Ofcom Consultation: Protecting people from illegal harms online

Global Partners Digital submission
February 2024

Table of Contents:

About Global Partners Digital

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

We welcome the opportunity to continue contributing to Ofcom's consultations on the implementation of the Online Safety Act. Please note that responses to all questions are not confidential.

Overarching comments

- *The Internet is not an aseptic environment*

The first article of the Universal Declaration of Human Rights recognises that everyone is *"endowed with reason and conscience"*, a principle developed further in human rights law to include, among other things, the protection of opinion, expression, belief, and thought. Article 19(1) of the International Covenant on Civil and Political Rights (ICCPR), also echoing the Universal Declaration, provides that *"everyone shall have the right to hold opinions without interference"*. Opinion and expression are closely related to one another, as restrictions on the right to receive information and ideas may interfere with the ability to hold opinions, and interference with the ability to hold opinions necessarily restricts their expression. As highlighted by the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, *"[t]he ability to hold an opinion freely was seen to be a fundamental element of human dignity and democratic self-governance, a guarantee so critical that the Covenant would allow no interference, limitation or restriction. Consequently, the permissible limitations in Article 19 (3) expressly apply only to the right to freedom of expression in Article 19 (2). Interference with the right to hold opinions is, by contrast, per se in violation of Article 19 (1)"*.[1]

As has been long-established in UK jurisprudence, the right to hold opinions without interference and the right to freedom of expression are principles that characterise a 'democratic society'. In the case of Handyside v. United Kingdom,[2] the European Court of Human Rights ruled that *"[f]reedom of expression constitutes one of the essential foundations of such a society, one of the basic conditions for its progress and for the development of every man. Subject to [legitimate restrictions] it is applicable not only to "information" or "ideas" that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the State or any sector of the population. Such are the demands of*

[1] David Kaye. Report "The use of encryption and anonymity to exercise the rights to freedom of opinion and expression in the digital age". A/HRC/29/32. Paragraph 19. Available at: https://www.ohchr.org/en/documents/thematic-reports/ahrc2932-report-encryption-anonymity-and-human-rights-framework

[2] *Handyside v United Kingdom*. ECHR (1976) 5493/72 (7 December 1976). Available at: https://hudoc.echr.coe.int/eng?i=001-57499

that pluralism, tolerance and broadmindedness without which there is no "democratic society". This means, amongst other things, that every "formality", "condition", "restriction" or "penalty" imposed in this sphere must be proportionate to the legitimate aim pursued."[3]

The legal framework summarised above is highly relevant to the consideration of Ofcom's duties to oversee the implementation of the Online Safety Act (OSA). It is of the utmost importance to draw the line between illicit speech and offensive, uncomfortable, or even shocking speech or actions in the use of digital platforms where the interchange of ideas and expressions shapes today's public debate.

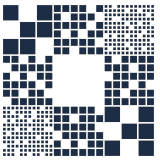
- *International Human Rights Framework on freedom of expression restrictions*

According to Article 19(3) of the International Covenant on Civil and Political Rights (ICCPR) harmful content that can be considered a legitimate restriction for freedom of expression must be "provided for by law"; be precise, public and transparent; and avoid providing state authorities with unbounded discretion to apply the limitation.[4] The OSA establishes legal obligations around illegal content categories and it has entrusted Ofcom with the duty to implement its provisions. However, in undertaking this task Ofcom should be careful to avoid unintentionally expanding their remit from assessing the risks of the spread of illegal content -which would be a part of the permissible restrictions of freedom of expression according to the ICCPR- to restricting conduct or activities that may or may not result in the exchange of illegal content, such a restriction would be incompatible with freedom of expression as it would amount to prior restraint of speech.

- *Stakeholder engagement*

Ofcom's obligation to oversee the implementation of the OSA will require the collection of a variety of input and perspectives, particularly on risk factors and the causes of online harm. As such, we would like to stress the need for Ofcom to engage with a broad range of stakeholders, providing a fair opportunity for the representation of their perspectives in a transparent way, and enabling stakeholders to track the impact of their contributions on Ofcom's thinking. This includes: sharing information on expert consultants that have been engaged, publishing submissions received (except for when the authors have requested confidentiality), and publishing consultation material in a language, format and extension that facilitates meaningful stakeholder engagement. Lengthy and difficult-to-digest materials pose a barrier to participation which can only be overcome by the best-resourced groups

[3] Ibid. at paragraph. 49.
[4] See Human Rights Committee, general comment No. 34 (2011). Paragraph 25. Available at: https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf



at the expense of the participation of others, including those who work on the ground, have specific thematic expertise or diverse perspectives that could be relevant to Ofcom's work.

We particularly encourage Ofcom to be transparent in how input received through consultation processes is weighted and how different stakeholders' positions continue to be integrated into the implementation of OSA.

Volume 2: The causes and impacts of online harms

1. End-to-end encryption and anonymity as "functionalities" posing particular risks (questions 1 and 2)

When exploring the causes and impacts of online harms, Volume 2 sets out analyses of different kinds of illegal harms and their associated risks on both U2U services and search services. End-to-end encryption (E2EE) is considered a risk factor that service providers must address in their risk assessments for twelve categories of illegal content in the OSA, including CSEA, hate crime, terrorism, drugs, immigration, sexual offences, extreme pornography, intimate image abuse, proceeds of crime, fraud, foreign interference and false communications.

Whilst acknowledging the role of encryption in safeguarding privacy, Volume 2 states that encryption *"stands out as posing a particular risk"* and that *"Offenders often use end-to-end encrypted services to evade detection. For example, end-to-end encryption can enable perpetrators to circulate* CSAM*, engage in fraud, and spread terrorist content with reduced risk of detection"*. Volume 2 also refers to *"some evidence"* that pseudonymity (where a person's identity is hidden from others through the use of aliases) and anonymity "*can embolden offenders to engage in a number of harmful behaviours with reduced fear of the consequences"*. There is recognition that *evidence* linking pseudonymity and anonymity to hate speech or cases of harassment and stalking is contested. At the same time, the guidance also references the importance of pseudonymity and anonymity for freedom of expression.

