25 March 2023

Ofcom Call for Evidence: Second Phase of Online Safety Regulation

Thank you for the opportunity to respond to the Ofcom Call for Evidence on the first phase of online safety regulation. In our response we have provided:

1. Background information about the Internet Commission. (Q1)

2. Answers to specific questions where we think we can contribute a helpful perspective

<u>1. Background to the Internet Commission (Q1):</u>

The Internet Commission is a non-profit organisation which promotes ethical business practice to counter online harms whilst protecting privacy and freedom of expression and increase platform accountability.

The Internet Commission was conceived by Dr Ioanna Noula and Jonny Shipp in 2017 in the context of their research for the Department of Media and Communications at the London School of Economics where they were both visiting fellows. The drivers at the time for such research were various events from Cambridge Analytica and Facebook's interference in the US election to Molly Russell's suicide following her exposure to harmful content on Instagram.

Ioanna and Jonny gathered multiple stakeholders such as senior academics from LSE, UCL and Imperial College, government representatives (UK Government Digital Service, Future Cities Catapult) as well as business representatives like Siemens, Telefonica and Pearson Education. The aim was to discuss the impact of social media platforms' failure to self-regulate and the need for the development of checks and balances that would increase the accountability of digital service providers, safeguard citizens' rights and wellbeing online, and restore stakeholder trust in tech.

On the back of this, in 2018, the Internet Commission was founded, and started a round of digital responsibility assessments with prominent businesses which led to their first public accountability report in 2021. The Internet Commission offers:

- independent evaluation of online intermediaries (social media, news sites, dating service providers, gaming service providers, digital education providers etc.) regarding their practices of content moderation;

- knowledge exchange where companies can discuss challenges and solutions related to tackling online harms; and
- a bank of good practices and reporting on the state-of-the art regarding governance and procedures of moderation of user-generated content (UGC) online.

Our comments to this consultation come from our experience from evaluating global online service providers' platforms across different online services and consider the insight the Internet Commission has generated by taking a closer look at procedures, resources and governance driving UGC moderation. Our research has explored critical challenges faced by service providers such as:

- achieving maximum efficiency by balancing human and automated moderation;
- understanding the implications of outsourcing content moderation services;
- addressing tensions emerging from users' rights online (digital rights); and
- ensuring content moderators' wellbeing.

Specifically, we share evidence from our evaluation of a diverse cohort of online services including two dating service providers, a gaming service provider, a live-streaming gaming service provider, a news services organisation, and a children's social media service provider. We retain a focus on procedural accountability; that consumer outcomes, particularly vulnerable communities, are best served by ensuring that processes and procedures are evaluated, and we use this information to identify emerging trends and issues. Being proactive in this fast-moving space is key and our approach allows us to flex against market requirements.

Our independent evaluation takes a look "under the hood" at processes, culture and technology that shape content moderation and offer industry benchmarks UK wide and internationally.

The Internet Commission was acquired by Ombudsman Services in 2022. Ombudsman Services is a not-for-profit private limited company established in 2002 which runs a range of discrete national Alternative Dispute Resolution (ADR) schemes across different sectors, including the sole ADR scheme in the energy sector, the Ofgem-approved Energy Ombudsman and the Communications Ombudsman, approved by Ofcom. Ombudsman Services is transitioning to a group structure under a parent company called 'Trust Alliance Group'.

2. Answers to specific questions.

Q3. What information do services have about the age of users on different platforms (including children)?

One partner organisation we worked with, explained to us that compliance with the ICO Age Appropriate Design Code (AADC) was the driver for recent compliance efforts to prevent under age users from accessing a service it operates that has an age restriction of 13+. The organisation is introducing an age verification process in Ireland and the UK. Parents/guardians of child users have the option to create child accounts for their children. When creating a child account, a parent/guardian must use a credit card to confirm that they are 18 or older.

