enough and in some cases any tools to either indicated they are dealing with an abusive 'pile on' by what we have termed 'Troll Commanders' users with high numbers of followers who effectively give a nod and a wink to their users to troll somebody (but the messages which instigate the pile on may have plausible deniability). In effect this is the use of the army of followers as a bully pulpit to silence others. Whilst X has some functionality for batch reporting of abusive posts this is not universal across platforms. Equally we do not think there is enough capacity to link reports if the complainant realises retrospectively that what they thought was a single abusive or problematic communication is actually now part of a wider campaign. It maybe that for context the complainant might wish to show that the abuse has been tracking them on many platforms. For clarity – we are not suggesting that platforms should be policing other platforms but simply that they might want to take into consideration the breadth as well as depth of the reported behaviour.
9. Building on the last point there needs to be a enhancement to the tool above to be able not only to link multiple cases but to link perpetrators who work together to abuse. We are specifically thinking of the complex relationship between "Troll Commanders" and their 'Troll Foot Soldiers'. We have seen numerous cases where members of those troll armies - these foot soldiers then reporting up to the 'commanders' in effect trying to get the troll commander to take up the cause to set the troll army on to a particular user. This is not just an issue for football this was prevalent online in the Amber Heard/Johnny Depp trial. We would like to see development of reporting mechanisms that allow these type of relationships to be captured and recognised in the reporting process. We want to explore how it might be possible for the platform sanctions to recognise the roll and impact of both those who incite mass bullying and those who enact it.
10. Ensure that activating protections and indeed sanctions do not destroy evidence or documentation needed for future prosecution.

11. We found that platform providers need to do more to ensure that perpetrator/users had less ability to destroy evidence of poor behaviour. A perfect example is account deactivation. From our understanding when accounts are deactivated by perpetrator-users, it prevents law enforcement from being able to 'option' account details which could lead to an identification and potential legal action. By simply deactivating their account, users can avoid detection and repercussions from the content they have posted and can then create a new account shortly after and continue posting abusive content. If account data for deactivated accounts remain available and archived for a specific time period after, this may help law enforcement positively identify individuals. We believe this archiving facility should be considerably longer if there is an open complaint from another user in progress which should suspend the timetable for deletion.
12. Offer effective complaint tracking tools so that users understand exactly where they are in the reporting process. This 'review dashboard' should be easily visible. It needs to be centrally located – probably near the user settings menu. It needs to be easy to locate and show all information about current and past cases in a simple format.
13. Platforms should commit to resolving complaints within a specific, reasonable timeframe. Delays should be minimized, and users should be informed of any expected wait times.
14. When a decision is made on a report the complainant should have a clear understanding of why that decision has been made. Feedback on why a report didn't meet the threshold set out in the TOS or the relevant policies is crucial.
15. Make the link between levels/types of abuse and sanctions very clear. Our players and others within football were unable to establish the criteria for either the removal of content or indeed account suspension.

(a) reporting or complaints routes for registered users, non-registered users and potential complainants (being affected persons who are not users of the service)

Response: We think that reporting pathways/routes should be clear to all users registered or not - so they can understand what support is available to them and what outcomes can be achieved.

(b) how to ensure that reporting and complaints mechanisms are not misused

Response: There is some evidence that people have been the victims of targeted reporting to try and silence them and act as a chilling effect on views that vocal groups do not agree with. There are examples of this on debates around Trans right and also high profile, controversial, events like the Johnny Depp/Amber Heard trial.

Platforms need to ensure that there can recognise and "de-fang" malicious reporting flash mobs whilst reserving the ability to recognise genuine event driven reporting spikes. X's new policy Freedom of Speech Not Reach could be an over correction in this area.

(c) the key choices and factors involved in designing these mechanisms
Response: Transparency of 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: 1. Again – we emphasise the importance of making reporting mechanisms clear and easy to follow. 2. Make it possible for others to report on behalf of the victim and ensure that those routes do not automatically diminish the degree of sanction meted out. 3. Do not rely on automation. The current process after reporting content is automated. We suggest that for reporting abuse relating to protected characteristics when a user is reporting on behalf of themselves or where others report due to suspicion of self-harm - that platforms should have additional direct human support to manage and support those who are reporting.
(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: All reporting should be anonymous and users should be informed of this. There seem to be very few circumstances where informing a party of the ID of someone who has reported their content would not lead to negative consequences.
(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: Platforms should provide search and response tools (AI enabled) where users can ask questions and queries about the reporting and complaints process and get high quality natural language responses that clarify any things that are unclear. Once exhausted there should be the possibility of escalation to human support service.

