Online Harms White Paper Consultation
Global Partners Digital Submission

GLOBAL PARTNERS DIGITAL
July 2019

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 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.

Table of Contents

Introduction

We welcome the consultation on the Online Harms White Paper (the "White Paper"), the first attempt made by a government to address all forms of online illegal and harmful content through a single regulatory framework. It is disappointing, however, that while the White Paper asks a number of questions about the proposals, it does not in fact consult on many of the most critical components, such as the scope of harms being considered and the regulatory model being proposed, namely a statutory duty of care enforced by a regulatory body. These proposals were not suggested in the government's earlier Internet Safety Strategy Green Paper, meaning that there has been no formal opportunity to input into and shape the specific regulatory model which is now being proposed.

We have significant concerns over the scope of harms included as well as the model being proposed, and the risks that they would pose to the rights to freedom of expression and privacy. Based on our analysis, we believe that the proposals, if taken forward in their current state, would likely put the UK in breach of its obligations under both international human rights law and the European Convention on Human Rights (ECHR), as incorporated into domestic law through the Human Rights Act 1998 (HRA 1998).

While we respond to the relevant questions posed in the White Paper consultation and make a series of recommendations on how the proposals should be refined, these refinements alone would still not be sufficient to ensure that the proposals as a whole did not seriously put these rights at risk. We therefore take this opportunity to set out these broader concerns as well, through a full human rights analysis of the proposals, and make further specific recommendations on how the proposals should be revised in order to mitigate those risks as far as possible, including through the incorporation of further safeguards for freedom of expression and privacy. These recommendations are also set out in full at the Annex to this response.

Framework for analysis of the proposals and consultation questions

Our analysis of the proposals in the White Paper and the consultation questions asked is based on international human rights law, the ECHR and the HRA 1998. The most relevant human rights impacted by the proposals in the White Paper are the rights to freedom of expression and to privacy. This is recognised by the government in the White Paper itself, where it states that "[o]ur vision is for (...) freedom of expression online" and that its intention is "to help companies ensure safety of users while protecting freedom of expression" (p. 7). The White Paper also recognises "the importance of privacy" (pp. 49 and 50) and that the government "takes both the protection of personal data and the right to privacy extremely seriously" (p. 26).

As is well-established under international human rights law, the ECHR and the HRA 1998, any measures which interfere either the right to expression (which includes the ability of individuals to seek, receive or impart certain forms of expression online) or the right to privacy (which includes the ability to communicate privately) will amount to a breach of those

rights unless they can be justified.[1] The situation is similar under the European Union law.[2] In order to be justified, any restriction must meet a three-part test, namely that (i) there is a clear legal basis for the restriction, (ii) it pursues a legitimate aim, and (iii) it is necessary and proportionate to achieve that aim.

It is important to remember that the UK's obligation to ensure that these rights are not unjustifiably restricted exists both in relation to restrictions which stem from the actions of the state itself as well as those caused by third parties, such as private companies. As such, it makes no difference from the perspective of the individual affected whether any restrictions are imposed and enforced directly by the state (e.g. through creating criminal offences which are enforced by the police and the courts) or through third parties, particularly when the third party is acting in order to comply with legal obligations.

With respect to the actions of private companies specifically, the United Nations Guiding Principles on Business and Human Rights (UNGPs) makes clear that a state's international human rights obligations include establishing a legal and policy framework which enables and supports businesses to respect human rights. Principle 3 notes that this general obligation includes ensuring "that (...) 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".

Given the impact that online platforms have upon the enjoyment and exercise of the rights to freedom of expression and privacy, the government has a clear obligation to ensure that these rights are respected by these platforms.[3] This includes ensuring that legislation and other measures do not constrain online platforms' ability to respect the right to freedom of expression or privacy themselves, nor should they directly or indirectly constitute a restriction on the enjoyment and exercise of those rights by those that use those platforms.