We welcome acknowledgement of the role of end-to-end encryption in safeguarding privacy but we respectfully disagree with the approach adopted by Ofcom of framing it as an enabler for a broad range of online harms. While encryption may facilitate or protect users from sharing illegal content without scrutiny, it remains an essential tool for protecting privacy and safety in the online environment.

The right to be able to communicate privately with others, free from interference, is a critical element of both the right to freedom of expression and the right to privacy. It is why, in the offline environment the Royal Mail does not read our letters, why telephone and mobile communication providers do not listen to our calls, and why our conversations are not surveilled by police officers. It is fundamentally important that in seeking to address illegal and harmful content online, our ability to communicate privately is not undermined. The UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression has detailed how encryption and other privacy-enhancing technologies provide the

security necessary for the exercise of the right to freedom of opinion and expression in the digital age:

> *"Encryption and anonymity, separately or together, create a zone of privacy to protect opinion and belief. For instance, they enable private communications and can shield an opinion from outside scrutiny, particularly important in hostile political, social, religious and legal environments. Where States impose unlawful censorship through filtering and other technologies, the use of encryption and anonymity may empower individuals to circumvent barriers and access information and ideas without the intrusion of authorities. Journalists, researchers, lawyers and civil society rely on encryption and anonymity to shield themselves (and their sources, clients and partners) from surveillance and harassment. The ability to search the web, develop ideas and communicate securely may be the only way in which many can explore basic aspects of identity, such as one's gender, religion, ethnicity, national origin or sexuality. Artists rely on encryption and anonymity to safeguard and protect their right to expression, especially in situations where it is not only the State creating limitations but also society that does not tolerate unconventional opinions or expression".*[5]

The identification of end-to-end encryption as an enabler of harm sits at the foundation of calls for weakening encryption. *As pointed out by the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression,* "[i]n the contemporary technological environment, intentionally compromising encryption, even for arguably legitimate purposes, weakens everyone's security online. [6]

Moreover, the OSA only refers to end-to-end encryption as it relates to CSEA, but the guidance by Ofcom lists it as a risk factor for other types of offences - suggesting that the scope could expand. The common themes are that encryption reduces the risk (from the criminal's perspective) of detection, and that encrypted services are therefore enablers or facilitators of these crimes. Providers are being asked by Ofcom to assess each kind of illegal harm occurring on their service, not only the presence of illegal content but an offence being committed or facilitated. If this approach is expanded, it would vastly increase the scope of what encrypted services will have to monitor, beyond CSEA, as understood to be the requirement under S.121 of the Online Safety Act. It would lead to other categories of content being scanned, removed or diverted. This broadening scope could lead to more risk-averse approaches to risk assessments and therefore content moderation, impacting freedom of expression.

[5] See above, note 1 Paragraph 12.
[6] See above, note 1 Paragraph 8.

Regarding illegal content which is shared on private channels, at present there is little data on the impact of encryption on the dissemination of content which is illegal or harmful to children. As highlighted by countless organisations and networks, including the Global Encryption Coalition[7], implementing measures that would compel online platforms to undermine encryption, such as requiring the use of accredited technologies to scan content shared on private channels, infringes the privacy of users and undermines the security of the whole system, leaving it vulnerable to exploitation by malicious actors. There is also a growing body of evidence[8] which confirms that there are no currently known technologies that can provide a high enough level of accuracy for client-side scanning without jeopardising the security of service.[9]

2. Data collection practices as key functionality for online harm risks (question 2)

When adopting a narrow approach to the term business model which refers only to the revenue model and growth strategy, Ofcom misses the opportunity to consider other aspects that make a huge difference in the way publicly funded or not-for-profit services determine the features of its services and with that influence the incentives for creation or prevention of online harms.

As we pointed out in a previous submission[10] some U2U services even with a large number of users are underpinned by different business models (such as publicly funded or not-for-profit services) and therefore tend to collect far less personal or behavioural data from users and don't use such data for personalised content or ad targeting. In light of the absence of an exemption for public interest platforms in the OSA,[11] we believe that as well as considering particular functionalities and the numbers of users to analyse risk the following factors may also be relevant:

- the amount of personal data collected by U2U services, for example, the number of data points;

[7] "Why is Encryption Essential?" Global Encryption Coalition, accessed 23 February 2024 https://www.globalencryption.org/resources/why-is-encryption-essential/

[8] Gabriëlle op 't Hoog, Linette de Swaart, Jan Essink et al, *Proposal for a regulation laying down the rules to prevent and combat child sexual abuse - Complementary impact assessment* (European Parliamentary Research Service, 2023), Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2023/740248/EPRS_STU(2023)740248_EN.pdf; Harold Abelson et al "Bugs in our pockets: the risks of client-side scanning" *Journal of Cybersecurity*, Volume 10, Issue 1, 2024, Available at: https://doi.org/10.1093/cybsec/tyad020; Jain, Shubham & Cretu, Ana-Maria & Montjoye, Yves-Alexandre. "Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning, 2021, Available at: https://www.usenix.org/system/files/sec22-jain.pdf

[9] See EDRI. Open letter: Hundreds of scientists warn against EU's proposed CSA regulation. Available at: https://edri.org/our-work/open-letter-hundreds-of-scientists-warn-against-eus-proposed-csa-regulation/

[10] Ofcom Call for Evidence: Second Phase of Online Safety Regulation Global Partners Digital Submission September 2023. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0026/277415/Global-Partners-Digital.pdf

