Joint Committee on Human Rights: Democracy, Free Speech and Freedom of Association
Global Partners Digital Submission

GLOBAL PARTNERS DIGITAL
May 2019

About Global Partners Digital

The advent of the internet – and the wider digital environment – has enabled new forms of free expression, organisation and association, provided unprecedented access to information and ideas, and catalysed rapid economic and social development. It has also facilitated new forms of repression and violation of human rights, and intensified existing inequalities. 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.

In this submission, we focus on the third set of questions asked by the Committee, "What is the role of social media in relation to free speech and threats to MPs?" and "How, if at all, should it be regulated?". We have analysed and answered these questions on the basis of the international human rights framework and, where relevant, the European human rights framework. In doing so, we also touch upon the first two sets of questions which look at the relationship between the right to freedom of expression and other human rights. We also review the proposals in the government's Online Harms White Paper given its relevance to the Committee's inquiry.

Question 3(i): What is the role of social media in relation to free speech and threats to MPs?

In answering this question, we look at the responsibilities that social media companies have under the international human rights framework rather than any *moral* obligations. As such, the starting point for determining the role that social media companies have when it comes to issues of free speech and threats to MPs is the United Nations Guiding Principles on Business and Human Rights (the UNGPs) for they set out the agreed international framework when it comes to human rights and the private sector.

The foundational principle of the UNGPs when it comes to businesses' responsibilities is that they should "respect human rights" (Principle 11) and, more specifically, should "avoid causing or contributing to adverse human rights impacts through their own activities, and address

such impacts when they occur" and "seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts" (Principle 13).

While soft law, the most comprehensive piece of guidance that looks at the role and responsibilities of social media companies when it comes to human rights is Recommendation CM/Rec(2018)2 of the Council of Europe's Committee of Ministers to member States on the roles and responsibilities of internet intermediaries (Recommendation CM/Rec(2018)2).[1] Recommendation CM/Rec(2018)2 highlights the responsibilities of social media companies to "respect the internationally recognised human rights and fundamental freedoms of their users and of other parties who are affected by their activities", noting that this responsibility "exists independently of the States' ability or willingness to fulfil their own human rights obligations."[2]

When it comes to the right to freedom of expression under Article 19 of the International Covenant on Civil and Political Rights (ICCPR),[3] this responsibility is relatively straightforward to set out in principle: it means, simply, that social media companies have a responsibility to respect freedom of expression and to avoid infringing upon that right as exercised by their users by censoring or otherwise moderating content inappropriately.

When it comes to threats to MPs, however, the responsibility is more complex to set out. The international human rights framework does not recognise an individual right to be free from abuse or threats, per se.[4] However, the right to security, under Article 9(1) of the ICCPR, has been interpreted to include protection from "intentional infliction of bodily or mental injury".[5] In relation to MPs (as well as those running for election) specifically, there is also Article 25 of the ICCPR, the scope of which covers public affairs, elections, and public service.[6] Article 25, paragraph (b) guarantees, the right of all citizens "[t]o vote and to be elected at genuine periodic elections which shall be by universal and equal suffrage and shall be held by secret ballot, guaranteeing the free expression of the will of the electors".

In its General Comment No. 25 on Article 25, the UN Human Rights Committee has stated that "[v]oters should be able to form opinions independently, free of violence or threat of violence,

---

[1] Council of Europe, Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries, 7 March 2018.

[2] *Ibid*., Para 2.1.1.

[3] The right to freedom of expression also exists within the European human rights framework in largely similar terms via Article 10 of the European Convention on Human Rights (ECHR).

[4] While Article 19(3) of the ICCPR and Article 10(2) of the ECHR both provide for permissible restrictions on the right to freedom of expression, these only set out the circumstances where freedom of expression *can* be restricted without it amounting to a breach of that right, not the circumstances where freedom of expression *should* be restricted. While restrictions on freedom of expression which relate to threats of violence, for example, may therefore be *justified*, they are not *required*.

[5] UN Human Rights Committee, General Comment No. 35: Article 9 (Liberty and security of person), UN Doc. CCPR/C/GC/25, 16 December 2014, Para 9. Article 5(1) of the European Convention on Human Rights also protects the right to liberty and security, although cases involving interferences with bodily and mental integrity have tended to be dealt with under Article 8's protection of the right to respect for private life.

