Online Harms White Paper: Global Partners Digital Briefing

About Global Partners Digital

Global Partners Digital (GPD) is a social purpose company dedicated to fostering a digital environment underpinned by human rights and democratic values. We do this by making policy spaces and processes more open, inclusive and transparent, and by facilitating strategic, informed and coordinated engagement in these processes by public interest actors.

Introduction

This briefing has been prepared to inform the proposals which will be set out in the government's upcoming Online Harms White Paper. In this briefing, we first set out some general considerations and concerns about risks to freedom of expression in any proposals. We then look at four specific proposals which are being considered, set out specific considerations and concerns, and make recommendations which would mitigate those risks.

We also wish to highlight the importance of non-legislative measures as a means of tackling online harms. Many of these were noted in the Internet Safety Strategy Green Paper and, in particular, we consider that there is significant potential when it comes to measures such as encouraging 'safety by design' when new products and services are developed, public education and promoting digital literacy, and specific education in schools on the safe use of the internet and technology. We would strongly encourage all non-legislative measures to be fully considered and developed, and for legislative measures to be taken only where alternatives insufficiently address the online harms which need to be tackled.

1. General considerations in relation to proposals to tackle online harms

a) Lessons from other countries

Any proposals which set out requirements or incentives for platforms to remove content naturally pose a risk to freedom of expression online. While certain forms of content can, of course, be justifiably restricted (such as child sexual abuse imagery, incitement to violence, etc.) as in pursuance of legitimate aims (such as the prevention of crime or the protection of the rights of others), our concerns relate to risks of content being removed which is in fact protected by the right to freedom of expression. It is important to note, therefore, that this right "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".[1] Any restrictions on this kind of content must, therefore, not be blanket, but proportionate and targeted only where actual harm is likely to occur as a result.

Proposals which require or incentivise the removal of content that is unlawful or harmful, if inappropriately designed and implemented, risk content being removed which is in fact neither unlawful nor harmful, and we see this risk potentially manifesting in a number of ways.

First, **any proposals which establish time limits for the removal of content which is unlawful or harmful.** There are a number of concerns here. The imposition of time limits would incentivise the use of automated processes for determining whether content is unlawful or harmful. However, automated processes are extremely poor at making determinations relating to the nature of content given their inability to determine context, and the difficulties in defining terms such as "bullying" or "insult".[2] The recent example of Tumblr which has used automated processes to identify content which breaches its standards on "adult content", with large swathes of innocent content being flagged shows how easily reliance on automated processes can lead to over-removal of content.[3] Even where human moderation is involved, short time limits risk rushed decisionmaking and an inability to fully consider context or obtain the necessary information and expertise in order to make an accurate determination.

Second, **any proposals which impose sanctions or penalties for failure to remove content further incentivise the removal of content**. If a platform is making decisions as to whether to remove content or not on the basis that it might potentially be unlawful or harmful, there will be a strong incentive to "play it safe" and simply remove the content rather than risk a sanction. As noted above, freedom of expression includes information and ideas which are offensive, shocking and disturbing; however, platforms will be strongly incentivised to remove all of these kinds of content on the basis that they might be unlawful or harmful, even if they are, in fact, neither. Evidence from the implementation of the Network Enforcement Act (NetzDG) in Germany in 2018 suggests that this would be the case: since the introduction of the tight timelines and heavy fines included in the NetzDG legislation (48 hours in the case of "manifestly unlawful" content), there have been a number of high-profile examples of Twitter, for example, removing tweets which were controversial, satirical and ironic, but not obviously illegal or even harmful.[4]

Third, **broad but undefined categories of what is considered 'harmful' content.** While some forms of harmful content have clear, legal definitions, many do not. Without definitions, and particularly if there are sanctions for non-compliance with any duties, platforms will be incentivised to interpret the terms broadly, rather than risk sanctions, and, therefore, remove an even broader range of content than is

[1] *Handyside v the United Kingdom*, Application No. 5493/72, 7 December 1976 (European Court of Human Rights)