[11] Despite proposals for an exemption for Public Interest Platforms presented by a coalition of organisations. For more information, see Wikimedia. Open call by UK civil society to exempt public interest projects from the Online Safety Bill. Available at: https://wikimedia.org.uk/2023/06/online-safety-bill-open-letter/

- characteristics or behaviours that are used to inform the U2U service's personalisation algorithms for organic and paid content;
- the proportion of service revenue that is generated by ad placements;
- the average amount of time that an average user spends on the service, and/or the percentage of content that the user views that they did not directly request or seek out. See, for example, Kevin Roose's research on the harmful properties of YouTube's "rabbit hole" qualities concerning extremist content[12]; and
- the degree of "polarisation" of content that average individual users or groups see on the platform; for example, Twitter's personalised timeline algorithm has been found to disproportionately amplify particular political opinions for particular users.[13]

We believe that the data collection features of U2U services are a strong predictor of the risks of online harms particularly for the categories of illegal content that benefit from virality in its spread. Data collection practices are also the foundation for the development of recommender systems that are listed as a relevant functionality for the determination of risk in Volume 2. Attending only to recommender systems but not to one of their essential components is a missed opportunity to establish the causal link between business models and risks of online harms. This is the fundamental difference between public interest platforms and commercial ad-revenue-driven platforms, which – even if the user numbers were the same – pose very different risks to users.

[12] Kevin Roose, 'The Making of a YouTube Radical', The New York Times, sec. Technology (8 June 2019). Available at: https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

[13] Ferenc Huszár. Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer and Moritz Hardt, 'Algorithmic Amplification of Politics on Twitter'(2021). Available at: https://cdn.cms-twdigitalassets.com/content/dam/blog-twitter/official/en_us/company/2021/rml/Algorithmic-Amplification-of-Politics-on-Twitter.pdf

Volume 3: How should services assess the risk of online harm?

3. Naming a person accountable to the most senior governance body for compliance with illegal content duties and reporting and complaints duties (question 3)

We have previously expressed our concerns about taking a personalised approach to the accountability mechanisms set for OSA enforcement.[14] We continuously worry that adopting a personalised approach creates incentives for the companies (and the human beings responsible for decision-making) to take an overly risk-averse approach to their responsibilities on illegal content and related government demands.

Moreover, named senior managers could face administrative or criminal prosecution under the OSA if they fail to comply with an Ofcom information notice. Ofcom should exercise this power in a necessary and proportionate manner, and as the last resort for compelling compliance.

Furthermore, it may be more challenging for smaller companies to be able to find or designate an individual willing to take on criminal liability for compliance with these duties, either because salaries are less competitive or because employees at smaller companies may fulfil a variety of different roles at the same time. It may be only the largest of technology companies that would be able to pay an attractive enough salary for someone to assume this high level of responsibility.

These factors are likely to further increase dominance by a small number of very large platforms over the UK market, stifling startups and innovation and further entrenching the power of big tech over public discourse and users' freedom of expression. For example, it is not clear how OSA responsibilities will apply to platforms like Wikipedia, where none of the 700 paid staff or contractors play a role in content curation or moderation, relying instead on a global community of volunteer moderators to make democratic decisions on content moderation informed by public discussion and negotiation.

[14] Jacqueline Rowe (Global Partners Digital) "The proposal to expand criminal liability for social media managers in the UK's Online Safety Bill" (2023) Available at: https://www.gp-digital.org/news/gpd-calls-on-uk-government-not-to-expand-criminal-liability-for-social-media-managers-in-online-safety-bill/

4. Types of services considered for governance and accountability measures (question 4)

We welcome the approach of imposing more requirements on large or multi-risk platforms only, and the rationale of not applying the same requirements for U2U and search services (given the different levels of risk, as elaborated in volume 2).

However, the range of obligations that apply to smaller, low-risk U2U and search services is still quite substantial. Given the implementation costs associated with compliance, smaller services may be incentivised to take a conservative approach and implement automated forms of compliance that can result in negative impacts on freedom of expression. Stringent compliance measures and associated costs can act as a barrier to entry for smaller companies and perhaps even an incentive for exiting the UK market altogether. This would undoubtedly negatively impact the plurality of actors including academic, public-interest or community-oriented platforms, or those of an associative nature that are significant for marginalised groups - which would not only significantly impact the diversity of services available to UK residents, but fundamentally the exercise of freedom of expression.

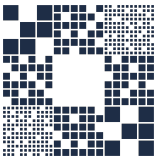
We call on Ofcom to further narrow and tailor obligations, giving particular attention to the implications for public interest platforms, community-led moderation approaches, and smaller niche businesses. As mentioned in our answers to Volume 2, the narrow definition adopted for the term 'business model' needs revision to provide a more nuanced approach to the obligations for the categories mentioned here, with particular consideration for non-profit, public interest and local community service as relevant factors for those determinations.

5. Service's risk assessments (questions 7 and 8)

Ofcom's proposed "four-step" approach encourages services to "assess risks" by considering the likelihood and potential impact of harms occurring on their services. What is very relevant to keep in mind, but sometimes lost in the explanations of the approach, is that this exercise of assessment is intended to determine "illegal content" risk in the services. Although many standards can be used to measure them and determine severity, the limited scope of the obligation to fulfil the OSA mandate should be explicit.

The second step broadly aligns with impact assessment mechanisms in the UN Guiding Principles of Business and Human Rights (UNGPs)[15] and OECD Guidelines for

[15] UN. Office of the High Commissioner for Human Rights, Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, 2011, Available at: https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf



Multinational Enterprises on Responsible Business[16]. However, as it is recognised in human rights impact assessment methodologies proposed by civil society groups,[17] in recognising the multiple competing priorities in assessing likely harms, the concept of severity is also key to orient the prioritisation among them as part of the third step proposed by Ofcom directed on mitigation. This is relevant again to ensure the risk assessment methodology can be tailored to match the size and context of the specific service in question.