Despite the requirements of the AADC, we do know that some organisations still do not conduct any age assurance at the point of access or thereafter. This means that they or third parties with whom they may share users' data could infer the age of users, including children in order to target advertising.

In our experience, it is often those services not relying on advertising which will actively implement age gates. The most common age gates are those which segment potential users into above-age, permissible users and underage, not-yet users. These are typically implemented at the point of registration.

Some organisations prevent users from re-submitting information once they find themselves ineligible for an account. They may block a child's credentials until they turn 18, based on the first date of birth entered. We have also seen the deployment of automated detection of underage users through analysis of photographs and self-descriptive free text associated with a user profile. Once a user is suspected of being underage, their account is suspended and can only be reinstated once an age verification process has been completed.

On detection of an underage user, some organisations alert the user and their parent/guardian, who is subsequently instructed to create a parental account which has more direct oversight over the child.

We do not have specific experience with organisations that segment user experiences on such a granular basis as ages 13-15 vs 15-17.

Q4. How can services ensure that children cannot access a service, or a part of it?

Age verification requires a fine balance of measures and costs for companies to ensure an inclusive yet safe process for online users. For services accessed by all, it is difficult to know which age groups will access the content so tailoring to age-appropriate needs is more difficult than if age verification were in place.

We do not have specific experience with organisations that segment user experiences on such a granular basis as ages 13-15 or 15-17. We have worked with an organisation that offers an online experience that is inclusive for all ages. Because of this, certain content may not be suited for

children as it caters to a general audience as opposed to acknowledging the specific risks to children online through this content being for all ages.
This leads to particular moderation and age gating challenges. Another challenge faced by this organisation is that some of their services are designed for multiple people, potentially of different ages to access at the same time.

Q6. Can you provide any evidence relating to the presence of content that is harmful to children on user-to-user and search services?

One organisation (with a strong child user base) did a pre-emptive risk evaluation of the features used by minors on their platform. After conducting this evaluation, they for instance determined that implementing a user-to-user chat function would not be worth the risk it would pose to children on the service. This shows how safety by design can pre-empt issues that are most high-risk for minors such as grooming or abusive language.

Q10. What are the governance, accountability and decision-making structures for child user and platform safety?

Central to those best practices we have identified in the course of our assessments has been integration: integration of external expertise, of user feedback and of internal stakeholder input.

External expertise, user feedback and stakeholder input are brought together in the policy development so that a range of professional and community-level perspectives is taken into account to better anticipate and mitigate potential issues prior to implementation.

A more holistic and encompassing approach to development enhances scalability by minimising friction. As the policy rolls out and formalises lines of communication and feedback mechanisms, it improves response times in relation to emerging issues.

A number of services that have participated in our assessments have emphasised the importance of de-siloing compliance and policy efforts across their business and leveraging the expertise of a range of internal stakeholders. They have reported more effective processes, a more collaborative culture internally and the successful translation of intelligence (regarding, for example, moderation data) and technology.

In our experience, the extent to which organisations communicate with users through structured processes and feedback mechanisms can indicate their maturity with respect to digital responsibility. Organisations that engage users when crafting policy and guidelines, and those that enable wider support communities, build trust and confidence by giving a meaningful voice to users. Practices we observed include user surveys, focus groups, and forums. The most mature practices involve engaging users in high-level policy and product decision-making.

When addressing complex topics such as freedom of expression, inclusivity and mental health harms, we often see third parties being involved to help organisations understand and prevent risks and harms. We also noticed that less mature organisations do not formalise governance processes and instead implement policy without thorough testing or consulting with external experts.

Q11. What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements for children (including children of different ages?

It is important that privacy and safety information is made available to users in age-appropriate formats. Clarity and accessibility of terms of service and public policy statements are two ways in which this can be achieved.