[2] See, for example, Center for Democracy & Technology, "Mixed Messages? The Limits of Automated Social Media Content Analysis", 28 November 2017, available at: https://cdt.org/insight/mixedmessages-the-limits-of-automatedsocial-media-content-analysis

[3] See, for example, Montgomery, S. J., "Here's Some of the Random Content Tumblr Is Flagging for Its No-Porn Policy", complex.com, 5 December 2018, available at: https://www.complex.com/life/2018/12/content-tumblr-is-flagging-for-no-adult-content-policy/; Romano. A., "Tumblr is banning adult content. It's about so much more than porn", Vox, 17 December 2018, available at: https://www.vox.com/2018/12/4/18124120/tumblr-porn-adult-content-ban-user-backlash.

[4] See, for example, Scott, M. and Delcker, J., "Free speech vs. censorship in Germany", Politico, 14 January 2018, available at: https://www.politico.eu/article/germany-hate-speech-netzdg-facebook-youtube-google-twitter-free-speech.

intended, including content which is protected by the right to freedom of expression. We are particularly concerned that this could create a perverse situation where speech which is lawful, but potentially harmful, is restricted when it is expressed online, but not when it is expressed in person.

Fourth, **child-focused requirements or incentives on content removal becoming the default.** We note that in an interview with the *Telegraph* in November 2018, the Home Secretary, Sajid Javid, proposed a duty of care for social media platforms, exclusively aimed at protecting children from online harms.[5] However, the online ecosystem does not generally distinguish between adults and children when it comes to its users. Indeed, children often use the same platforms and websites as adults. As such, there is a risk that in complying with any proposals, such as a duty of care, platforms would simply remove any and all content that could be harmful to children, thereby making such content unavailable for adults as well, despite the fact that no harm would be caused by its availability to them.

To mitigate these risks, we therefore recommend:

- Proposals **should not include time limits for the removal of content**, unless it is in relation to content which has already been identified as unambiguously unlawful or harmful, regardless of context (such as child sexual abuse imagery);
- Proposals **should not include penalties or sanctions for non-compliance as a first step**, but only if alternative efforts to ensure compliance have been unsuccessful (see below for further detail), with the same exception as above in relation to content which has already been identified as unambiguously unlawful or harmful, regardless of context; and
- Proposals **should ensure that all forms of harm which are to be addressed by platforms have clear, precise definitions**.

We would further recommend that, to ensure that the right to freedom of expression is not just built into the processes which are developed, but is an explicit consideration at all stages of any process, the following are considered:

- Any legislative proposals – such as a Code of Practice or a duty of care – **explicitly state that the importance of protecting and respecting the right to freedom of expression is to be taken into account** when platforms make decisions and when compliance is being assessed; and
- Any regulatory guidance, whether in secondary legislation or produced by the government or a regulator, which is developed for the purposes of ensuring compliance **includes a section on the importance of protecting and respecting the right to freedom of expression**.

b) A lack of transparent and accountable decisionmaking by platforms

The same principles which underpin permissible restrictions on freedom of expression apply online as they do offline. This means that restrictions, including the removal of online content, should only take place following a clear, transparent and rights-respecting process, with appropriate accountability and the possibility of an independent appeal process.

When it comes to unlawful content, proposals which contain requirements for platforms to remove content would shift judicial and quasi-judicial functions to platforms, or their nominees, forcing them to make determinations regarding whether particular forms of content are legal or not. This would include determinations on whether certain content constituted, among others, hate speech, defamation and incitement to violence or terrorism. However, unless mandated, there would be no guarantee that there would be mechanisms for accountability or

[5] Hymas, C., "Facebook, Google and other tech giants could be forced to accept a duty of care", The Telegraph, 8 November 2018.

safeguards in place, as there are when decisions are made by public authorities or the judiciary.

Where the proposals extend to "harmful" but lawful content, such as bullying or insulting other individuals, there are similar concerns over whether platforms are well-placed and able to make determinations as to what content is harmful, particularly if no clear, precise definitions are provided (see above). The sheer scale of content means that in-person reviews are unlikely to be feasible, and we have highlighted above how automated processes are notoriously poor at making decisions at identifying this kind of content.[6] And, as with unlawful content being removed, there would not necessarily be any mechanisms for accountability nor safeguards in place to challenge decisions.