[16] OECD, Guidelines for Multinational Enterprises, Available at: https://www.oecd.org/corporate/mne/
[17] Eliška Pírková (Access Now), Marlena Wisniak and Karolina Iwańska (European Center for Not-for-Profit Law), "Towards Meaningful Fundamental Rights Impact assessments under the DSA" (2023) Available at: https://ecnl.org/sites/default/files/2023-09/Towards%20Meaningful%20FRIAs%20under%20the%20DSA_ECNL%20Access%20Now.pdf

Volume 4: How to mitigate the risk of illegal harms - the illegal contents Code of Practice

6. Illegal content Codes of Practice (questions 12, 13, 14, 15 and 16)

We are encouraged to see that Ofcom's approach to developing its illegal content Codes of Conduct aligns with industry best practices, by focusing on ensuring that services have the required systems and services in place to meet their duties, rather than on regulating individual pieces of content. This approach will allow different types of services to develop standards that work best for their systems.

However, there is a risk that services will struggle with the absence of proper legal benchmarks in the guidance, meaning that they may take an overly risk-averse approach to compliance. This issue is further exacerbated by the broad list of priority offences that services will have to contend with. Whilst the guidance documents provide useful information and parameters for service providers, for small services this may create an additional compliance burden. We, therefore, recommend that Ofcom works with smaller services to assist them in coming to grips with the guidance and developing and implementing the illegal content codes of practice guidance.

We welcome alignment with the EU regulations when determining the size of large platforms, however as outlined in our response to Volume 3, we consider that size should not be the sole determinant element for platform classification. Some services even those with a very high number of users are underpinned by different business models -such as publicly funded or not-for-profit services. These services tend to collect far less personal or behavioural user data and don't use such data for personalised content or ad targeting decreasing which risks of the spread of illegal content. In light of the absence of an exemption for public interest platforms in the Online Safety Act, we believe that as well as considering the numbers of users when determining the regulatory burden placed on a particular service, it is also important to consider how the type of business model may impact the level of risks - namely the difference between public interest platforms and non-profit services to commercial ad-revenue-driven platforms. By failing to acknowledge the different types of platforms when setting regulatory standards, there is a risk that not-for-profit and public interest platforms, which may not have the same level of resources as other large platforms, will struggle to ensure compliance and therefore be forced to exit the UK market, harming diversity and perpetuating existing market dominance of a few large actors.

7. User-to-User content moderation (question 18)

We welcome the language in terms of requiring the "awareness" of services for taking down content, and we appreciate that there is no general requirement for services to have dedicated reporting mechanisms (except for large services at medium to high risk of fraud). The creation of dedicated reporting mechanisms, plus ensuring numerous well-trained flaggers would take time and be an additional compliance burden for services. Also, the inclusion of enforcement agencies as trusted flaggers (as is the recommendation for fraud set out in the guidance) can be problematic, as it creates a privileged channel for government control of content.[18] This underscores the need for transparency for dedicated reporting mechanisms, particularly when they involve privileged channels for government requests.

We welcome clear and easily understandable guidance for services on setting and implementing content policies and appeals processes. Also, we value Ofcom's flexible approach, which recognises that different services will rely on different forms of moderation which can include a combination of both human review and automated systems - rather than proscribing a strict approach.

We'd like to further emphasise the importance of human review and moderator training to ensure accuracy. Some recommendations for best practices include that:

- 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 decision-makers, 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;
- 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 particular content type, and should be assessed for psychological suitability for deployment on that content type before working on it.

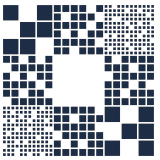
[18] See Electronic Frontier Foundation. Enforcement Overreach Could Turn Out To Be A Real Problem in the EU's Digital Services Act. Available at: https://www.eff.org/deeplinks/2022/02/enforcement-overreach-could-turn-out-be-real-problem-eus-digital-services-act

We would also like to emphasise the disproportionate impact on vulnerable and marginalised communities who are at a higher risk of over-enforcement[19] and having their content removed for allegedly violating policies. For example, research has found that it can be difficult for both human moderators and automated moderation systems to understand the nuances in content relating to activism and counter-speech,[20] which can lead to the censorship of marginalised voices, impacting rights to freedom of expression and access to information.[21] In light of these increased risks, we encourage Ofcom to conduct further research on the differential impact for such communities and find ways to mitigate that risk. This could include, for example, consultations with particular at-risk groups and communities. A further suggestion would be to provide incentives for accurate illegal content removal. Given the broad scope of the OSA, the amount of content being removed is likely to increase exponentially, and with this comes greater room for errors and the over-removal of legal content.

8. Automated content moderation (questions 20, 21, 22 and 23)

We underscore the importance of restricting the use of CSAM hash matching to publicly communicated content only, as the current guidance has set out. Whilst we acknowledge the urgent task of combating online CSEA, approaches to tackling CSEA are often accompanied by calls for the need to scan all content - even content which is communicated privately via encrypted services. As noted in our response to Volume 2, the right to communicate privately without interference is a critical element of freedom of expression and the right to privacy. Furthermore, there is research[22] confirming that currently there are no technically feasible means of allowing access to end-to-end encrypted channels without affecting the security of the system as a whole. Such measures have also been deemed to be insufficiently accurate in their detection, leading to risks of false positives, as well as the potential vulnerability of hash databases with public algorithms being subjected to 'poisoning attacks[23].' The complementary impact assessment of the EU Digital Services Act also recently came to this conclusion on the availability and accuracy of existing technologies to scan encrypted content.[24]