The way a service interface is designed must incorporate the opinion of the trust and safety team such that relevant updates to terms of service agreements and public policy statements are notified to child users. One organisation makes safety by design an integrated part of its service and policy development pathways. It plays a major role in the organisation's preventive, risk-based approach. This is achieved by:

- embedding multi-stakeholder reviews into the product development process,
- ensuring project managers receive an online safety education course,
- conducting mandatory reviews of product specifications focussed on risk,
- setting positive expectations for user safety, and
- aligning on safety priorities across functional teams throughout the organisation.

Further, the format of the communication must be simplified, for instance by including a summary of policy statement or the terms of service. For instance, when considering the format in which children are best able to digest relevant changes, an audio-visual format may be preferred if that is the key means of sharing content on said service.

Another organisation provides key privacy information to children in a form and language that children can engage with and understand. This takes the shape of summarised privacy information for young players which explains in plain language how the organisation collects and processes their personal information, and their data protection rights. Additional information is also provided regarding how children's data is processed on the organisation's website through relevant notices.

One relevant point of reference in relation to adjusting the statements for children of different ages is the Californian Age Appropriate Design Code. The legislation will compel online platforms to proactively assess the privacy and protection of children in the design of any digital product or service that they offer.

The age bands in the Code will inform the way in which services approach the task of “provid[ing] any privacy information, terms of service, policies, and community standards concisely, prominently, and using clear language suited to the age of children likely to access that online service, product, or feature”. The UK AADC has a similar approach.

We have seen examples of this approach among participants in our assessments. One organisation offers differentiated privacy notices for users creating and/or using a child account. Such notices are further split into those for younger children and older children. The notice for younger children uses plainer language and uses visuals to both further convey meaning and to keep younger users’ attention. The Rules or Terms of Service of the platform are similarly differentiated across age groups.

Q13. What can providers of online services do to enhance children’s accessibility and awareness of reporting and complaints mechanisms?

Organisations have developed and strengthened their appeals process. One organisation we worked with has updated its appeals process significantly by launching an online appeals system. Appeals are reviewed by a senior member of staff who was not involved in the original decision and users are kept informed about the decision by email.

Q14. Can you provide any evidence or information about the best practices for accurate reporting and/or complaints mechanisms in place for legal content that is harmful to children, or users who post this content, and how these processes are designed and maintained?

The overarching best practice regarding accurate reporting and complaints mechanisms is a willingness by organisations to adjust their products and processes in the light of shortcomings or after improved industry practices come to light. With several organisations undergoing two separate Internet Commission accountability reports, we were able to identify improvements after organisations noted our feedback and reassessed their approaches accordingly.

Examples of such improvements include the introduction of an online user appeals process and the establishment of a centralised team to ensure that trust and safety is factored into product development across the organisation as part of a revised technology governance structure.

Q15. What actions do or should services take in response to reports or complaints about online content harmful to children (including complaints from children)?

Services must, as a minimum, ensure their reporting processes are legally compliant in relation to upcoming UK Online Safety Bill and EU Digital Services Act online safety rules.

However, identifying leading industry practices is also key to effective reports and complaint processes for children. Based on our expertise, the more informal the process is, the more difficult it may be to identify – especially for a child user – so it is crucial that trust and safety teams work in tandem with those responsible for interface design. It is important for these processes to be evaluated in comparison with other emerging industry practices as well as via feedback of users of the service.

Q16. What functionalities or features currently exist that are designed to prevent or mitigate the risk or impact of content that is harmful to children?

The Internet Commission's second Accountability Report evaluates many practices designed to prevent or mitigate the risk of harmful content to children. It was clear that certain practices might be considered as erring on the side of caution. For instance, one partner organisation with a strong child user base uses a system of automated removal, whereby potentially harmful content is immediately suppressed. We noted that platforms which are asking their users to review their potentially "harmful or offensive" content before sending, helps them learn from their negative behaviour patterns. Although erring on the side of caution with what it removes from its platform, the organisation believes that combined with the promotion of space for positive engagement is a net gain for its child user base. Further, pre-emptive measures are key for providers with a large child user audience. This is achieved by 'designing out' certain features such as private messaging and thereby pre-empting the risks of grooming or obscenity.