Our analysis of the regulatory measures proposed in the White Paper and our subsequent recommendations are based on these frameworks. Given the limited existing interpretation and case-law of these frameworks as they apply to measures comparable to those proposed in the White Paper, we also make reference, as appropriate, to 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),[4] and relevant commentary from the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (the UN Special Rapporteur). These guidelines and commentaries provide detail on the obligations of states with respect to the protection and promotion of human rights in the digital environment, with a particular focus on any legal frameworks that apply to internet intermediaries.

---

[1] See, in particular, Articles 17 and 19 of the International Covenant on Civil and Political Rights, and Articles 8 and 10 of the European Convention on Human Rights (ECHR). The rights to freedom of expression and privacy are also protected in other treaties, such as Articles 13 and 16 of the Convention on the Rights of the Child. Under section 6 of the Human Rights Act 1998, it is, of course, unlawful for a public authority, including a government department, to act in a way which is incompatible with a right protected by the ECHR.

[2] See, in particular, Articles 7, 11 and 52 of the Charter of Fundamental Rights of the European Union (the Charter). The Charter is of particular importance since it is directly binding upon EU member states, including the United Kingdom, when implementing EU law. Given that any legislation in this field is likely to impact upon the United Kingdom's implementation of the E-Commerce Directive (Directive 2000/31/EC), it must be compatible with Articles 7, 11 and 52 of the Charter.

[3] The Online Harms White Paper refers to those companies within its scope as "companies that provide services or tools that allow, enable or facilitate users to share or discover user-generated content, or interact with each other online". This consultation responses uses the term "online platforms" as shorthand.

[4] 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.

Though not a framework for the purpose of our analysis, we note that the UK has, through its membership of the Freedom Online Coalition, signed up to a number of commitments which are relevant to the subject. These includes commitments made in the "Recommendations for Freedom Online, Adopted in Tallinn, Estonia on April 28, 2014 by Ministers of the Freedom Online Coalition":

> "We, the members of the Freedom Online Coalition
>
> 4. Dedicate ourselves, in conducting our own activities, to respect our human rights obligations, as well as the principles of the rule of law, legitimate purpose, non-arbitrariness, effective oversight, and transparency, and call upon others to do the same,
>
> (…)
>
> 6. Call upon governments worldwide to promote transparency and independent, effective domestic oversight related to electronic surveillance, use of content take-down notices, limitations or restrictions on online content or user access and other similar measures, while committing ourselves to do the same".[5]

More recent commitments were made in the Freedom Online Coalition's "Joint Statement on Internet Censorship":

> "In 2017, the world witnessed state-sponsored Internet censorship in various forms: states have manipulated and suppressed online expression protected by international law, have subjected users to arbitrary or unlawful surveillance, have used liability laws to force ICT companies to self-censor expression protected by international law, have disrupted networks to deny users access to information, and have employed elaborate technical measures to maintain their online censorship capabilities. Further unlawful efforts included state censorship in private messaging apps and systematic bans of news websites and social media. Likewise certain states have introduced or implemented laws which permit executive authorities to limit content, on the Internet broadly and without appropriate procedural safeguards. Individuals who may face multiple and intersecting forms of discrimination, including women and girls, often faced disproportionate levels of censorship and punishment.
>
> (…)
>
> he FOC firmly believes in the value of free and informed political debate, offline and online, and its positive effects on long term political stability. The Coalition calls on governments, the private sector, international organizations, civil society, and Internet stakeholders to work together toward a shared approach - firmly grounded in respect for international human rights law - that aims to evaluate, respond to, and if necessary, remedy state-sponsored efforts to restrict, moderate, or manipulate

[5] Recommendations for Freedom Online, Adopted in Tallinn, Estonia on April 28, 2014 by Ministers of the Freedom Online Coalition, available at: https://www.freedomonlinecoalition.com/wp-content/uploads/2014/04/FOC-recommendations-consensus.pdf.