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 revieweing 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: Deployment of generative AI but also linking to verified or self-declared user characteristics. The issue of reclaimed language is legitimate but often used as an excuse to not deal with problematic content. Equally terms have been removed that were legitimate for the author to use. Particularly in academic contexts. Reports could be referenced against author user data to help improve this.
(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: There could be a benefit to have people who have an expertise (academic and researchers) to supplement the work of moderators who are experts in applying the internal policies to a specific case but might not be subject experts so might struggle with the difference in context between LGBTQIA+ issues playing out in Scotland versus Ghana (where homosexuality receives legal sanctions). We think platforms could meaningfully reflect on what they could do to utilise subject matter experts either on quality assurance processes on decisions but also perhaps for capacity building for the core moderation staff.
(c) any other information.
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)

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: Moderators with knowledge of the area they are working on do a far more accurate job. However, often those with the greatest knowledge share characteristics that

have led to the abuse in the first place so this also exposes them to greater risk of trauma in dealing with it.
(b) how services can/should/do assess the accuracy and consistency of human moderation teams;
Response: Sampling quality control should be deployed across individuals and also teams.
(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: We know that content moderation for online social platforms relies heavily on workers from the Global South, especially from the Philippines, India, and several countries in Latin America and Africa. The Philippines is a major hub due to its large English-speaking population and lower labour costs, making it a key location for companies like Facebook and Google. India also plays a critical role, leveraging its vast pool of English-speaking professionals and cost-effective labour market, particularly for regional language moderation. Additionally, countries in Latin America, such as Mexico and Colombia, and several African nations, including Kenya and Nigeria, are increasingly important for their economic advantages and multilingual workforce. These regions provide essential support for global tech companies, ensuring that content moderation is conducted around the clock and across various languages. However, this work often comes with significant challenges, including poor working conditions, psychological stress, and insufficient support for the moderators handling potentially traumatic content. As an organisation that is human rights sensitive particularly around labour violations that can and have happened in support of the beautiful game we think it is crucial that platforms consider how they are taking care of the wellbeing of a largely non-white workforce. Providing 24/7 access to counselling, resiliency training, and ensuring workflow flexibility can help mitigate the negative impacts. Regular breaks and avoiding the moderation of highly egregious content from home are also recommended to protect moderators from burnout and secondary trauma

(Sendbird) (SpectrumAI)(Sendbird) (Strixus).

(d) whether training is provided and/or updated (including for emerging harms), and the frequency of these updates;

Response: Training is vital. The landscape of abuse and threat is also evolving so training should be refreshed regularly and be an evolving process.

(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: We do not have enough intelligence to make suggestions on this point.

(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: Employing issue and moderation experts will greatly enhance services.

(g) any other information.

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: You have made this a specific question that asks about the effectiveness of platforms' appeals processes where journalistic content is at question – implying that general appeals processes are functioning well. We do not agree. We think there is a case for considering huge improvements to complaint, appeal processes for all users and relating to all reporting of abusive or discriminatory content.
Current mechanisms for user appeal of content moderation decisions are insufficiently effective. It is our view that even though people may be using social media on a daily basis and may have made a report they may not have realised that seeking resolution or redress following a disputed content moderation issue was even possible. We need this to change. Users need to know this is an option available to them.
(a) examples of effective redress mechanisms that you consider important

Platforms should:

1. Have robust, independent appeal process where users can challenge decisions post report. Appeals should be reviewed by a different moderator or an external body to ensure impartiality.
2. Establish independent oversight bodies to review complaints and appeals, ensuring unbiased resolutions and holding platforms accountable for their moderation decisions.
3. Engage civil society organizations in the oversight process to ensure that diverse perspectives and community standards are considered.
4. Consider creating dedicated support teams to handle appeals, and users should receive clear, detailed explanations of decisions
5. Provide comprehensive guidelines on appeal processes and conduct awareness campaigns about user rights and available redress mechanisms

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

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: In general we do not think that platforms do enough research or transparency reporting relating to any aspect of the evolution and nature of their a) tools development

processes, b) the evolution and improvement of their moderation processes c) how they have improved changed their reporting mechanisms d) how they are working on their appeals processes e)how they engage civil society to QA their work f) the overarching statistics relating to the volumes/numbers and types of issues they deal with. I think Ofcom should consider whether there is a certain minimum level of reporting/briefing and essentially audit that firms should be mandated to do on an annual or biannual basis.