[6] At the European level, there is no equivalent right, although Article 3 of Protocol 1 to the European Convention on Human Rights (which places an obligations upon ratifying states "to hold free elections at reasonable intervals by secret ballot, under conditions which will ensure the free expression of the opinion of the people in the choice of the legislature") has been interpreted as including in its scope the right to stand for elections, although it has not ruled on what requirements Article 3 includes, if any, to restrict threats of violence or other abuse against candidates running for election.

compulsion, inducement or manipulative interference of any kind."[7] While the General Comment does not specifically address violence or threats against those running for election, there is no reason why the same considerations should not apply. Using such an interpretation, it can be argued with a degree of confidence that the right to be elected necessitates candidates' freedom from violence or threat of violence, or manipulative interference of any kind.

States are the primary duty-bearers under international human rights law, and so it falls to state actors to take the lead in ensuring that these rights are protected and respected. This means, at a minimum, setting out clear definitions of what conduct is prohibited under legislation, whether criminal or civil, and enforcing that legislation. However, in its recent Scoping Paper on Abusive and Offensive Online Communications which looked at many of the forms of prohibited conduct which are relevant to this Committee's inquiry, the Law Commission stated that there is "considerable scope to improve the criminal law in this area"[8] and that many of the criminal provisions in this area were unclear, ambiguous or overly broad.

Drawing a line between speech which is protected by the right to freedom of expression on the one hand, and speech which amounts to violence (or threats of violence) is extremely difficult. When such speech relates to political candidates, this makes this area even more challenging. The UN Human Rights Committee has made clear that "[t]he free communication of information and ideas about public and political issues between citizens, candidates and elected representatives is essential" and that "in circumstances of public debate concerning public figures in the political domain and public institutions, the value placed by the [ICCPR] upon uninhibited expression is particularly high". As such, "the mere fact that forms of expression are considered to be insulting to a public figure is not sufficient to justify the imposition of penalties".[9]

It is therefore even more critical that governments ensure both that legislation provides sufficient clarity over when speech is prohibited, and that legislation does not require or incentivise restrictions on speech which is protected by the right to freedom of expression.

Companies, including social media companies, are not the primary duty-bearers, but do nonetheless have a responsibility, although not a legal obligation, to address abuse and threats of violence directed toward those running for election which is facilitated by their platforms. In doing so, however, they should be particularly cautious before removing or otherwise moderating content relating to "information and ideas about public and political issues", including "public figures in the political domain". However, where the platform is being used to facilitate threats of violence or intimidation against those running for election, and where that high threshold is unambiguously crossed (as opposed to cases involving robust criticism or insult), social media companies can be reasonably expected to take steps to ensure that such speech can be reported and removed. It is worth noting, of course, that it is only on a very small number of platforms where threats of violence or intimidation against those running for election take place.

The means by which social media companies can reasonably be expected to meet that responsibility will vary depending on their size, audience, and the types of content that their

---

[7] UN Human Rights Committee, *General Comment No. 25: Article 25*, UN Doc. CCPR/C/21/Rev.1/Add.7, 27 August 1996, Para 19.

[8] Law Commission, *Abusive and Offensive Online Communications: A Scoping Report*, Law Com. No. 381, 1 November 2018, Para 13.11.

[9] At the European level, the European Court of Human Rights has said that the right to freedom of expression "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". See *Handyside* v *the United Kingdom*, Application No. 5493/72, 7 December 1976 (European Court of Human Rights).

platform facilitates. Given the wide range of social media companies, and the fact that the particular issue is one that faced by only a small number of platforms, we do not think it sensible to prescribe or suggest specific actions which all companies should be considered as having to take to fulfil their responsibilities set out above. However, we do think that there are sufficient commonalities among social media companies that facilitate the generation and sharing of content, and where this issue exists, for the following to be general ways by which their responsibilities can be fulfilled:

a. Social media companies should clearly set out their content moderation policies and with a sufficient degree of specificity that allows users to know what content is and is not permitted;
b. Those content moderation policies should include detail on content which might amount to threats of violence, abuse or intimidation against other users or individuals;
c. Affected individuals, whether users or otherwise, should be able to report content which they believe breaches those content moderation policies in a simple and straightforward manner, and for a decision to be made as to what action, if any, will be taken within a reasonable period of time. It should be possible for affected users to challenge these decisions; and
d. Where a report is made which alleges threats of violence or intimidation against those running for election, such reports should be considered more urgently. As a precautionary measure, content to which such reports relate (unless very clear permitted by their content moderation policies) could be temporarily removed in the first instance, pending a final decision.