> online content, and that calls for greater transparency of private Internet companies' mediation, automation, and remedial policies."[6]

Human rights analysis of the Online Harms White Paper

Our analysis of the White Paper looks at four aspects of the proposals: the harms within scope (Chapter 2), the regulatory model (Chapters 3 and 4), the means of enforcement (Chapter 5, 6 and 7), and the safeguards for freedom of expression and privacy. Although we comment on each of those four aspects separately, the overall impact upon the rights to freedom of expression and privacy necessitates an analysis of their cumulative impact, which we undertake as the fifth part of this section.

1. The scope of harms covered

Chapter 2 of the White Paper sets out an "initial list of online harmful content or activity" (p. 31) which are in scope, however it also notes that this list is "neither exhaustive nor fixed" (p. 30). Our analysis of the scope of harms covered is limited to those which are listed in the White Paper as we are not in a position to assess further potential types of harmful content or activity which are not mentioned.

As we note in our analysis of the duty of care below, the White Paper makes clear that, in order to fulfil that duty, online platforms will be expected to take steps to remove, restrict, or otherwise moderate content and activity which is "harmful". While the White Paper does state that the regulatory approach "will impose more specific and stringent requirements for those harms which are clearly illegal, than for those harms which may be legal but harmful, depending on the context", (p. 42) nothing in the White Paper excludes the possibility that the duty of care will nonetheless require the removal, restriction, or moderation of content which is "legal but harmful". We therefore undertake our analysis of the scope of harms covered on the assumption that the duty of care will require online content or activity which is "harmful", whether illegal or legal, to be restricted in some way.

Harms with a clear definition

Where content is going to be restricted on the basis that it is "harmful", it must, to meet the first limb of the three-part test, be restricted "by law". This means, in the context of Article 19 of the International Covenant on Civil and Political Rights, for example, that the restriction:

> "(...) must be formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly and it must be made accessible to the public. A law may not confer unfettered discretion for the restriction of freedom of expression on those charged with its execution. Laws must provide sufficient guidance to those charged with their execution to enable them to ascertain what sorts of expression are properly restricted and what sorts are not."[7]

Similarly, as Recommendation CM/Rec(2018)2 states:

> "Any legislation applicable to internet intermediaries and to their relations with States and users should be accessible and foreseeable. All

[6] The Freedom Online Coalition, Joint Statement on Internet Censorship, available at: https://freedomonlinecoalition.com/wp-content/uploads/2018/05/FOC-Joint-Statement-on-Internet-Censorship-0518.pdf.
[7] UN Human Rights Committee, General Comment No.34, Article 19: Freedoms of opinion and expression, UN Doc. CCPR/C/GC/34, 12 September 2011, Para 25.

> laws should be clear and sufficiently precise to enable intermediaries, users and affected parties to regulate their conduct."[8]

With respect to those forms of harmful content which have a "clear definition", (i.e. those in the first column of Table 1 on p. 31) and which constitute various criminal offences, these requirements are largely met, at least when they are being enforced by state actors. However, in its recent Scoping Report on Abusive and Offensive Online Communications which looked at many of the criminal forms of conduct which are relevant to the White Paper, the Law Commission stated that there is "considerable scope to improve the criminal law in this area"[9] and that many of the criminal provisions in this area - such as those relating to harassment and disclosing private sexual photographs or films ("revenge pornography") - were unclear, ambiguous or overly complex. Despite the assertion of the White Paper, therefore, not all of these forms of content and activity do, in fact, have "clear" definitions.

If there is uncertainty over the extent to which particular forms of content or activity are prohibited among those tasked with their enforcement, it is difficult to see how they could be considered as either "clear" or "sufficiently precise" for the purposes of the first limb of the three-part test. They should not, therefore, be included in the Online Harms Bill until such uncertainty has been removed.

Recommendation 1: To ensure that the requirements of legal clarity and precision are met, the Online Harms Bill should not include those forms of harmful content and activity which are criminal offences highlighted by the Law Commission as needing review. Such forms of content and activity should only be included in the Bill (or, if it has passed into law, added to the Act) once the Law Commission has completed its review of communications offences and the relevant criminal offences subsequently amended in line with any recommendations.