To mitigate these risks, we therefore recommend:

- Proposals **should ensure that all forms of harm which are to be addressed by platforms have clear, precise definitions**; and
- Proposals **should require platforms to ensure that any decisionmaking about content takes place following a clear, transparent and rights-respecting process**. This should include, at a minimum, (i) enabling affected users to be informed of content that has been flagged for review, and able to input into that decisionmaking process, and (ii) ensuring that there are independent appeal mechanisms for affected users to challenge decisions.

c) Proposals being adopted in other jurisdictions with more harmful consequences

A trend of states passing copycat legislation relating to the internet, including that regulating online content, has been gathering momentum over the last twelve months. For example, shortly after the introduction of the NetzDG in Germany, a near-identical version was put forward in the Russian Duma. However, while there are certainly concerns in relation to the German legislation, the adoption of the legislation in Russia would be even more problematic given the absence of any effective national human rights framework and the existence of criminal laws which prohibit expression in violation of international human rights standards.

As such, any proposals which are put forward in the UK have the potential to be adopted in other states which could then point to the UK framework for justification. In states where speech which should be protected under international human rights law is criminalised or where there are no effective safeguards, such as an independent judiciary or a national human rights institution, for example, the effects could be far more restrictive than they would be in the UK.

To mitigate these risks, we therefore recommend that proposals **should explicitly set out all of the safeguards that exist to ensure that the right to freedom of expression is not adversely impacted** and which can be pointed to if and when the proposals are considered in other jurisdictions. These could include, in addition to the recommendations listed above:

- Involving the Equality and Human Rights Commission in the development of any guidance, in the establishment of any new regulator, in the enforcement of any requirements and with a role of reviewing the overall process to determine impacts upon freedom of expression;
- Ensuring that the regulator and its decisions are open to judicial review, and are considered a public authority for the purposes of section 6 of the Human

[6] See, above, note 2.

Rights Act 1998, which prevents a public authority from acting in a way which is incompatible with the rights under the European Convention on Human Rights.

2. Considerations on specific proposals

a) A Code of Practice

With respect to the draft Code of Practice published alongside the government's response to the Internet Safety Strategy Green Paper, given its narrow focus on empowering users to be able to report content and activity, such as bullying, and for greater transparency on decisionmaking, there are few concerns from a human rights perspective.

Problematic, however, is the draft Code's reference to "using a mix of human and machine moderation" as an example of good practice when it comes to moderating content. There are real concerns, as noted above, over the use of automated processes for determining whether content is harmful, given the inability of machines to determine context, and the difficulties of defining terms like "bullying" or "insult".[7] There is a high risk that legitimate content could be removed through automated processes, which means that there should always be human assessment at some point in the process. The draft Code of Practice makes no reference to the need for human involvement at some point during moderation processes which rely on, or start with, automated review.

We therefore recommend that:

- The final Code of Practice should make clear that **where automated processes are used, there is also human involvement at some point of the decisionmaking process**; and
- Any legislation establishing a duty of care should be **consistent with our recommendations under 'General considerations'** and, in particular, the need to avoid time limits and sanctions, the importance of making sure that any specified harms are clearly defined and that child-focused requirements or incentives on content removal do not become the default.

b) A duty of care

In principle, we acknowledge that imposing a duty of care on platforms appears to be a pragmatic solution to address online harms: it is an established legal doctrine that would be relatively straightforward to legislate. Additionally, it also holds the potential to be flexible enough to cover the wide variety of platforms that host user-generated content. However, there are also a number of potential risks to human rights, which arise from the establishment of a duty of care, which are set out above, where the duty of care – or interpretative guidance – includes time limits, sanctions for non-compliance, broad and unclear definitions of specified harms or would lead to an approach whereby all content, regardless of audience, is assessed on the basis of potential harms to children.

We therefore recommend that:

- Any legislation establishing a duty of care **should be consistent with our recommendations under General considerations** and, in particular, the need to avoid time limits and sanctions, the importance of making sure that any specified harms are clearly defined and that child-focused requirements or incentives on content removal do not become the default.