Question 3(ii): How, if at all, should it be regulated?

States have a duty to ensure that the human rights of those within their jurisdiction are respected and protected, including by third parties (such as businesses). The UNGPs make clear that this includes establishing a legal and policy framework which enables and supports businesses to respect human rights. Principle 3, in particular, sets out the general obligations on states on this point:

> "3. In meeting their duty to protect, States should:
>
> (a) Enforce laws that are aimed at, or have the effect of, requiring business enterprises to respect human rights, and periodically to assess the adequacy of such laws and address any gaps;
> (b) Ensure that other laws and policies governing the creation and ongoing operation of business enterprises, such as corporate law, do not constrain but enable business respect for human rights;
> (c) Provide effective guidance to business enterprises on how to respect human rights throughout their operations;
> (d) Encourage, and where appropriate require, business enterprises to communicate how they address their human rights impacts."

Given the impact that social media companies have upon the enjoyment and exercise of the rights to freedom of expression and to free and fair elections, the state does, therefore, have an obligation to ensure that these rights are respected by such companies. Critically, though, any legal and policy framework established should not constrain social media companies' ability to respect human rights themselves, nor should it directly or indirectly constitute a restriction on

the enjoyment and exercise of the human rights that use those platforms. While the framework may include certain forms of regulation, it is important for non-regulatory responses to be pursued as well.

The government's Online Harms White Paper (the White Paper) contains a set of proposals to regulate social media companies, as well as many other actors, in order to tackle a range of harms which are experienced online, including threats to MPs. There are, unfortunately, a number of serious risks to freedom of expression stemming from the proposals, and areas where they are clearly incompatible and inconsistent with the government's international and European human rights obligations. These relate, in particular, to the scope and definitions of the harms included; the specific model proposed; and the lack of safeguards in relation to freedom of expression.

Scope and definitions of harm

While the scope of this committee's inquiry is threats to MPs, it is important to highlight the extremely broad scope of the "harms" that the White Paper seeks to address. The core element of the model proposed in the White Paper (which we look at later in this submission) is a statutory duty of care on companies which provide online platforms "to take reasonable steps to keep users safe, and prevent other persons coming to harm as a direct consequence of activity on their services".[10] In practice, this will mean the removal or moderation of online content which is either illegal or legal but "harmful". The White Paper sets out a list of forms of examples of illegal content or behaviour which would be within scope, such as harassment and the incitement of violence. It also sets out a list of forms of content or behaviour which are not generally illegal, but considered "harmful" and "with a less clear definition", including cyberbullying, trolling and disinformation.

International and European human rights law is clear that restrictions on freedom of expression must meet the test of legality, which includes clarity and sufficient precision. As Recommendation CM/Rec(2018)2 states, "[a]ny legislation applicable to internet intermediaries and to their relations with States and users should be accessible and foreseeable. All laws should be clear and sufficiently precise to enable intermediaries, users and affected parties to regulate their conduct."[11] This is particularly important when it comes to third parties who will be involved in assessing whether content is illegal or not, given that they are unlikely to have legal expertise. The broad range of harms identified in the White Paper, however, falls far short of this requirement.

With respect to the illegal harms identified, the White Paper includes a number that are relevant to this inquiry, particularly harassment, hate crime and the incitement of violence, all of which are criminal offences. The Law Commission's report in this area highlighted a number of aspects of these, and other relevant criminal offences where there is a lack of clarity, ambiguity, potential impacts upon freedom of expression. In relation to communications offences under the Malicious Communications Act 1988 and Communications Act 2003, for example, the Law Commission concluded that "elements of these offences are broadly defined and rather ambiguous on some of the proscribed speech, with much left to the courts and prosecutorial discretion".[12] A number of specific terms used in different criminal offences were singled out, including "grossly offensive" (described as "vague and unclear")[13] and "indecent"

---

[10] HM Government, *Online Harms White Paper*, April 2019, p. 42, available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf.

[11] See above, note 1, Para 1.2.1.

[12] See above, note 8, Para 4.154.

[13] *Ibid*., Para 5.95.