Even where the definitions are clear, it is important to note that not all forms of content and activity which might technically be illegal actually result in prosecution or other form of enforcement. This is because many criminal offences are deliberately broad and provide the police and the Crown Prosecution Service (CPS) with discretion in deciding whether or not to charge or prosecute an individual. The Law Commission's Scoping Report on Abusive and Offensive Online Communications highlighted a number of forms of prohibited conduct which are relevant to the White Paper as including speech which is protected by the right to freedom of expression, and where prosecution would constitute an unjustified interference with this right. These criminal offences included offences of harassment, stirring up hatred, and other hate crimes.[10]

Indeed, consideration of whether a prosecution would constitute an unjustified restriction on the right to freedom of expression is one factor considered by the CPS in the exercise of its discretion. As the CPS Guidelines on prosecuting cases involving communications sent via social media make clear, prosecutions should only take place when "there is sufficient evidence that the communication in question (...) crosses the high threshold necessary to protect freedom of expression, even unwelcome freedom of expression".[11] This means, for example, that speech which is technically illegal, will not be prosecuted where it is simply:

- Offensive, shocking or disturbing;
- Satirical, iconoclastic or rude;

[8] See above, note 4, Para 1.2.1.
[9] Law Commission, Abusive and Offensive Online Communications: A Scoping Report, Law Com. No. 381, 1 November 2018, Para 13.11.
[10] See, for example, *ibid*., Paras 8.131, 8.187 and 9.20.
[11] Crown Prosecution Service, Social Media - Guidelines on prosecuting cases involving communications sent via social media, Para 28.

- The expression of unpopular or unfashionable opinion about serious or trivial matters, or banter or humour, even if distasteful to some or painful to those subjected to it; or
- An uninhibited and ill thought out contribution to a casual conversation where participants expect a certain amount of repartee or "give and take".[12]

There is therefore a clear recognition that some forms of speech and activity which are technically criminal offences with clear definitions are nonetheless protected by the right to freedom of expression and should not be prosecuted since to do so would amount to an unjustified restriction on that right.

Further protection of the right to freedom of expression in the enforcement of criminal offences is provided by the fact that the police, the CPS and the courts are all public authorities for the purposes of the HRA 1998. As such, it is therefore unlawful for them to act in a way which is inconsistent with the right to freedom of expression, and individuals can bring proceedings under the HRA 1998 to a court to challenge that action. These bodies, and the individuals that work for them, are also provided with resources, guidance and training on human rights and how to exercise their functions in a way which is consistent with those rights.

There is a fundamental concern in principle with tasking private companies, in this case online platforms, with the role of enforcing the criminal law. While these companies may not be prosecuting individuals, or making determinations of guilt, it is hard to see how the duty of care proposed in the White Paper could be undertaken without these companies having to make decisions as to whether particular forms of content are legal or illegal, and then taking steps to restrict content if it is illegal. They are therefore undertaking a role which has, until now, been reserved to public authorities, but without the requisite safeguards and mechanisms of accountability listed above.

The White Paper, however, proposes no equivalent safeguards or mechanisms of accountability which would ensure that online platforms are able (and required) to exercise similar discretion in their interpretation of harms with a clear legal definition in order to safeguard the right to freedom of expression. The proposals run contrary, therefore, to the assertion that "it is also important to make sure that criminal law applies online in the same was as it applies offline" (p. 34), as an entirely different regime for its enforcement is being proposed.

Recommendation 2: The Online Harms Bill should explicitly provide that all online platforms must take sufficient measures to safeguard the right to freedom of expression when complying with the duty of care and any other obligations under the Bill, such as complying with codes of practice. This would help ensure an equivalent level of protection to that provided by section 6 of the HRA 1998.