Beyond this there should be some thought about the level of access researchers and third parties should have so that they can conduct meaningful independent research on their work.

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: No comment at this point
(a) whether there are any additional measures/ safeguards that can be put in place during local or national elections
Response: No comment at this point

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.
No comment at this point

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:

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: Platforms deploy a variety of mechanisms to help control content and interactions on platforms. For adults this is primarily restricted to limiting or moderating comments on your posts or page, limiting people tagging you in posts/photos or blocking users you find objectionable. There are sensitivity filters, for example on Instagram, however these do not restrict content from being serviced, but rather present it in a censored way that allows a choice to reveal.

Some platforms including only seeing replies from verified accounts. Or restricting comments only to those who follow you. Essentially we know that users need to be able to restrict (hide abusive content) block (cut off all access between a perpetrator and their account) or mute (hide content from themselves personally but not necessarily others:

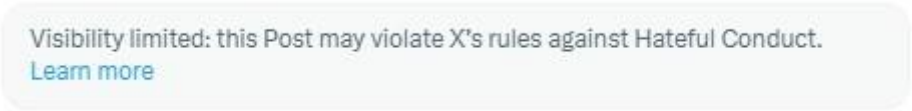
Below we list a number of features that we find helpful are listed below:

X (formerly known as Twitter)

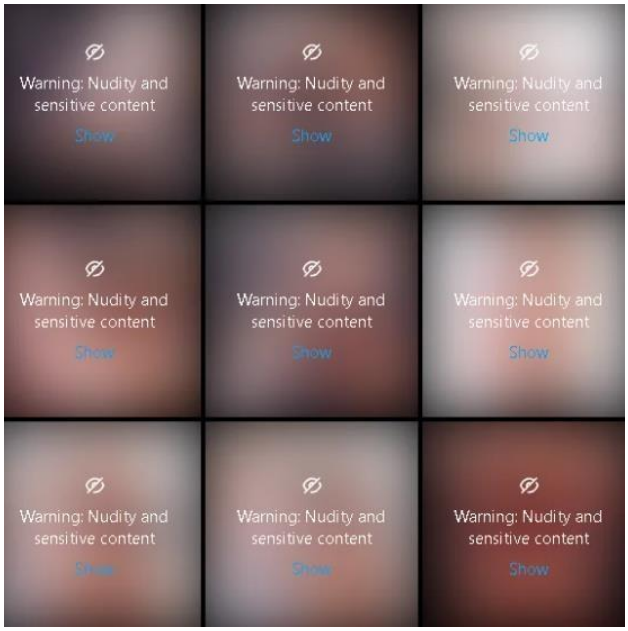
We think the function where users can manage their own settings for what they can see online is helpful. We think that it is positive that users can adjust settings if they want to post sensitive content, and then label the content with either Nudity, Violence or Sensitive markers. In some cases these tools do not work across all types of content we think this should happen so any shield/bespoke privacy security settings should work across feeds, comments and direct messages.

These configurations must be customisable and it should be possible for a user to toggle between different settings. We think where practicable the tools should incorporate 'preview' function that allows users who are re-configuring their stings to show how the adjustments affect, reach/appearance of post.

- Visibility labels are applied to content which is harmful but not illegal.

