Emailed to: OS-CFE@ofcom.org.uk

10 October 2022

Dear Sir/Madam,

Ofcom Call for Evidence: First 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 Ombudsman Services;

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

1. Background to Ombudsman Services:

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. We are also one of two ADR schemes in the Communications sector approved by Ofcom and we run an appeals service for private parking.

We operate at a critical juncture between suppliers, consumers and the Government to resolve disputes. Each scheme is funded by the members and our service is free to consumers. We share data and insights to support suppliers to deliver better innovation and positive outcomes for consumers. This practice enables us to drive up standards in the industry by encouraging collaborative approaches to making improvements, managing expectations and informing policy.

Tasked with serving some of the UK's most disruptive consumer sectors, we work hard to proactively respond to the continuously evolving requirements of sectors, including our people expanding their technical expertise and our systems expanding their technological functionalities to reflect the changing way consumers are interacting with suppliers, their services and products.

We are a purpose-driven organisation which exists to **build, maintain and restore trust** and confidence between consumers and businesses and that technology, data and insight are key enablers in the delivery and efficacy of service. We have used data to deliver insights that support suppliers in changing their approach to customer service. In the dispute resolution area, we promote and operate what we call strategic redress. By working with all parties – including businesses and regulators – we can perform an important role to support the improvement of consumer outcomes at a macro level. Integrating redress with processes upstream (such as the way businesses deal with complaints) helps to create the right cultures and practices in businesses and helps to foster trust and confidence in the broader market.

As we have acknowledged the need to evolve, we are also moving to a new group structure under a parent company called “Trust Alliance Group”. Our Group purpose strives to:

- help businesses improve service, change culture and build confidence;
- help industries resolve systemic industry-wide issues; and
- harness technology and data analysis to improve services and reduce detriment.

Within our Group we have created a software development company called Lumin Tech Limited with a mission to empower businesses to be more consumer-centric via the development of technological solutions for dispute resolution, including Case Management System (CMS) platforms. We have also delivered a data modernisation programme designed to unlock the full potential of our data and insights for the betterment of sectors, suppliers and consumers.

We recently acquired the Internet Commission, a non-profit organisation which promotes ethical business practice to counter online harms and increase platform accountability. Our response to this consultation is underpinned and comes from the knowledge and expertise from this part of our wider organisation. We refer to platforms as service providers throughout this response as we think it places the consumer at the centre of the conversation and highlights the function of that service.

2. Below we have made some further comments on specific consultation questions:

Q1. Please provide a description introducing your organisation, service or interest in Online Safety. For providers of online services, please provide information about:
• The type of service and functionalities you provide;

The Internet Commission, recently acquired by Ombudsman Services, is a non-profit organisation which promotes ethical business practice to counter hate speech, abuse and misinformation online, whilst protecting privacy and freedom of expression.

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.

Q2. Can you provide any evidence relating to the presence or quantity of illegal content on user-to-user and search services? We are particularly interested in evidence about how this might vary across different services or types of service, or across services with particular users, features or functionalities.

Q5. 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.

Q7. What can providers of online services do to enhance the transparency, accessibility, ease of use and users' awareness of their reporting and complaints mechanisms? Please submit evidence about what features make user reporting and complaints systems effective, considering:
• Reporting or complaints routes for registered users;
• Reporting or complaints routes for non-registered users; and
• Reporting routes for children and adults.

As highlighted above, we work with a range of service providers - including dating service provider, gaming service provider, news services, and children's social media service providers. We have seen service providers implement different ways in which they enhance the transparency, accessibility and awareness of reporting and complaint mechanisms. These include:

- Ensuring there is a formal right of appeal process and that it is clear to users.
- Sharing details of content which has been identified as inappropriate or harmful and information on the appeals process. This approach aims to treat users as trustworthy contributors, with a focus first on users' intentions when reaching a judgement about the suitability of their posts.
- Apology mechanisms that are followed for users which have been found via the appeals process to have been wrongfully banned. This can encourage a shared sense of accountability.
- Progress updates on appeals and in the case of one organisation, a forthcoming dashboard for appeals.
- Integrating enforcement and appeals system. In instances where this doesn't happen, users cannot connect an appeal with a specific enforcement action and moderation staff must spend time checking across the two systems to validate the appeal. This may mean questionable – or simply incorrect – moderation decisions simply go unchallenged

The reporting routes for children and adults are not clear at the moment in the sector but some providers are looking at simplifying their appeals process to make it more accessible to vulnerable groups and we believe this is an important step. We are keeping this area under review.

Q11. Could improvements be made to content moderation to deliver greater protection for users, without unduly restricting user activity? If so, what? Please provide relevant evidence explaining your response to this question. Please consider improvements in terms of user safety and user rights, as well as any relevant considerations around potential costs or cost drivers.

In our experience, improvements to content moderation could be made by considering:

Moderator training and support

The moderation process should be respectful to users: when a post is removed, both the user that created the post and the "flagger" of the problem are notified, with details of which content was removed, the rule broken and information about the appeals process. This could follow a well-developed process for broadcast television and radio, which includes a clear escalation path which dovetails with the established complaints process.

Quality Assurance

Appeals processes help get the balance right between safety and freedom of expression. Moderators and automated processes can remove too much or too little content. Holding regular quality assurance sessions where a sample of decisions can be checked, and feedback could be provided particularly on contentious issues should be part of a running dialogue in the organisation. Quality assurance checking should ensure consistency across moderators at different periods of time. The number of appeals should be tracked and evaluated by specialist quality assurance teams.

Integrated enforcement and appeals systems

Users need to be able to understand what activity causes a particular enforcement action to understand where they went wrong and be able to appeal if necessary. This also has impacts for moderation staff who must spend time checking across the two systems to validate the appeal. A disconnected approach may lead to questionable – or simply incorrect – moderation decisions. Moreover, educating the user through more transparency could minimise the impact of online activity that requires further sanctions.

Signposting mental health support

We are aware of a service provider who partnered with a mental health service. Users may also text the name of the organisation to the mental health service provider to be connected with a counsellor immediately. It is also beneficial to consider mental health support for content moderators.

Q14. How are sanctions or restrictions around access (including to both the service and to particular content) applied by providers of online services? Please provide evidence around the application and accuracy of sanctions/restrictions, and safeguards you consider should be in place to protect users' privacy and prevent unwarranted sanction.

Service providers needs to ensure there is a balance between technology and human intervention in the building of their systems and processes, so it's important to have:

- moderation technology that detects and responds to new kinds of harm. ; and
- humans engaged in moderating content. This is important when deploying automated tools, to ensure fairness and avoid biases online.

Q22. What age assurance and age verification technologies are available to platforms, and what is the impact and cost of using them? In particular, please provide evidence explaining:
• how these technologies can be assessed for effectiveness or impact on users' safety;
• how accurate these tools are in verifying the age of users, and effective in preventing children from accessing harmful content;
• steps that can be taken to mitigate any risk of bias or exclusion that may result from age assurance and age verification tools;
• the costs involved in implementing such technologies; and
• the safeguards necessary to ensure users' privacy and access to information is protected, and over restriction is avoided.

Age-gating checkpoints

Establishing and maintaining an age-appropriate online environment is another tough challenge. Organisations that do not restrict access to their services by age focus on careful moderation and reporting mechanisms to create an environment suitable for all users. A wide age range presents the problem of how to best set expectations of acceptable content, conduct and contact. Conversely, organisations that offer services that are only suitable for adults must find robust ways to protect children by preventing them from gaining access. One dating service provider operates a strictly 18+ service with an age-gate at the point of registration. If a user inputs a date of birth which puts them under 18 years of age, their credentials are locked until they reach 18 according to the date of birth originally inputted. Beyond the point of registration, it uses multiple checkpoints to detect and remove underage users. Any such user is automatically suspended pending age verification in which the user's identity document is verified. If verified, then reinstatement is immediate.

We know that there are currently no completely reliable tools to stop children accessing harmful content.

Q27. For purposes of transparency, what type of information is useful/not useful? Why? In particular, please consider:
• Any evidence of public information positively or negatively affecting online user safety or behaviours, how this information is used, and by whom;
• What information platforms should make available, considering frequency, format and intended audiences;
• What information Ofcom should make available through its transparency report, considering frequency, format, intended audiences and potential use cases by external stakeholders;
• The benefits and/or drawbacks of standardised information and metrics; and
• Any negative impacts or potential unintended consequences of publishing certain types of information, and how these may be mitigated.

We think that the key information includes:

- An emphasis on procedure, or procedural accountability. Rather than companies addressing specific harm, there should be an emphasis on robust governance that prevents regulation by outrage.
- A commitment to a cycle of independent review and evaluation and publication of those results. By committing to review and evaluation, particularly of procedures and the publication of those results, companies demonstrate a commitment to transparency and their duty of care by sharing best practice towards advancing the industry standard.
- Clear communication of rules in user agreements, with regards to how it processes appeals.
- Quality assurance to ensure the consistent application of enforcement.

Accountability can be achieved by going beyond the current industry standard in Transparency Reporting as seen in the Transparency Reports published by Very Large Online Platforms (VLOPs) such as Twitter, Facebook of Google. The content of the reports and the way information is communicated (i.e. structure, format, ease of access) should increase the knowledge of the public and generate insight that can be

used by stakeholders and contribute to the development of pertinent regulation, the minimisation of harms and risks.

Additionally, service providers should be prepared to clarify or respond to further requests through dedicated transparency teams services qualified to provide further orientation and insight. This approach to Transparency Reporting will further empower regulators and create consistency across service providers regarding the information provided and the resource required by regulators to undertake audits and analysis.

The Internet Commission's Accountability Reports are evidence that evaluation methodologies aimed at delivering procedural accountability are technically feasible and allow the documentation of good practices and challenges regarding UGC moderation. We would be happy to discuss our findings in this area in more detail.

Q28. Other than those in this document, are you aware of other measures available for mitigating risk and harm from illegal content? We would be interested in any evidence you can provide on their efficacy, in terms of reducing harm to users, cost and impact on user rights and user experience.

We have seen a couple of different measures used by service providers, which may help to mitigate risk and harm from illegal content:

These are:

- **Pre-filtering of all public-facing content**: One dating service provider mitigates the risk of harm by automatically pre-filtering all publicly available photos and user descriptions before being published, so inappropriate profile pictures and biographies are identified and removed and, in serious cases, malicious actors are detected and removed.

- **User empowerment through user-friendly feedback:** One of the dating service providers explains to the user where they have gone wrong. In offering to users the opportunity to become more informed about breaking the rules, it is offering a more equitable pathway to improving the safety of users, delegating agency to users, and cultivating a sense of responsibility that allows users to align their behaviour with the service providers' expectations before more severe enforcement action is required. The company has found a significant degree of success in this strategy, having seen very low rates of repeat offending and a significant reduction in bans. User empowerment has been at the heart of the recent language update in reporting flows. The articulation of possible violations in user- friendly language which maps "real-life" harms onto policy language is enhancing the role users play in improving the quality of the service by better capturing their negative experiences on- and off- platform.

Should you wish to find out more, you may find our reports helpful The Internet Commission – advancing digital responsibility through independent evaluation. (inetco.org) and please do not hesitate to contact us if you would like further information regarding our response. Our response is not confidential.