(with reference to its "vagueness").[14] The relevant CPS guidance on social media makes clear that many communications which technically meet the conditions set out in legislation should not be prosecuted, as prosecution would constitute an unjustified interference with the right to freedom of expression.[15]

The report also highlighted the difficulties of applying other criminal provisions to online speech, such as offences of assault (which require a person to apprehend immediate violence, without any caveat or conditionality)[16] and harassment (which requires a "course of conduct comprising at least two instances of offending behaviour).[17]

"Legal but harmful" forms of content are not defined at all in the White Paper. There are no legal definitions to turn to because these "harms" are, indeed, legal.

If there are no definitions provided of what specific content is harmful, and the legal definitions of different forms of illegal content are vague, unclear, or include speech which is protected by the right to freedom of expression, then it is extremely difficult (if not impossible) for companies to know which content they should remove or moderate. Even if only the illegal forms of content were removed by companies, this would still amount to restrictions on the right to freedom of expression because, as is acknowledged by the Law Commission and the CPS, the criminal offences do indeed include speech which is protected, and that prosecutions should not take place in such circumstances. This would 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. It would further be inconsistent with the government's stated intention of ensuring that the law applies equally online as offline.

The model proposed

The regulatory model proposed - a duty of care, enforced by a regulatory body with the power to develop binding Codes of Practice - would require (or incentivise) online platforms to remove content which is potentially harmful. When it comes to regulation, Recommendation CM/Rec(2018)2 provides that:

> "Any request, demand or other action by public authorities addressed to internet intermediaries to restrict access (including blocking or removal of content), or any other measure that could lead to a restriction of the right to freedom of expression, shall be prescribed by law, pursue one of the legitimate aims foreseen in Article 10 of the [European Convention on Human Rights], be necessary in a democratic society and be proportionate to the aim pursued. State authorities should carefully evaluate the possible impact, including unintended, of any restrictions before and after applying them, while seeking to apply the least intrusive measure necessary to meet the policy objective."[18]

However, the nature of this model, as set out in the White Paper, poses serious risks to freedom of expression online by incentivising the removal of content which is protected by that right. The model is neither proportionate, nor the least intrusive measure that could achieve the policy objective.

---

[14] *Ibid*., Para 6.86.

[15] Crown Prosecution Service, Social Media - Guidelines on prosecuting cases involving communications sent via social media, Para 28, available at: https://www.cps.gov.uk/legal-guidance/social-media-guidelines-prosecuting-cases-involving-communications-sent-social-media.

[16] See above, note 8., Para 7.26.

[17] *Ibid*., Para 7.69.

[18] See above, note 1, Para 1.31.

(i) A preventative approach

The duty of care model would require social media companies to take reasonable steps to protect users from harm. The White Paper suggests that this would mean going beyond simply ensuring that users are able to report content which is harmful so that it can be removed, but also taking steps to prevent users from coming across that harmful content in the first place.19 As such, the model proposed implies a preventative approach (sometimes referred to as "prior restraint"), rather than a reactive one.

There are two main ways that online platforms could, in theory, prevent users from coming across harmful content which would be consistent with this preventative approach. The first is to prevent it from every being made available on the platforms through checking all content beforehand (in practice, through machines); the second is to proactively and continuously monitor all content on the platform and removing harmful content as soon as it is identified with the hope that it will not have been seen.

Were the equivalent measures proposed in the offline world, they would be terrifying and unquestionably violations of the right to freedom of expression. The first is equivalent to requiring all individuals in the UK to have what they would like to say approved before they can say it, in case they wish to say something harmful. The second is equivalent to having everything anyone in the UK says monitored in case it is harmful. Such proposals would, without question, be considered disproportionate ways of addressing illegal and harmful speech. This should be no less true simply because they are being proposed in relation to what is said online, rather than offline. Indeed, Recommendation CM/Rec(2018)2 is clear that, "[s]tate authorities should not directly or indirectly impose a general obligation on intermediaries to monitor content which they merely give access to, or which they transmit or store, be it by automated means or not."[20]

(ii) Broad categories and unclear definitions

As noted above, the definitions of many of the illegal forms of content are unclear or overly broad, and the "legal but harmful" forms of content are not defined at all, (although the White Paper does suggest that these may be defined through Codes of Practice developed by the regulatory body). Many of the particular forms of illegal content with which the Committee is concerned are either vaguely defined, defined overly broadly and include speech which is protected by the right to freedom of expression, or difficult to apply to online speech.

Without clear, legal definitions, and particularly given the serious potential sanctions for non-compliance with the duty of care, social media companies will be incentivised to interpret the terms broadly, rather than risk sanction, and therefore remove an even broader range of content than is intended, including content which is protected by the right to freedom of expression.