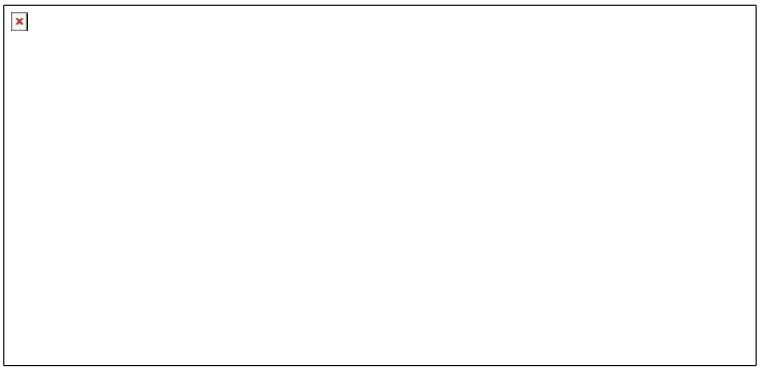


- Blur images in Media section which may be harmful, users must opt in to seeing the images.



- Blurring restricted accounts so the user must opt in to seeing the full profile.

- Users can block accounts, mute accounts, and mute words to protect themselves from seeing particular content.

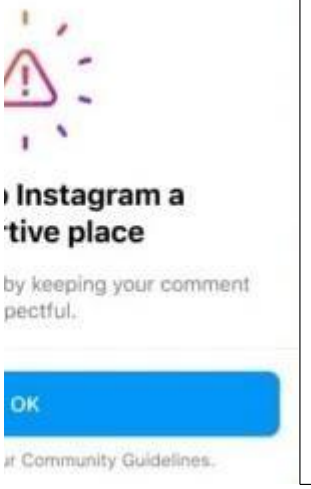


Instagram

- Users can control what sensitive content they can see from accounts you don't follow.

- offer some educational points as to

made to think and question about what they were going to post. Th
be harmful.



Other examples of prompts include:

informed discourse.

Did you mean to share this without reading?" when users try to sh
rs to read content before sharing, promoting

retweet.

s that contain potentially misleading information, guiding users to
em to click through to read more before they

YouTube: Implements age-

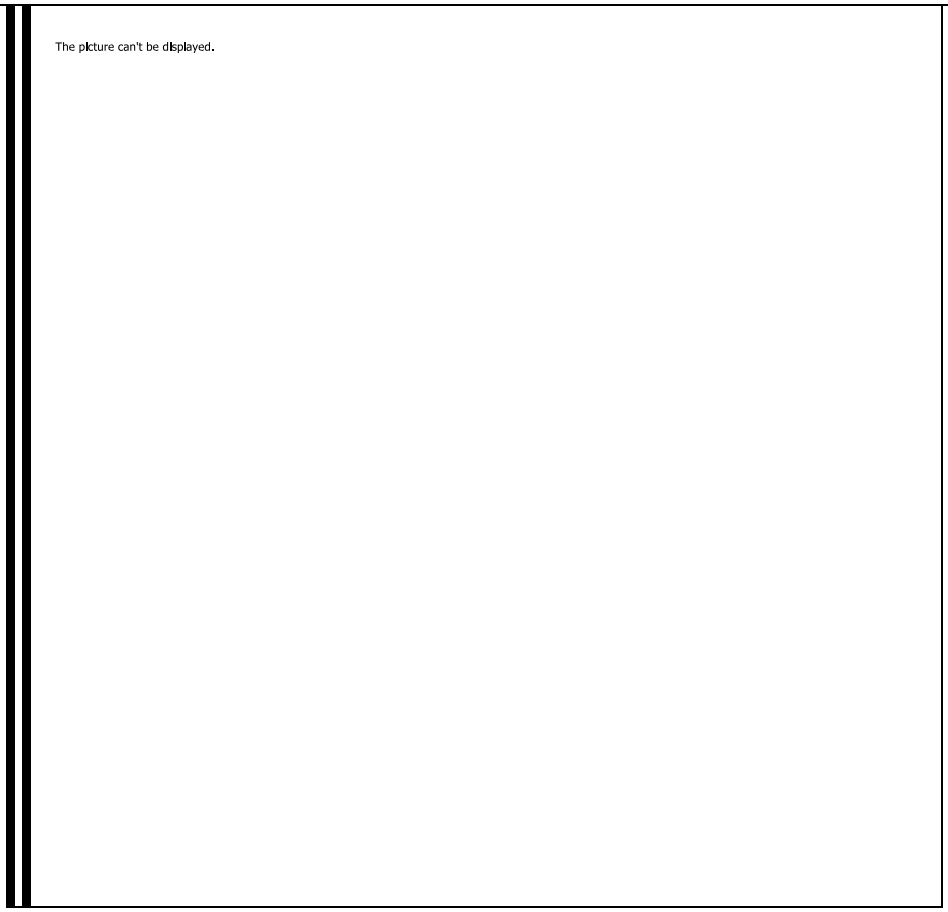
estriction labels and community guideline reminders before users
certain types of content, nudging them