[19] Ángel Díaz and Laura Hecht-Felella, *Double Standards in Social Media Content Moderation* (Brennan Centre for Justice, 2021) Available at: https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf
[20] see Electronic Frontier Foundation. One Database to Rule Them All: The Invisible Content Cartel that Undermines the Freedom of Expression Available at: https://www.eff.org/deeplinks/2020/08/one-database-rule-them-all-invisible-content-cartel-undermines-freedom-1
[21] See Human Rights Watch. Meta's Broken Promises: Systemic Censorship of Palestine Content on Instagram and Facebook. Available at: https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and
[22] See above, note 8.
[23] Seny Kamara, Mallory Knodel (Center for Democracy and Tech) et al. Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems (2021) Available at: https://arxiv.org/ftp/arxiv/papers/2202/2202.04617.pdf
[24] See above, note 8.



As set out in our previous response,[25] automated systems using hashing technology to detect known CSAM imagery are relatively reliable and can be deployed at scale. Such systems should still be regularly reviewed and assessed for accuracy and impact on users. Affected users should always be informed when a decision that affects them is made by automated systems, and should always be allowed 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. Where services rely on hash matching, there are measures they can take to improve systems and accuracy, such as:

- Ensuring that the databases of known illegal content scanned by the hashing algorithm are either verified by a trustworthy, independent party (such as the Global Internet Forum to Counter Terrorism's hash-sharing database for terrorist content) or are securely maintained by the online service provider itself, subject to regular audits to ensure that all matches generated by the hashing system are for genuinely illegal content; and
- Ensuring that hashing systems can still flag known illegal images that have been cosmetically altered, for example by cropping or changing image contrast, through perceptual hashing techniques.

In light of the above, we urge Ofcom to retain a narrow scope for the use of perceptual hash matching. Broadening out to use such technology on private communications - whether server-side or client-side - entails privacy risks, even in situations where the technology is simply checking for a match without learning anything about the material itself. By expanding to include private communications, the risks of false positives are also heightened.

As well as focusing on public communications only, it is important to retain a narrow focus on illegal content to which detection of such technologies will be applied. In its current guidance, Ofcom has restricted this to CSAM and determining the legality of a piece of content relating to CSAM is considerably more straightforward in comparison to other more complex criminal offences set out in the OSA. As such, we urge Ofcom to be wary of scope creep and the requirement of platforms to deploy hash matching to other forms of illegal content, as this would create more room for error and risk the widespread removal of legal content.

Finally, we suggest further clarity on what would be classed as a 'private communication' in annex 9 - which relates to whether the communication is public

[25] Ofcom Call for Evidence: Second Phase of Online Safety Regulation Global Partners Digital Submission March 2023. Available at: https://gp-digital.org/wp-content/uploads/2023/06/GPD-Ofcom-submission-March-2023.pdf

or private, rather than whether the content is private. This includes adding explicit reference to encrypted services and taking them into account in the deliberation of whether the content is communicated privately, as this will help distinguish private messaging from other types of service and the consequent mitigations needed to comply with the regime.

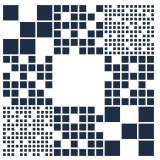
9. Reporting and Complaints (question 28)

We welcome Ofcom's guidance on the need for services to introduce robust reporting and complaints mechanisms. As set out in our earlier submission,[26] providers of online services should make reporting and complaint routes available for both users and non-users, given the fact that harmful content may impact a broad range of individuals who are not users of a particular online service, particularly those which are smaller or medium-sized. Any content which is available to or visible to non-registered users should be accompanied by relevant reporting and complaints systems which are also available to registered users.

There are also specific steps which services to should take enhance the transparency, accessibility, use experience and awareness of their reporting and complaints mechanisms such as:

- Clearly explaining to users submitting a complaint what will happen to the complaint at each stage and how long they can expect the process to take;
- Notifying the individual or entity responsible for the cause of the complaint that a complaint has been made, and explaining how it will be reviewed and what the potential outcomes will be;
- Providing the individual or entity responsible for the cause of the complaint a chance to rebut or provide counter-evidence or context;
- Providing a clear explanation and justification to all relevant parties for any decision made or action taken in response to the complaint, referring to the specific sections of the terms of service where a violation has been identified;
- Informing all parties if the review of the complaint or report has been undertaken by an automated tool, and allowing any party to request a human review of the merits of their complaint;
- Ensuring that appropriate safeguards and verification measures are in place to protect complaints and appeals systems from misuse or abuse by malicious actors (e.g. in an attempt to censor content that they do not like); and
- Explaining clearly how any data or content shared as a result of a complaint will be stored, assessed and deleted. This is particularly important to complaints or appeals over content shared on private or encrypted services,

[26] Ofcom Call for Evidence: First Phase of Online Safety Regulation Global Partners Digital submission September 2022. Available at: https://gp-digital.org/wp-content/uploads/2022/09/Ofcom-Call-for-Evidence-.pdf



or over complaints relating to certain forms of content such as the non-consensual sharing of intimate images. Online service providers should ensure that rigorous safeguards and protections are in place for user privacy throughout the complaints and appeals process;