(iii) Automated processes

Given the scale of content which is generated and shared on online platforms, it would be impossible for human moderators either to review all content before upload or proactively and continuously monitor it post upload. As such, it is inevitable that companies would have to turn to automated processes, such as artificial intelligence (AI), to meet their obligations under the duty of care, possibly by filtering illegal and harmful content prior to upload, or identifying

---

[19] See, for example, Para 3.3 of the White Paper which states that "This statutory duty of care will require companies to take reasonable steps to keep users safe, and prevent other persons coming to harm as a direct consequence of activity on their services."

[20] See above, note 1, Para 1.3.5.

and removing it once uploaded. Indeed, the White Paper repeatedly refers to the use of AI to tackle certain forms of illegal and harmful content. AI is, however, at a very nascent stage when it comes to analysing speech, and can only accurately identify a very small number of categories of speech which don't require an assessment of context or other nuances. As such, there are particular risks to freedom of expression which stem from the use of automated processes in order to determine whether content is illegal or harmful.

First, it is simply not possible to develop accurate automated processing to identify particular forms of content, if at all, if the definitions of those forms of content are not clear, as is the case with many of the forms of illegal content, and all of the forms of "legal but harmful content" set out in the White Paper. As such, automated processing will lead to inaccurate results and either the removal of legal and/or harmless content, or a failure to remove to illegal and harmful content.

Secondly, making a decision about whether a particular piece of content is illegal or harmful requires an understanding of the context; however, automated processes are unable to determine context (or factors such as sarcasm, satire or irony).[21] For example, it is impossible to know without context whether an online post which simply states "I'll see you in Shoreditch on Friday. Be ready!" is threatening violence, or simply a friend arranging to see another. An automated process could easily identify such a statement as a threat of violence and either remove it or prevent it from being uploaded at all. A video of violent and graphic war crimes could be terrorist propaganda or important evidence shared by human rights defenders. An automated process would not be able to tell the difference.

(iv) Time limits

The White Paper suggests that time limits may be imposed to remove content[22] and the government has been advocating for a one hour time limit for the removal of terrorist content at the European Union level. Recommendation CM/Rec(2018)2, however, states that processes should not be "designed in a manner that incentivises the take-down of legal content, for example due to inappropriately short timeframes".[23] The imposition of time limits incentivise rushed decisionmaking and stifles any ability to fully consider context or obtain the necessary information and expertise in order to make an accurate determination. If the time limits don't allow the necessary contextual information to be obtained, and there is a risk of sanctions, the platform is likely just to remove the content, even though it may in fact be neither illegal nor harmful.

(v) Penalties and sanctions

The White Paper proposes heavy sanctions for non-compliance with the duty of care, including high fines, and consults on others, including criminal liability for individuals working at the social media companies, and blocking platforms to UK users. These heavy sanctions skew incentives and exacerbate the risks outlined above. If a social media company is making decisions as to whether to remove content or not on the basis that it might potentially be illegal or harmful, there will be a strong incentive to 'play it safe' and simply remove the content rather than risk a sanction. Noting this risk, Recommendation CM/Rec(2018)2 states that "[s]tate authorities should ensure that the sanctions they impose on intermediaries for non-compliance with regulatory frameworks are proportionate because disproportionate

[21] 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. Recommendation CM/Rec(2018)2 also highlights the fact that "automated means, which may be used to identify illegal content, currently have a limited ability to assess context".

[22] See, above, note 10, p. 42.

[23] See above, note 1, Para 1.3.7.

sanctions are likely to lead to the restriction of lawful content and to have a chilling effect on the right to freedom of expression."[24]

Evidence from the implementation of the Network Enforcement Act in Germany in 2018 suggests that this would likely be the case: since the introduction of the tight timelines and heavy fines included in the law (48 hours in the case of "manifestly unlawful" content), there have been a number of instances of social media companies such as Twitter and Facebook, for example, removing pieces of content which were controversial, satirical and ironic, but not obviously illegal or even harmful.25

(vi) A lack of transparency and accountability

There are also concerns at a more principled level over how decisions to remove online content should be made. 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 illegal content, the White Paper's proposals to require social media companies to decide on whether content is illegal shift judicial and quasi-judicial functions to those companies, or their nominees. In the context of the abuse of MPs, this would include determinations on whether certain content constituted a criminal offence, such as threats of violence or harassment. However, the White Paper makes no proposals to guarantee that there would be mechanisms for accountability or safeguards in place, as there are when equivalent decisions are made by public authorities or the judiciary.