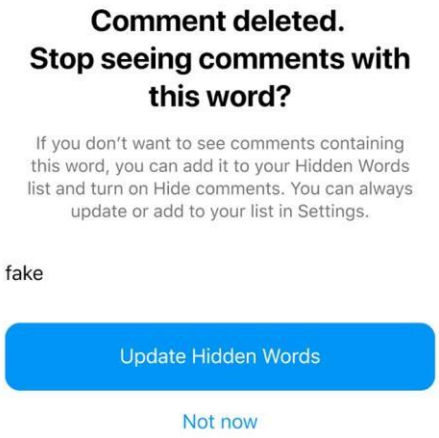
- edit. the hidden words list.

nents, hide comments, and have a custom words list which they ca
nts with the same word in, they can add this to

to

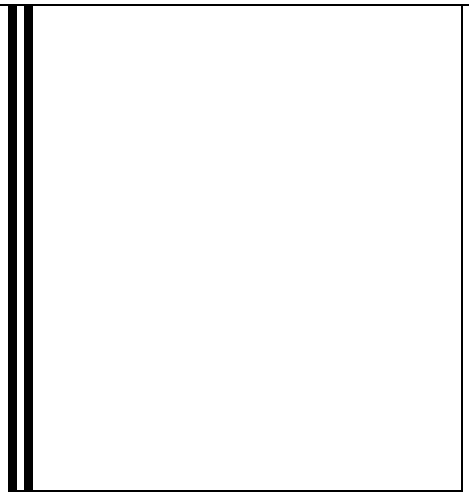
ules.





- Hide

  unwanted comments and messages from people.



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

Response: **Better tools for response to images used to discriminate and abuse**

The abuse experienced by the players we represent often has a significant visual element. Images and videos with discriminatory and harmful connotations feature on social media and can gain traction quickly among social media users. When these image files are copied it allows them to gain further traction. However, there is currently no method of monitoring for abusive imagery, monitoring through APIs can only allow for monitoring of written abuse. We believe that players/users need the ability to monitor for abusive images more easily.

We would like platforms to create the ability to block copies of particular images to help prevent further spread of the content. The Bukayo Saka monkey meme is a good case in point. If this type of tool was in place when the meme started to circulate it would have enabled the image to be reported to the relevant social media company for a blocklist. Our vision is that this type of tool would prevent the republishing/re-circulation of the image and its copies in the future.

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

Response: Image filters (Instagram) and warnings that some content may be abusive (X) are deployed. However, for adults there is very little limit on the type of content that can be exposed in feeds or search. A good example of this is the switch in policy re adult content on X. A trending topic like “Cheat” relating to a sports allegation of cheating is often littered with adult content related to “cheat” partners.

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

Response: As above, limiting or moderating comments on your posts or page, limiting people tagging you in posts/photos or blocking users you find objectionable. Restrictions can limit to followers or friend accounts only or just to verified accounts.

One area where we think there needs to be new features to help users be able to control their interactions with other users is the current flaw by which an individual can be added to Lists or tagged on platforms like X without a) being asked permission and b) without receiving a notification that they have been added/tagged. These lists often contain offensive or discriminatory epithets (e.g. Jewish Baby Killers). The only way the targeted user can find out if they have been added to such a list is to look and then try to remove themself. This is both traumatizing for the individual and highly time consuming. Platforms need to make this easier to manage. This could be solved by a) users receiving notifications when added to lists giving them an option to remove themselves b) a function monitoring the names of lists c) having an opt in option for 'listing' or @'ing people – so that only people that make a request to you can tag or add you to a list.

Message Request Abuse Filtering – We are aware of a number of circumstances where users unknown to the victim have used the medium of message requests to deliver abusive content. We think that it should be possible to have an automatic filter that picks up hateful and abusive content from individuals with which an account holder does not have an existing relationship and ensure that message requests aren't a 'soft underbelly' for accessing targets of abuse.