Finally, certain types of providers are not focused on a younger audience and may, by their nature, place younger users in a vulnerable environment, so age verification plays a crucial role. One of our partner organisations opted for an automated end-to-end age verification process. As an exclusively 18+ service, age-gating measures are built into the registration process. Any user entering an underage date of birth sees their credentials blocked until they turn 18 according to the date of birth entered. Machine Learning (ML) based tools are used to detect underage users through their photographs, biographies and private messages. Once a user is suspected of being underage, their account is suspended and can only be reinstated once an age verification process has been completed.

Q17. To what extent does or can a service adopt functionalities or features, designed to mitigate the risk or impact of content that is harmful to children on that service?

Please see question 16.

Q22. How are human moderators used to identify and assess content that is harmful to children?

Human moderation is a key part of content moderation. Within human moderation, there are several styles of moderation. Firstly, one organisation with a strong child user base uses limited community moderation. This enables creators to facilitate online sub-cultures adapted to their streams. Creators can appoint trusted users to act as channel moderators who then set the level at which automated moderation tools filter content and can blacklist specific terms. Since most of the content on the service is public (‘one-to-many’), this layered approach of internal and community enforcement must operate coherently. The organisation has sought to empower community moderators whilst keeping devolved moderators in line with the organisation’s broader standards for content moderation.

Another organisation curates its content in an editorial manner to promote public debate around current affairs. The organisation proactively identifies areas of its website that would benefit from and support online user interaction. Once an area is open for user comments, the internal moderation management team invites and curates contributions. They keep user engagement going by opening and closing content threads. Moderation is done in such a way to encourage users to participate and comments are treated with care and respect. Decisions err on the user’s side, focusing first on users’ intentions when reaching a judgement about the suitability of their posts. They have promoted a culture of openness and respect that encourages public debate and upholds the value of audience input into key issues.

Finally, organisations that are impacted by safety concerns brought about by user fraud or identity theft use human moderation to have some oversight of content flagged by automated tools. The organisation uses ML-based tools to scan private messages and detect and flag anything potentially harmful or inappropriate. In the organisation’s specific context, individual preferences and other factors can affect how a comment is intended or received in a way that ML cannot always detect. The tool was therefore designed to first ask the recipient if they perceive a flagged message to be harassment and if so, direct the user to report it. This means that user feedback is also incorporated into the moderation process and used to further refine the tools. This has generated significant insight for the organisation about what may constitute offensive or harmful content and has assisted the development of new prompts; for instance, a feature which prompts users to consider before they send a message whether it might be perceived as harassing.

Q23. What training and support is or should be provided to moderators?

From our experience, the best practice that we have seen for the training and support of moderators includes:

- Knowledge of the organisational values and policies,
- How to use moderation tools;
- Assessments to check and update moderator understanding with access to retraining where appropriate,
- A combination of theory and classroom work, with supported hands-on experience, and
- Performance reviews and conversations.

We have seen organization adopt a wellness programme which gives moderators access to a range of support tools such as counselling, wellness apps and Cognitive Behavioural Therapy (CBT) The same organisation also rotates the channels on which the moderator works, so they are not continually exposed to the most harmful content.

Q27. Where children attempt to circumvent mitigations in place on a service, what further systems and processes can a service put in place to protect children?

Two partner organisations emphasised the importance of the front-facing image of the platform in order to minimise the interactions which children have with the service itself. There is a greater risk to child users, if a service is presented in a manner which is appealing to children and an ineffective age verification tool is in place, the risks presented to child users are far greater. Thus, the remedy to such a risk is marketing or advertising the service in such a way that does not appeal to children. Within the service itself, one partner organisation tailors its advertisements via user segmentation which may have an indirectly positive impact in ensuring underage users do not see unsuitable or harmful advertising.

Please do not hesitate to contact us if you would like further information regarding our response. Our response is not confidential.