[7] See above, note 2.

c) A regulator

We recognise that the imposition of any new duties, such as a binding Code of Practice or a duty of care – requires some means of enforcement. The establishment of an independent regulatory body is therefore an understandable proposal. Our general concerns in relation to the establishment of any such regulatory body are set out above under General considerations but we would further recommend the following:

- Any regulatory body established **must be fully independent from government and multistakeholder in nature**. It should include all relevant stakeholders including government, platforms, academia and civil society;
- Any legislative proposals establishing the regulatory body **must explicitly state that protecting and respecting the right to freedom of expression is one of its functions**, or a consideration to be taken into account when carrying out any of its statutory functions;
- The Equality and Human Rights Commission **should be involved in the process of establishing the regulatory body** and able to review its work; and
- The actions and decisions of the regulatory body **should be open to judicial review**, and are considered a public authority for the purposes of section 6 of the Human Rights Act 1998, which prevents a public authority from acting in a way which is incompatible with the rights under the European Convention on Human Rights.

When it comes to enforcement, we believe that such a graduated approach, which incentivises and/or requires full transparency and improved action being taken through self-regulation, should be the starting point. Only when this has clearly not been sufficiently complied with should sanctions be a potential penalty open to the regulatory body. For example:

- As a first step, **platforms should be required to provide sufficient detail on what action they are taking in relation to specified harms through transparency reporting**. The template for that transparency report could be developed by the regulatory body. Where the regulatory body considers that the platform has fully complied with its transparency reporting requirements, and that they demonstrate sufficient action is being taken, the platform should be immune from any sanctions, deprioritised in some way when it comes to reviews or be able to rely on this fact in any enforcement process;
- As a second step, where there is a failure to comply with transparency reporting requirements, or these do not demonstrate sufficient action being taken, the regulatory body **should have the power to demand such reporting or set out specified actions that should be taken to ensure compliance**; and
- Only as a third and final step **should a platform be subject to sanctions** for failure to comply with any duties relating to content removal.

d) Transparency

As noted above, we believe that transparency should be the starting point in any regulatory proposal and only where such transparency does not show that sufficient action is being taken should further enforcement mechanisms be undertaken. However, the precise form of the transparency reporting requirements will have a significant impact on whether they encourage positive behaviour, or simply

incentivise the removal of further content.

Transparency reporting requirements which simply require platforms to set out how many reports of harmful content they have received and what percentage they have taken down, for example, would risk inadvertently incentivising platforms to take steps to increase those numbers so that they appear successful or are making "progress." This will be the case particularly if any legal requirements explicitly or implicitly indicate that a certain percentage of content under each category of harm reported should be removed, or that the percentage should go up over time. Further, any reporting requirements which relate to the time taken for content to be reviewed and removed after being reported risk incentivising those periods to reduce over time. This would encourage either the use of automated decisionmaking or for quicker decisions to be made by human moderators, risking greater inaccuracy.

In contrast to purely quantitative forms of transparency which, as noted above, risk simply encouraging the removal of content and more quickly, qualitative forms of transparency could create opportunities for greater respect for freedom of expression by platforms, as well as more effective tackling of online harms. At present, it is not always clear how platforms make decisions about what content to remove, the standards and processes that are employed, those involved in the process, and how quality of decisionmaking is ensured. As a result many have raised concerns that platforms are taking too much content down, not taking enough down, and failing to be sufficiently transparent about how they take down content in the first place. Mandatory transparency reporting requirements could help address these concerns. They would encourage platforms to develop clear terms of service which explain what content is and is not allowed on the platform, and how decisions are made relating to content removal. Good practice could be more easily identified and adopted by other platforms. And qualitative reporting requirements on steps taken to improve processes would encourage platforms to make better and more consistent decisions, rather than simply remove more content more quickly.

We therefore recommend that:

- Any transparency reporting **should focus on qualitative reporting, and require platforms to set out what they are doing to tackle specified unlawful and harmful forms of content**; what further steps they are planning to take; what opportunities there are for people to report unlawful and harmful content; what process is undertaken to determine whether content is unlawful or harmful; and what opportunities there are to challenge decisions.