(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:

Our players fall into the latter category ‘users likely to be affected by encountering relevant content. We think that the following proactive tools/features which are either unavailable or only available on limited platforms could help them keep themselves and other users better protected whilst online.

Repeat Offenders – Clearer sanctions

There have been multiple occasions where we have identified individuals who have posted abuse towards players several times across the Season, or in high volumes within a short period of time. As we stated earlier platforms should make it much easier for users that experience abuse to link these complaints within their reporting systems. But platforms/companies should have in place:

- a threshold for how many times a user can post abusively in succession under a single post or profile (see Spamming below)

- a threshold for how many times they can be found posting abusive content overall . This could either be through how many times there have been confirmed reports for harmful content, content that has been blocked for being uploaded or based on how many nudges the account has received *and* ignored (see below).

Spamming/Abuse Campaign
In some cases, we see users spam-post abusive content online towards players during moments of anger and passion. Limiting how often a user can comment/ respond to a certain piece of content or profile within a certain time may help reduce the volume of explosive, spam-like messages. Again in this case we should ensure that the reporting tools better capture this type of repetitive abuse and that there is clear guidance of the sanctions that this type of behaviour might typically attract.

Behavioural Nudges
This is currently active on Meta but want this approach applied to all social media. If a user ignores a certain number of the nudge related to posting online abuse, the account and user should receive a penalty for ignoring it such as two-week suspension, no accounts can be created under any information related to their banned account e.g., phone number, and made to complete a mini education pack and test to get access to their account. The effort of going through the training should deter people from posting abuse, as such losing access to any of their accounts would.

'Opt in to abuse' Policy
Every social media company should automatically set individuals accounts to filter out abuse that can be received through comments or messages. This will help protect individuals from receiving abusive content. The only way they can see the abuse is to go into settings and manually opt-in to seeing this. This acts as a protective measure for all users, rather than a responsive measure from social media companies. While it appears X already has this in place to some extent, their thresholds for that is harmful/ sensitive content is much higher and allows all kinds of abuse to remain available.

Hidden words on Meta platforms is a good example. It can help protect those with protected characteristics by limiting abusive language in comments. However, many users object to the fact that they are being left in charge of having to implement this service rather than the platforms tackling the issue outside of providing individual tools.

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: **'Know your customer' policy.**

We think about this as being fundamentally about platforms knowing who they serve. Our first argument is that identification verification methods are beneficial because they allow platform provides the ability to truly hold registered users accountable for the activity that takes place on a particular account. In the case of our described 'behavioural nudge' tool (see empowerment section) In the instance a user wants to post abusive content, they would also have to identify themselves by filling out an automated form which pops up that covers name and contact details. Contact details would need to be verified to reduce the possibility of fake information being used. We believe the act of logging their details could as a secondary 'nudge' as it will indicate the platforms interest in them and a sense that they are being monitored. The information has a second purpose in that is will allow the Platforms to contact perpetrators and apply sanctions for posting specific harmful pieces of abuse. We believe that this type of intervention will hold users accountable and hopefully deter them from following through with posting.

(b) what documentation you understand to be necessary for different types, or levels, of identity verification on user-to-user services

Response: We would like to reflect on this and give you our more considered view further down the line

(c) whether you believe there are there any other circumstances where identity verification should be offered on user-to-user services.

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

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 ongoing 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-touser services? In your response to this request, please provide information relating to (a) – (d) where relevant.
Response: Most users want platforms to know who is posting so that action can be taken if they are threatened, abused or scammed
(a) whether they value verification being offered on a service

Response: There is also a decreasing value placed on public verification as it is perceived to have lost its authenticity. (see below)

(b) whether verification influences user behaviour, such as whether they perceive identity verification to signify authenticity

Response: This used to be the case. However, since X and other platforms reformed what it means to be verified this perceived authenticity has been eroded. In recent months it has become true that some of the most abusive accounts, as well as prolific spam accounts have used verification services as a shield from account action.

(c) attitudes towards non-verified, anonymous or pseudonymous users and the willingness to engage with them

Response: There is no clear evidence of users engaging less with non-verified users. If anything platforms like X see behaviours where non-verification is used to deride and chastise users. I.e. – if you are not verified you will often be attacked as being a "bot" or "fake".