When it comes to "legal but harmful" content, there are similar concerns over whether companies are well-placed and able to make determinations as to what content is harmful, particularly if no clear, precise definitions are provided. The sheer scale of content means that in person reviews are unlikely to be feasible, and we have highlighted above how automated processes are poor at making decisions at identifying this kind of content. As with unlawful content being removed, there would not necessarily be any mechanisms for accountability nor safeguards in place to challenge decisions.

Given this, we believe there is a critical role for transparency when it comes to any regulation. 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 the quality of decisionmaking is ensured. Mandatory transparency reporting requirements would encourage companies 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 companies. Qualitative reporting requirements on steps taken to improve processes would encourage companies to make better and more consistent decisions, rather than simply remove more content and more quickly.

---

[24] *Ibid*.

[25] 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, and Kinstler, L., "Germany's Attempt to Fix Facebook Is Backfiring", The Atlantic, 18 May 2018, available at: https://www.theatlantic.com/international/archive/2018/05/germany-facebook-afd/560435/.

Safeguarding freedom of expression

Given the risks to freedom of expression set out above, it is critical that safeguards are fully considered and integrated into any regulation developed, and there are a number of ways that this can be done.

First, as recommended by CM/Rec(2018)2), it should be made clear in legislation that, as is the case now, social media companies cannot be held liable for third-party content which they merely give access to or which they transmit or store, save where they do not act expeditiously to restrict access to content or services as soon as they become aware of their illegal nature.

Second, any statutory provisions which set out requirements on social media should make clear that compliance is focused on ensuring that they have appropriate terms of service and content moderation policies that deal with harmful content, and that these are enforced effectively, consistently, and with decisions made as soon as is reasonably practicable. They should also make clear that compliance does not require any forms of review of content prior to upload, nor any form of proactive or continuous monitoring of content.

Third, any statutory provisions should explicitly state that the importance of protecting and respecting the right to freedom of expression is to be taken into account when social media companies make decisions and when compliance with any duties is being assessed.

Fourth, the Codes of Practice developed for the purposes of ensuring compliance with any duties should also include a section on the importance of protecting and respecting the right to freedom of expression. The Equality and Human Rights Commission should be involved in the development of any guidance.

Firth, social media companies should be required 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.

Sixth, the Equality and Human Rights Commission should be involved in the establishment of any new regulatory body, in the enforcement of its duties and functions, and be given a role of reviewing the overall process to determine impacts upon freedom of expression. The government should proceed with caution in this novel area of policy and focus first on ensuring that any new regulatory body has the sufficient skills and understanding before they start exercising powers.

Seventh, any statutory provisions establishing a regulatory body should explicitly state that protecting and respecting the right to freedom of expression is one of its statutory duties and ensure that a human rights framework is embedded in its decisionmaking, including in its development of Codes of Practice.

Eighth, any regulatory body established, and the decisions it makes, should be open to judicial review. As such, the body should be designated as 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.

Ninth, any regulatory body established should be fully independent from government and political direction from ministers. It should be multistakeholder in nature, and include all relevant stakeholders including social media companies, academia and civil society.

Tenth, when it comes to enforcement, we believe that such a graduated approach, which focuses on 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 sanction be open to the regulatory body. For example:

a. As a first step, social media companies 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 should be developed by the regulatory body in consultation with other stakeholders. Where the regulatory body considers that the company has fully complied with its transparency reporting requirements, and that they demonstrate sufficient action is being taken, the company 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;
b. 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;
c. Only as a third and final step should a company be subject to sanctions for failure to comply with any duties relating to content removal.

Eleventh, transparency reporting requirements should focus on qualitative reporting, and require social media companies 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.

Risks of copycat legislation being adopted elsewhere

Finally, the internet is global in nature, and we wish to highlight the international dimension to this issue. There has been a recent trend of states passing copycat legislation relating to the internet, including that regulating online content, 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.[26] 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. This would be hugely damaging for the UK's reputation as a strong proponent of a free, open and secure internet.

[26] Reporters Without Borders, "Russian bill is copy-and-paste of Germany's hate speech law", *rsf.org*, 19 July 2017, available at: https://rsf.org/en/news/russian-bill-copy-and-paste-germanys-hate-speech-law.