- Ensuring through software design that users can easily report or make a complaint about any content that they encounter, in any format, including comments, private messages, multimedia and content shared within closed groups, as well as public posts and public webpages;
- Using unambiguous and non-technical language for all reporting and complaints mechanisms, instructions and supplementary information, that is 3 understandable to the average user (and has been tested with users to ensure this is the case);
- Translating all reporting and complaints mechanisms, instructions and supplementary information into all languages in which the online service is used and available, including those spoken by minority groups and immigrant communities (in consultation with local experts);
- Ensuring that reporting and complaints mechanisms, instructions and supplementary information are hosted in a way which is compatible with assistive technologies used by individuals with disabilities, and/or creating audio or video versions of the documents (working in consultation with those with disabilities to ensure effective solutions);
- Providing users with pre-prepared options or categories for their complaint as well as an open complaint category (in cases where the user is not sure which category to use or feels that no categories are suitable);
- Allowing users to provide more detail on the context or substance of their complaint if it is not clear from the original content itself;
- Confirming receipt of the complaint, ideally with a reference number that users can use to follow up easily;
- Offering users the option of downloading or having a copy of their complaint sent to them (provided that non-registered users consent to providing relevant contact details);
- Ensuring that appeals mechanisms are designed to be just as clear, accessible, transparent and easy to use as the primary complaints mechanisms, in the ways outlined above;
- Including along with any decision issued clear information about each affected party's right to appeal the decision, including both internal and external appeals processes;
- Regularly (e.g. once per year) reminding users through a pop-up or notice of how to use the reporting and complaints mechanisms (this may only be possible for registered users, and may have to be randomised frequency for non-registered users); and

- Where a piece of content has been identified as suspicious, for example, by an automated tool or by viral activity, the online service provider might prompt users about their reporting and complaints mechanisms concerning that specific piece of content (e.g., "Are you concerned about this content? Report it here.").

Reporting systems should recognise that vulnerable users, in particular children, may not be competent or able to make use of the reporting and complaints mechanisms designed for adult users. This may be due to a lack of awareness that particular content is wrong (for example, in the case of child grooming), a lack of knowledge of reporting and complaints mechanisms (for example, if the child does not know about this feature of the platform), or a lack of understanding of how to use the reporting and complaints mechanism (for example, if the child does not know which category their complaint falls into or does not understand the instructions). In addition to the above recommendations, there are further steps providers should consider to help children access and use such mechanisms effectively:

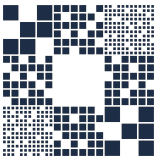
- Creating age-appropriate content regarding digital safety for child users to learn from, either upon signing up for a service or regularly (e.g. once per month) during their use of the service, which could be accompanied by games or quizzes for the child to complete which tests their understanding;
- Creating more straightforward mechanisms for underage users to lodge complaints, including simpler or more clearly explained categories, simpler language, graphics and visuals to aid explanation and instructions; and
- Enabling adults to make complaints on behalf of a child under specific circumstances, such as when the adult is a parent or guardian or otherwise responsible for the child, or if the child has given the particular adult permission to make a complaint on their behalf.

10. Terms of service and publicly available statements (questions 29 and 30)

We welcome the obligations in terms of services and guidance on how to organise the substance and present terms of service to make them clearer and accessible.

We would like to draw attention again to our earlier submission[27] which sets out specific suggestions to enhance the clarity and accessibility of terms of service and public policy statements including:

[27] Ibid.



- Providing users with a high-level summary of terms of service or public policy statements (e.g. in the form of simple bullet points), with the option for users to seek more information and detail should they desire;
- Providing clear definitions of important terms, such as "hate speech", "violent content", and "graphic content", wherever such terms are used, with examples of what is and is not included within the definition. This could involve explaining any thresholds that the online service applies when determining if a piece of content is prohibited, and/or providing examples or additional detail to demonstrate what is meant by each term;
- Publishing lists of any organisations or individuals for which content affiliated with or supporting such entities would be in violation of their policies;
- Providing information about what enforcement actions the online service may take in the case of each type of content violation and in case of repeat violations;
- Informing users clearly of how their data will be used, both for routine use and operation of the online service, including for complaints or appeals that relate to the user or the user's content;
- Explaining clearly whether the company will treat public figures differently when it comes to enforcement of its terms of service and if so, how;
- Explaining clearly what exemptions or allowances may be made for violations of the terms of service for journalistic purposes and how such cases are assessed;
- Providing users with reasonable notice of any new policy documents or any changes to terms of service before they take effect; and
- Requiring explicit acknowledgement of the changes in terms of service by users, beyond simple pop-up banners or windows, which are often ineffective means of relaying information as users often ignore or quickly bypass such mechanisms.
- Hosting all terms of service and public policy statements in a centralised location with clear signposting towards different types of documents and information;
- Ensuring through interface design that the location of the terms of service is easily accessible, and that users can search for the relevant information within terms of service or public policy documents (e.g. through a help centre or chatbot function);
- Using unambiguous and non-technical language for all terms of service and public policy statements that is understandable to the average user;
- Using age-appropriate language for terms of service and public policy statements relevant to children using the online service, including graphics, videos, or other creative means of communicating terms of service where appropriate;

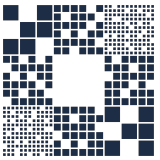
- Translating the terms of service and public policy statements into all languages in which the online service is used and available, including those spoken by minority groups and immigrant communities; and
- Ensuring that terms of service and public policy statements are hosted in a way which is compatible with assistive technologies used by individuals with disabilities, and/or creating audio or visual versions of the documents, as well as working in consultation with those with disabilities to find other effective solutions.

11. U2U default settings and support for child users (question 31)

We understand that additional proposals will be published later this year on the deployment of age assurance technology on U2U services as well as higher standards of age verification for services which have higher users, however, we would like to take this opportunity to reiterate our concerns[28] around the potential adverse impacts on individuals' human rights that such mechanisms may pose[29]:

- Any mechanisms which require the sharing or upload of official identification documents or sensitive biometric data pose risks to user privacy. Even where the online service provider does not retain copies of these documents, the risk of malicious actors hacking or otherwise intervening in such data exchanges remains salient and could result in abuse of personal information. This could also adversely affect vulnerable groups, including children;
- Any mechanisms which require the sharing or upload of official identification documents or sensitive biometric data would remove the possibility of individuals being able to use services anonymously, which may be vital for certain vulnerable or persecuted groups to be able to access and share information online without fear of reprisal;
- Any mechanisms which require the sharing or upload of an official or up-to-date ID may adversely affect the freedom of expression of some of the most vulnerable users, who may not have access to an ID due to financial limitations, homelessness, or due to being a victim of human trafficking or controlling partnerships;
- Any mechanisms which rely on machine learning tools for age estimation will contain a margin for error which, even if small, would adversely impact individuals' right to freedom of expression by preventing them from accessing or sharing information when they should be able to do so;

[28] Ibid.
[29] See EDRi. Position paper: Online age verification and children's rights (Oct, 2023) Available at: https://edri.org/wp-content/uploads/2023/10/Online-age-verification-and-childrens-rights-EDRi-position-paper.pdf



- Any mechanisms which rely on machine learning tools for age estimation or verification may pose risks to individuals' right to non-discrimination, as such tools are less accurate for particular racial groups or genders; and
- Any age verification systems run the risk of creating a two-tiered internet, as well as serving as a deterrent for many adults accessing legal content.

Where online service providers are required to use age verification measures, whether these are designed in-house or outsourced to an external company, providers should ensure that:

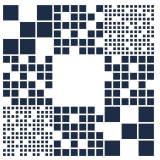
- The highest standards of data privacy are in place for users sharing personal IDs or sensitive biometric data, and no such data is retained longer than the period necessary to conduct the age check;
- Individuals who do not wish to or cannot share a personal ID or biometric data are provided with alternative means of verifying their age, or are provided with alternative means of accessing adult portions of the site;
- Users can appeal any determinations or estimations of age made by an automated tool, and are provided with alternative means of verifying their age where they claim that the decision of the automated tool is incorrect; and
- All age verification measures are assessed for potential impacts on human rights and potential biases, and any such impacts or biases are addressed before roll-out. It would also be valuable for online service providers to collect and publish data on exactly how effective age verification measures are at preventing children from encountering harmful or illegal content online, as well as any information on how underage users may be circumventing the age checks to access adult content intentionally.

12. User access (question 40)

We welcome the consideration of the human rights impact of each sanction that impacts user access, particularly freedom of expression and association.

As outlined in our previous submission,[30] sanctions or restrictions around access are applied by providers of online services through various means. These may include restricting the accessibility or shareability of content, de-amplifying or deprioritising it in ranking algorithms of content, providing warnings or flags over the content, redirecting users away from the content, removing the content entirely, and temporarily or permanently removing a user or entity or removing certain functionalities or services available to them. In some cases, particular content types may be referred to law enforcement. These sanctions may be determined either by a human moderator or human review team or by an automated tool. These sanctions

[30] See above, note 26.



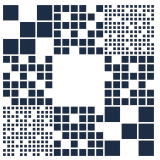
and enforcement mechanisms may have considerable adverse impacts on users' human rights. Further guidance on sanctions and enforcement mechanisms that online platforms should be aware of includes that:

- Unwarranted sanctions, whether imposed by an automated tool or by a human moderator, may result in the temporary or permanent disabling or removal of content which is not unlawful or against terms of service, which may have a detrimental impact on the ability to impart and receive information of all kinds;
- Passing on suspected illegal content to law enforcement poses significant risks to user privacy, and should only be justified where explicitly required by law and in relation to the most serious forms of illegal online content. In each case, the decision should be made by a human reviewer, and the relevant user notified of the action being taken;
- Implementing "three-strike rules" or similar means of assessing repeat offenders on the online service before taking action against a particular user requires the retention of user data and violative content shared by the user, as well as data on those who have submitted complaints relating to the user in question. This may require the processing and storing of personal information on individuals wishing to remain anonymous, posing risks for individuals' privacy and personal data; and
- Enforcing sanctions inconsistently across different users or groups may result in a disproportionate level of removals or de-platforming of particular groups, particularly in cases where the sanctions are erroneous. This may threaten individuals' right to non-discrimination.

All of these human rights risks should be carefully assessed in accordance with the potential harms caused by not implementing such sanctions and enforcement policies, in consultation with experts on free expression, privacy and other affected human rights. Wherever an automated tool cannot make a determination with a high degree of certainty, it should be passed on to a human moderator. Similarly, wherever a human moderator is at all uncertain of the correct course of action or how to apply the terms of service in a particular case, there should be the possibility of passing the case on to a more experienced or specialist moderator, to reduce the likelihood of unwarranted sanctions.

13. Cumulative assessment of proposed measures (questions 45 and 46)

As mentioned in our response to Volume 3, the requirements for small and micro companies are substantial. Given the implementation costs associated with compliance, smaller services may be incentivised to take a conservative approach and implement automated forms of compliance that can result in negative impacts



on freedom of expression. Even for some smaller actors, these compliance measures and associated costs can act as a barrier to entry or even an incentive for exiting the UK market altogether.

We understand the need for additional requirements for small and micro businesses that have significant risks of illegal content, in the interest of minimising the risk of users accessing illegal content. However, as Ofcom has recognised in the guidance, the cumulative impact of these provisions can be expensive and discouraging. Requiring small businesses with more limited means to implement expensive technologies such as hash matching, could increase market concentration in niche services. To attempt to address the burdensome nature, we ask that Ofcom provide additional hands-on guidance and support for small businesses in this position to assist with the transition.

Volume 5: How to judge whether content is illegal or not? (Illegal Content Judgements Guidance)

14. Illegal content determination is bound by what's in the OSA - expanding beyond this remit will not fulfil the requirements for breaching freedom of expression (questions 49, 50 and 51)

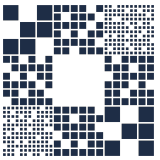
The illegal content judgements must refer exclusively to a relevant offence within the OSA. The Illegal Content Judgements Guidance must make clear to providers that going beyond this remit will not fulfil the requirements for breaching freedom of expression. This is particularly important given that providers may be incentivised to be overly cautious in their takedown duties, to not fall foul of Ofcom's rules.

The Illegal Content Judgements Guidance should acknowledge that in all cases the final authority for legal determinations on the legality of content is the judiciary. Outsourcing this power to private entities, who do not have the same obligations to the public, could lead to incorrect determinations and inconsistencies inside and across different platforms.

A further problem with this approach is that it undermines the critical role that the criminal justice system plays in ensuring accountability for the commission of criminal offences. Taking into consideration examples where individuals have been convicted of posting hateful or of sharing CSAM, had their posts been deleted by platforms instead, they may not have been investigated, prosecuted and sentenced. There would have been no accountability for the criminal behaviour, as well as the corresponding deterrence factor for both the convicted individual and others considering similar conduct. While it is reasonable to expect online platforms to remove content which has been identified as illegal by an authoritative body, such as a court, expecting platforms to make those determinations themselves gives them a level of authority to make decisions for which they're not equipped.

Requiring platforms to determine the legality of content and particularly the mental element present to fulfil a new threshold of "reasonable grounds to infer" is burdensome and can also lead to mistakes - further risking an overly risk-averse approach to content removal and impacting users' right to freedom of expression. We have previously raised similar concerns[31] about the difficulties of determining the legality of a particular piece of content. Determining whether a particular piece of speech is illegal or not is not simple, and whilst for certain offences it can be relatively straightforward (such as for CSAM), for many relevant offences it is not. The OSA covers a broad range of offences - requiring platforms to understand the

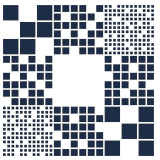
[31] Joint Committee on the Draft Online Safety Bill. Global Partners Digital submission September 2021. Available at: https://gp-digital.org/wp-content/uploads/2022/09/Ofcom-Call-for-Evidence-.pdf



complexity of each offence, as well as the various contextual factors that need to be considered to determine legality, with no legal expertise, is no easy feat. In particular, determinations of the 'mental element' of an offence can be extremely difficult to make without a clear understanding of the relevant offences and thorough investigation. Once again, the challenges will be more acute for academic, public-interest or community-oriented platforms, or those of an associative nature that are significant for marginalised groups.

Given the complexity of this task, it is inevitable that platforms will make rushed and incorrect decisions frequently; and given the potential outcomes for a failure to comply with the OSA, platforms are likely to err on the side of being overly cautious, which could lead to the censorship of legal speech. This issue is heightened by the fact that the Illegal Content Judgements Guidance does not cover what can be understood as 'reasonable grounds to infer' where a general defence is available. We appreciate that it may be difficult to access and understand the contextual elements in user interactions of an alleged piece of illegal content, however, that difficulty seems to be no greater than conducting other contextual assessments advised by the Guidance, for example around the "appearance of being under age", "journalistic or academic purposes" or "humour and satire". Therefore additional consideration should be given to the inclusion of guidance in the general defences.

Finally, we express our concern that given the extensive list of what constitutes a priority offence that platforms will need to make determinations on, there is a risk that platforms will fail to address the most severe forms of illegal content that cause the most harm simply by a matter of resource allocation. Here we come back to our call that in the risk mitigation stage of the risk assessment, prioritisation of action should be given a central place.



Volume 6: Information gathering and enforcement powers and approach to supervision

15. Transparency in the implementation of information notices and enforcement action (questions 52 and 53)

As an overarching point, we would like to emphasise the importance of transparency in the implementation of information notices and enforcement action. Government involvement in the determination of illegal content poses a real risk of censorship and interference with freedom of expression. Whilst Ofcom is an independent body, given the wide-ranging powers conferred by the OSA, demonstrating a proportionate exercise of those powers through transparency mechanisms is crucial to building trust with services and users.

We remain concerned with the imposition of criminal liability for senior managers[32] and urge Ofcom to only trigger criminal liability as a proportionate last resort, as emphasised by the Joint Parliamentary Committee report,[33] and only in cases where there is considered to have been a serious breach. As already outlined in response to Volume 3, adopting a personalised approach creates incentives for the services (and the human beings responsible for decision-making) to adopt risk-averse approaches to removing illegal content and compliance with demands from government bodies and enforcement agencies.

We are also troubled by Ofcom's ability to request remote access to live systems and live user data, which could pose serious security concerns and privacy risks due to the disclosure of data.[34] This could also be seen as a workaround to access data that would ordinarily require a warrant to be obtained. We urge Ofcom to exercise caution in the implementation of such powers, and to introduce additional safeguards to mitigate the potential security risks, for example by limiting access to a testing environment where there is no user data, or by not using it to access live data.

Finally, concerning the power to obtain a skilled person report, we request clarity on who would be classified as a 'skilled person' and how such a determination will be made by Ofcom. One suggestion is to adopt a similar approach to the FCA's Skilled Persons review[35] which includes a publically accessible list of 'skilled persons' based on different categories - regulated firms can select from the list.

---

[32] See above, note 14.
[33] Joint Committee on the Draft Online Safety Bill, Draft Online Safety Bill: Report of Session 2021–22, https://committees.parliament.uk/publications/8206/documents/84092/default/, paras. 360-369
[34] See TechUK. Statement on the Online Safety Bill, Available at: https://www.techuk.org/resource/techuk-statement-on-the-online-safety-bill.html
[35] "Skilled persons reviews," Financial Conduct Authority, Accessed 23 February 2024 https://www.fca.org.uk/about/how-we-regulate/supervision/skilled-persons-reviews#section-appointing-a-skilled-person