(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: The methodology of any identity verification would make a significant impact on who could use it and who would be 'vulnerable' to exclusion and other impacts. This includes those less likely to hold official identification. It would be weighted against younger and minority demographics. There is also vulnerability in terms of those who may feel forced into identification in order to utilise a platform, but feel they are exposing themselves to risk when doing so. For example, people using platforms under pseudonyms who may discuss issues such as their sexuality when choice or cultural issues have kept them from doing so under their full identity. Even if identity verification would not be exposed to wider users on the platforms, the fear of that potentially happening could cause users not to verify and therefore be excluded from the platform.

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

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

For providers of online services

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

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

Question 41: Please provide information regarding the proportion of successfully identified fraudulent advertisements that are identified via:
(a) automated systems
Response:

(b) human processes
Response:
(c) user reports
Response:
(d) other (please provide further detail).
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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

Question 43: Proactive technology

For all respondents

Question 43: Please provide any evidence you have regarding proactive technologies which could be used to identify fraudulent advertising activity. In particular, we are interested in information related to the following points:

(a) The kinds of proactive technology which are/could be applied to identify or prevent fraudulent advertising
Response:
(b) A brief description of how these technologies are/could be integrated into the service
Response:
(c) The effectiveness, accuracy and lack of bias of such technology (including compared to alternative proactive and non-proactive methods) in relation to detecting fraudulent advertising and accounts which post fraudulent advertising material
Response:

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

Question 44: Advertising onboarding and verification

For all respondents

Question 44: Please provide any evidence you have regarding the processes for advertiser onboarding and verification related to protections against fraudulent advertising. In your response, please indicate whether these processes are currently implemented in respect of services which are in scope of the Act or whether they stem from another sector In particular, we are interested in information related to the following points:
(a) The criteria which advertisers are verified against, including documentation/evidence used to support verification, and what advertisers are required to declare
Response:
(b) The role of (a) automated processing and (b) human processing in the verification process, and how they interact
Response:
(c) The costs associated with advertiser verification and how those costs vary as scale increases
Response:
(d) The percentage of advertiser accounts that are verified
Response:

e) Whether advertisers are permitted to publish advertisements on the service while the verification process is ongoing
Response:
f) Whether there are additional/specific verification checks for advertisers placing adverts of certain kinds or targeting certain audiences, such as about specific products or services, or targeting users under the age of 18
Response:
g) Whether the verification of an advertiser account expires after a certain amount of time or certain activity, such as when advertisers make changes to their account or profile
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

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

For all respondents

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

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

Question 46: Advertiser appeals of verification/review decisions

For all respondents

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

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

Question 47: User reporting mechanisms

For all respondents

Question 47: Please provide any evidence you have regarding user reporting mechanisms for fraudulent advertising on services in scope of the Act. In particular, we are interested in information related to the following points:
(a) What user reporting tools there are for paid-for advertisements, and how these tools differ from those for user-generated content and/or search results and other search functionalities that are not paid-for advertising
Response:
(b) What percentage of user reports of advertisements relate to suspected fraudulent content, and the processes for taking action in relation to such reports
Response:
(c) Any statistics you can share on (i) the number of user reports of suspected fraudulent advertising received and resolved over a specific period and (b) the number of initial decisions appealed by users who made the report
Response:
(d) The criteria used to classify and prioritise user reports
Response:
(e) The median and/or average time it takes to respond to a user report, and any measures that are in place to ensure timely and accurate responses to user reports
Response:
(f) Any measures taken to make user reporting tools accessible, easy to use and easy to find for users
Response:
(g) How transparency and communication is maintained with users who have submitted reports
Response:
Is this response confidential? (if yes, please specify which part(s) are confidential)
Response:

Question 48: Use/involvement of third parties

For all respondents

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

Question 49: Generative AI and deepfakes

For all respondents

Question 49: Please provide any evidence you have regarding the impact of generative AI developments and deepfakes on the incidence and detection of fraudulent advertisements on services in scope of the Act. In particular, we are interested in information related to the following points:
(a) The frequency of deepfake fraudulent advertisements' occurrence, in absolute terms and/or as a proportion of all fraudulent advertisements, and how you expect this to evolve in the future
Response: We would like to revisit the issue of deepfakes with you in future
(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 thirdparty, 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: