Ofcom Call for Evidence: Second Phase of Online Safety Regulation

Global Partners Digital Submission
March 2023

About Global Partners Digital

Global Partners Digital is a social purpose company dedicated to fostering a digital environment underpinned by human rights

Response

We welcome the opportunity to provide comments on Ofcom's call for evidence to inform its second phase of online safety regulation for protecting children from legal content that is harmful to them, through access assessment duties, risk assessment duties and draft codes of practice. Below we respond to the questions most closely linked with matters related to human and children's rights online.

Question 5: What age assurance and age verification or related technologies are currently available to platforms to protect children from harmful content, and what is the impact and cost of using them?

There are several age assurance and age verification technologies in varying stages of development and deployment by online platforms and third party age assurance vendors. These include verification of paper ID documents or of users' details against public records, social vouching, and biometric age estimation technologies, as well as estimation of users' ages based on behaviour or content shared.

While it is important to find ways of protecting children from content that is harmful to them online, age assurance and age verification technologies and methods may have potential adverse impacts on individuals' human rights, as we highlighted in our submission to the first call for evidence (see response to Q22).

We reiterate here our concerns over blanket mandating of age assurance and age verification tools across all providers to fulfil their child safety duties. If online service providers are still required to use age verification measures, there should be robust safeguards in place around the collection, processing and storage of personal data by these systems, as well as minimum standards for critical infrastructure to ensure that such systems do not make personal data vulnerable to misuse or cyber attack. Individuals should not be required to share their personal data in order to access parts of a site or application if they do not want to and wish

to remain anonymous – this is particularly important for vulnerable or persecuted communities seeking to enter safe spaces for expression and community online.

Finally, platforms should be transparent with users about where age assurance or age verification tools are being used and how their data is being processed through such systems, as well as providing users with the option to appeal any age determinations or inappropriate usage of their personal data.

Question 19: With reference to content that is harmful to children, how can a service mitigate any risks to children posed by the design of algorithms that support the function of the service (e.g. search engines, or social and content recommender systems)?

Online service providers that design content ranking or recommendation algorithms to maximise user engagement will inevitably produce systems which amplify and promote divisive, shocking, offensive or disturbing content. These content ranking and recommendation systems are based on maximising profit for the online service provider through advertising, rather than on maximisation of the health of the information environment provided by the online service provider.
A service should mitigate risks to children by content ranking and recommendation systems by:

- Designing algorithms to rank high-quality and verified content from reputable sources more highly in recommendation pathways;
- Not using children's personal data or interaction histories to target them with specific content types;
- Ensuring that, where a child has liked or interacted with a piece of content which has later been judged to be either illegal or to fall into one of the designated categories of harmful content for children, that such engagements with this content do not continue to serve as input on which the algorithm bases its future recommendations;
- Providing children with tools to easily reject or switch off content which is recommended to them which they do not like or wish to see;
- Providing children with easy access to reporting tools for harmful content that is prompted in a recommendation or ranking system;
- Designing services and spaces intended specifically for children with minimal collection of personal data and without targeted advertising.

Question 20: Could improvements be made to content moderation to deliver greater protection for children, without unduly restricting user activity? If so, what?

We provided a number of recommendations on improvements to content moderation processes without unduly restricting user activity in our response to Q11 of the first call for evidence, as well as listing a range of content moderation

tools and platform design features which may help platforms to deliver greater protections for users in our response to Q18.

With respect specifically to content moderation to deliver greater protection for children, we note that the UN Committee on the Rights of the Child has advised states to respect children's evolving capacities in relation to their engagement with digital technologies (General Comment No. 25 (2021) on children's rights in relation to the digital environment, CRC/C/GC/25, paragraphs 19-21). The Committee points out that "the risks and opportunities associated with children's engagement in the digital environment change depending on their age and stage of development", and recommends that "States parties should take into account the changing position of children and their agency in the modern world, children's competence and understanding, which develop unevenly across areas of skill and activity, and the diverse nature of the risks involved."

In line with this international guidance, and in light of the fact that might be harmful to one child may be a useful source of information for another, online service providers should be empowered to provide children with greater agency over the types and amounts of content that they see, and it should be left to children and parents to determine what types of content they are comfortable seeing and interacting with. Platforms should also empower community-led moderation in groups and forums, allowing users to set their own terms of engagement and to define what speech or content is prohibited.

Platforms should also ensure that children are clearly informed of how to opt out of seeing particular content types that they do not want to see, and of how to report harmful content easily to the platform. We provided additional guidance on accessible reporting routes for children in our response to Q10 of the first call for evidence. Platforms can also ensure that children who are searching for or consuming illegal or damaging content are recommended or re-directed towards alternative content, such as helplines or resources.

Question 21: What automated, or partially automated, moderation systems are currently available (or in development) for content that is harmful to children?

Primary priority content deemed harmful to children will be defined in secondary legislation, but is likely to include content which is pornographic or which encourages self-harm or eating disorders, as well as legal suicide content. These content types are generally image- and video-based, and are primarily detected with image classifiers developed through machine learning techniques. These tools have been known to erroneously flag non-harmful images and videos as harmful – for example, by flagging a mother breastfeeding as a pornographic image – and have also been shown to discriminate against women by disproportionately flagging women's images as pornographic compared to men's.

Priority content which is harmful to children will also be defined in secondary legislation, but is likely to include online abuse, cyberbullying and harassment,

harmful health content and content depicting or encouraging violence. These content types may be a mixture of text/speech and image/video content, and tend to require multimodal classifiers incorporating natural language processing technology for text and speech elements as well as image recognition tools, sometimes alongside metadata analysis. These tools have been shown to struggle with interpreting context, as well as to erroneously flag speech from marginalised groups or language varieties as abusive or explicit.

Non-designated content which is harmful to children may take any format, and therefore automated tools to address it face the same problems as those outlined above.

Crucially, for all of these content types, typically the content in question is not already known by the platform to be content which is harmful to children, rendering more accurate automated moderation systems – such as hashing of known illegal images – virtually impossible. Furthermore, the complexity of these content types and their context-sensitivity makes them harder to detect than, for example, spam or malware.

In light of the limited accuracy, known bias and contextual blindness of most of the automated tools mentioned above, we provided recommendations as to how such automated content moderation tools could be improved and governed in our response to Q11 in the first call (see page 8). We strongly recommend that where such automated content moderation tools are in use that they are accompanied by rigorous human oversight and are not authorised to remove content independently of review by a human.

We also strongly recommend that platforms are not required to apply such automated content moderation tools to content posted by users online in a way which would constitute a requirement for general and proactive monitoring. Such a requirement would contradict normative guidance from the UN Special Rapporteur on the promotion and protection of the right to freedom of expression; in his 2018 report to the Human Rights Council (A/HRC/38/35) he stated that "States and intergovernmental organisations should refrain from establishing laws or arrangements that would require the "proactive" monitoring or filtering of content, which is both inconsistent with the right to privacy and likely to amount to pre-publication censorship".

Question 22: How are human moderators used to identify and assess content that is harmful to children?

Human moderators are used by many platforms to identify and analyse content which is suspected to contravene the terms and conditions of the platform. In some cases, human moderators are not employees of the platforms but are participants of the online com-munity, as demonstrated by community-governed platforms like Wikipedia and Reddit.

For some platforms, the content types prohibited in the terms of service include content types which are likely to be designated as content which is harmful to children in second-ary legislation, such as pornographic content or cyberbullying. However, other platforms – such as adult-content sites or some gaming servers – have fewer restrictions on speech or content, meaning that human moderators may not be used at all at present to identify and assess content that is harmful to children, instead focusing only on content which is illegal or which constitutes incitement to violence.

Question 23: What training and support is or should be provided to moderators?

Moderators should be provided with training on how to interpret and consistently apply platform terms and conditions, as well as on any changes or updates to those terms and conditions on an ongoing basis. Moderators should also be trained on how to escalate contentious cases to more senior decisionmakers, and empowered to raise concerns about the application of particular aspects of the platform terms and conditions in practice based on their experience with managers.

We made a number of recommendations on improvements to human moderator systems in our response to Q11 of the first call for submissions (see p.6). Moderators should be provided with decent pay and support for psychological wellbeing such as therapy and counselling and regular breaks. They also should not be required to work towards unreasonable daily or hourly quotas so as not to force hasty decisions on more nuanced or difficult pieces of content. These principles should apply whether or not the moderator is employed in-house by the platform or by a third party service provider on behalf of the platform.

Content moderators should also be able to specialise and progress in expertise on a par-ticular content type, and should be assessed for psychological suitability for deployment on that content type prior to working on it.

Question 24: How do human moderators and automated systems work together, and what is their relative scale? How should services guard against automation bias?

Automated systems using hashing technology to detect known CSAM imagery are rela-tively reliable and can be deployed at scale. Such systems should still be regularly re-viewed and assessed for accuracy and impacts on users.

Automated systems using machine learning or statistical inference to classify or flag con-tent require a greater degree of human oversight. Affected users should always be in-formed when a decision that affects them is made by automated systems, and should al-ways be given the opportunity to request a human review of the decision. Platforms should collect and analyse data on the accuracy and consistency of any such automated systems that they deploy, taking into account the number of decisions made which were subsequently appealed and overturned and comparing the accuracy of decisions made for different content types and

formats.

Automation bias is a well-recognised phenomena. Services can guard against this by:

- not telling a moderator whether they are reviewing the decision of an automated tool or the decision of another moderator;
- providing training for moderators on how automation bias manifests and how they can be aware of it;
- including step-by-step questions or principles that a moderator can follow when reviewing an automated decision, such as "Is it likely that this is a positive result because of biased input data to the model?" or "Is this decision consistent with other decisions I have made on similar content types before?";
- increasing the time allocated to moderators to make decisions on reviews of au-tomated systems to encourage more critical thinking.