Recommendation 3: Before any statutory duty of care or other legal obligations come into force, the Online Harms Bill should require the regulatory body to produce detailed guidance to all online platforms on how to fulfil any obligations in a way which will fully protect the right to freedom of expression. Online platforms should be given sufficient time, once such guidance has been published, to adopt the necessary policies and processes which will allow them to do so, before any duty of care or other legal obligations will start to apply. This period of time should be no less than six months.

Recommendation 4: The guidance should be sufficiently detailed and provide specific guidance in relation to each of the particular forms of harms listed in the Online Harms Bill, and be tailored for different sizes of online platforms, and different types of online platforms.

[12] *Ibid*.

The guidance should be developed in collaboration with other stakeholders with relevant expertise, including civil society organisations and the Equality and Human Rights Commission. The guidance should be published in draft form and open to consultation, with sufficient time provided for feedback to be received and incorporated.

Recommendation 5: The Online Harms Bill should make clear that any online platform within scope has, at any time, the right to require further guidance from the regulatory body where the platform reasonably believes that fulfilment of the duty of care or any other legal obligations would undermine their ability to safeguard the right to freedom of expression.

Recommendation 6: The Online Harms Bill should also explicitly provide that in relation to any potential enforcement proceedings or action taken under the Bill, whether by the regulator or another body, an online platform is able to argue that, with respect to the alleged non-compliance, it was reasonably acting in accordance with its duty to take sufficient measures to safeguard the right to freedom of expression. Where such an argument is raised by an online platform, the regulator or other body enforcing the Bill should be required to reconsider its enforcement proceedings or action, and withdraw them if the online platform was in fact reasonably acting in accordance with its requirement.

Harms with a less clear definition

The second column of Table 1 on p. 31 of the White Paper accepts that the forms of harmful content and activity listed therein have "a less clear definition". However, as noted above, for any restriction on the right to freedom of expression to be justified, the first limb of the three-part test requires the restriction to be set out in law and be clear and precise. Without such a clear and precise definition, in law, restrictions on the right to freedom of expression based on these "harms with a less clear definition" will not meet that first limb of the test, and would not be justified.

In order to comply with the first limb of the test, the Online Harms Bill would therefore need to contain clear and precise definitions of these particular forms of harmful content and activity. However, this would not entirely solve the problem and, in fact, would raise further concerns. The government has repeatedly stressed that its intention is for equivalence between that which is prohibited offline and online.[13] Introducing new categories of "harmful" content and activity that potentially require removal, restriction of other form of moderation when they occur online, but not offline, would be wholly inconsistent with that intention. The White Paper does not propose, for example, that there should be a general prohibition - through the criminal law or otherwise - of "bullying" or "disinformation", however the duty of care will potentially require online platforms to take steps to remove, restrict or moderate such content and activity when it occurs via their platforms. The White Paper is therefore proposing the creation of two different standards of permissible expression depending on whether it occurs offline (e.g. in person, or via printed media or television) or online.

There are two ways this problem can be addressed. The first would be for these forms of "harmful content or activity" to be prohibited generally and clearly defined in the Online Harms Bill or other legislation. We would not, however, support such an approach. We also note that there has been some suggestion that the harms themselves would not be set out in the Bill, but left entirely to the discretion of the regulatory body to determine and define. This does not remedy the problem highlighted above and, indeed, would be even more problematic since it would delegate decisions about what forms of expression can and cannot be exercised online to an independent regulatory body, when such decisions should be taken by a democratically elected legislature. Further, any guidance from the regulatory body would not

[13] On p. 4 of the Internet Safety Strategy Green Paper, for example, the government set out as one of its underlying principles, "What is unacceptable offline should be unacceptable online."

have the status of "law" for the purpose of the first limb of the three-part test for permissible restrictions.

The second way the problem could be addressed is through making it explicitly clear in the Online Harms Bill that while the duty of care applies in relation to certain forms of "harmful content or activity" that are not illegal, complying with the duty does not require the removal, restriction or moderation of such content or activity, and that the duty can be fully complied with through other actions.

This second approach recognises that the list of "harms with a less clear definition" are legitimate public concerns, and that it is legitimate for the government to take steps to address them. It may also be the intention of the government that these harms be dealt with differently and in a way that does not involve the removal, restriction or moderation of content. For example, disinformation online which causes public harms could be addressed through better transparency over the sources of information. Images of self-harm online which could potentially encourage others to harm themselves could be preceded by warnings of the images, or accompanied by information on where a viewer could seek help if they were considering harming themselves. Such measures would not amount to restrictions on the right to freedom of expression, however it would still, of course, be essential for these forms of "harmful content or activity" to have clear and precise definitions in the Bill.

As noted above, though, it is not clear in the White Paper whether such a different approach is expected in relation to "harms with a less clear definition". Indeed, the suggestions of what might be in codes of practice on both of these examples include areas which indicate that certain forms of content would be removed or restricted.[14] It therefore appears to be the case that there is an expectation that, to some extent, forms of content and activity under the "harms with a less clear definition" category will be restricted as a result of the duty of care. If this is not the case, then this should be made clear in the Bill.

Recommendation 7: Unless the government proposes that the "harms with a less clear definition" be prohibited generally and clearly defined in the Online Harms Bill or other legislation (a proposal which we would not support), then, ideally, the Online Harms Bill should not contain any such forms of "harmful content or activity". They should instead be dealt with by a separate regulatory framework and other measures, distinct from those proposed in the White Paper.

Recommendation 8: If "harms with a less clear definition" are not to be addressed through a separate regulatory framework, distinct from that proposed in the White Paper, then the Online Harms Bill should explicitly state that the duty of care does not require online platforms to remove, restrict or moderate such forms of "harmful content or activity" and that the duty of care can be complied with through other actions (a "modified duty of care").

Recommendation 9: All forms of "harmful content or activity" that are to be addressed through the duty of care should be specified in the Online Harms Bill itself. They should also all be clearly and precisely defined in the Bill itself, or the Bill should make reference to other pieces of legislation which set out clear and precise definitions.

Recommendation 10: The regulatory body should not be given any power to introduce further forms of "harmful content or activity" or to require companies within scope to take

[14] See, for example, p. 71 of the White Paper, where the suggestions for a code of practice on disinformation refer to "making content which has been disputed by reputable fact-checking services less visible to users"; and p. 72, where the suggestions for a code of practice on self-harm and suicide include "steps companies should take to ensure that their services are safe by design, including (...) measures to block content".

any action in relation to any forms of "harmful content or activity" which are not specified in the Bill. The regulatory body could, of course, be given the power to make recommendations to the government that certain forms of "harmful content or activity" should be added to the Act, once passed.

Recommendation 11: If further forms of "harmful content or activity" are to be added to the Act, once passed, this should be done ideally via primary legislation. Alternatively, but less satisfactorily, this could be done via secondary legislation subject to the affirmative procedure. If secondary legislation is to be used to amend the list of forms of "harmful content or activity", the Online Harms Bill should require the government to consult beforehand on the particular forms that it is considering including. The Bill should explicitly state, however, that if further forms of "harmful content or activity" are added that are not already prohibited generally, whether through criminal law or otherwise, then the modified duty of care (**Recommendation 8**) will apply.

2. The regulatory model proposed

To address this diverse range of "online harms", Chapter 3 of the White Paper proposes one blanket solution: a statutory duty of care, accompanied by codes of practice relating to different harms. Before turning to that model specifically, we would note that each of the different forms of "harms" is a distinct public policy issue, requiring targeted and specific responses. Trying to deal with them through a single regulatory response fails to recognise the very different considerations that each requires. We recognise that the different codes of practice may well look very different, and set different expectations in relation to different forms of harm, but we are not convinced that this allows for the sufficient degree of tailoring when it comes to responding to different forms of harm, not least because they will only look at their online manifestation. We would have preferred to see, and continue to believe that these issues should have been considered separately, with regulatory responses that are both more carefully tailored and which address the issues holistically – including both offline and online dimensions – rather than solely focusing on their internet-related dimensions.

However, given that no alternative to the regulatory model proposed appears to be under consideration by the government, we have considered how the model is likely to work in practice, and have grave concerns over its potential impacts on the rights to freedom of expression and privacy.

A duty of care

In the UK, a duty of care is a legal obligation owed by one person (or entity) to another to ensure that the latter, in situations where there is a degree of proximity between the two, does not suffer any reasonably foreseeable harm or loss as a result of the former's acts or omissions. The legislature and the courts have shied away from establishing or recognising a duty of care to prevent others from suffering harm caused by the acts of third parties, unless a person or organisation has voluntarily assumed responsibility for their safety. Duties of care are almost exclusively found only in relation to physical spaces, and largely in relation to risks of physical harm or financial loss.

The duty of care which is proposed in the White Paper bears little resemblance to this existing understanding of what a duty of care means. While the idea of a duty of care on online platforms to take reasonable steps to protect users from harm may, at first glance, seem to be a logical extension to existing duties of care, there are a number of very real differences between what currently exists, and what it is proposed, which renders such an extension unsuitable.

First, existing duties of care only exist in relation to harm that is caused by the individual or entity's acts or omissions, and not those of third parties, even where they are on their

premises or using their services.[15] The owner of a premises is therefore not liable if a person on those premises commits a criminal offence through his or her speech. The manufacturers of cameras and those offering photographic processing services are not held liable if an individual takes photos which are illegal, such as indecent images of children. It is the individual who commits the prohibited action who is held liable in law, not those who owned the premises, or provided the products or services which facilitated that criminal offence.

The duty of care proposed in the White Paper is wholly different, therefore, from existing duties of care because it would make online platforms liable for the actions of third parties, i.e. for individuals generating or sharing particular forms of content or acting in a certain way. The fact that online platforms could, in theory, restrict or moderate such forms of content and activity does not change the situation. The owners of premises can, in theory, remove people who commit criminal offences through speech, and those who offer photographic processing services could, in theory, review all photographs and refuse to process those which contain illegal images. But they are not held liable if they fail to do so. To hold online platforms liable simply because content is generated or shared via their platforms, or users act in a way which is illegal or harmful, would run entirely contrary to existing duties of care and create an inconsistent system of liability.

Second, existing duties of care only exist in relation to harm which is objectively measurable, such as physical harm or financial loss. However, the proposals in the White Paper largely relate to harms which are not objectively measurable, such as emotional harm, distress, or impacts upon electoral processes, particularly those harms which have a "less clear definition". While the risk of a person being physically injured does not particularly depend on any aspect of that person, the risk of a person feeling "bullied", "trolled", "coerced" or "intimidated" will depend greatly upon various aspects of that person's personality and other characteristics, as well as the context in which the words or behaviour occur. It is therefore far more difficult for an online platform to determine the likelihood of such a risk materialising, and how to prevent it from happening in a way which would not also curtail identical or similar words or behaviour which do not cause harm.

Third, existing duties of care only exist where the risk is created by the actions or omissions of the individual or entity subject to the duty. However, the duty of care proposed in the White Paper will apply to companies merely by offering an online platform by which individuals can share or discover user-generated content or interact with each other, which is not in and of itself a risk. Just as those who manufacture paper, pens, mobile telephones, cameras and video recorders all develop tools that could, in theory, be used by individuals to say or do things which are illegal or harmful, there is no duty of care owed by those manufacturers to those who might be harmed. While the specific actions or omissions of an online platform might create a risk, for example if they use particular tools to filter or curate content, the duty of care would not only apply in such circumstances, but simply through the provision of a platform in and of itself. Such an approach suggests that enabling individuals to create and share content, and interact with each other, is in and of itself the creation of a risk, however no duty of care has ever existed, nor has been proposed, in relation to such circumstances offline.

Fourth, existing duties of care do not pose risks to human rights. Obligations to protect individuals from physical harm or financial loss, for example, do not create any risks to human rights through compliance or even over-compliance. They do not, therefore, require the same sorts of safeguards that are necessary when actions are taken which risk restricting human rights. As such, while simple measures such as a requirement to act proportionately may be sufficient when it comes to existing duties of care to mitigate risks related to compliance (such

[15] There is a limited exception to this general principle where a person or organisation has voluntarily assumed responsibility for their safety when it comes to the acts or omissions of third parties (see, for example, *Home Office v Dorset Yacht Co Ltd* [1970] UKHL 2). Such an exception would not apply to online platforms who do not take responsibility for the acts or omissions of third parties.

as disproportionate expenditure), these are not sufficient when for a duty of care which impacts upon and creates risks to freedom of expression, as would the duty of care set out in the White Paper.

Fifth, and related to the above, existing duties of care require a preventive approach, i.e. to identify where risks might occur and then take steps to mitigate them. Such an approach might be suitable in relation to risks of physical harm in a particular physical environment, it is not appropriate when it comes to restricting different forms of online content and activity. The duty of care proposed in the White Paper would require online platforms to take reasonable steps to protect users from harm and, as noted above, the proposals 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.[16] 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 remove 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 unquestionably be violations of the right to freedom of expression and the right to privacy:

- 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 disproportionate ways of addressing illegal and harmful speech, and therefore fail the third limb of the three-part test set out above. 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."[17]

General and specific monitoring

On this point, we have particular concerns over proposals in the White Paper for codes of practice to require "specific monitoring that targets where there is a threat to national security or the physical safety of children, such as CSEA and terrorism" (p.43). There is no logical or conceptual distinction between an obligation to undertake general monitoring of all forms of illegal content, and an obligation to undertake "specific" monitoring of two specific forms of illegal content. As soon as an online platform is required to undertake monitoring of all content on its platform in relation to a particular type of illegal content, it is undertaking general monitoring. We are unconvinced by the argument that this is, instead, "specific" monitoring.

[16] See, for example, the repeated suggestions in Chapter 7 of the White Paper that codes of practice will contain guidance on what steps will be required to proactively prevent new content from being made available to users, i.e. filtered at the point of upload

[17] See above, note 4, Para 1.3.5.

The use of automated processes and artificial intelligence

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 content and activity once uploaded. 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 content prior to upload, or identifying and removing it once uploaded. Indeed, the White Paper repeatedly suggests that the codes of practice will include details on how technology, such as AI, should be used to prevent certain forms of harmful content or activity.[18]

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.[19] AI has had some success in relation to images, as opposed to speech, with its most successful application being to identify copies of images already identified by humans as constituting child sexual abuse and exploitation. Using AI to identify new images of potentially illegal or harmful content or activity is far more difficult. 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.[20] Over-removal is even more likely when it comes to speech, given that context is even more relevant. 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 the UN Special Rapporteur has noted:

> "The weight of scientific research definitively indicates that '[a]rtificial intelligence-driven content moderation has several limitations, including the challenge of assessing context and taking into account widespread variation of language cues, meaning and linguistic and cultural particularities.'"[21]

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.

---

[18] See, for example, p. 66 of the White Paper on terrorist use of the internet ("guidance on proactive use of technological tools, where appropriate, to identify, flag, block or remove terrorist content.), p. 67 on serious violence (guidance on the content and/or activity companies should proactively identify, to either prevent it being made publicly available or prevent further sharing"), and p. 68 on hate crime ("guidance on the content and/or activity companies should proactively identify, to either prevent it being made publicly available or prevent further sharing.").

[19] 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-ofautomatedsocial-media-content-analysis.

[20] 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/contenttumblr-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/tumblrporn-adult-content-ban-user-backlash.

[21] UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression; and the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, UN Doc. OL AUS 5/2019, 4 April 2019.

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). 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.

Further, as the UN Special Rapporteur has also noted, "'artificial intelligence applications are often grounded in datasets that incorporate discriminatory assumptions,' and may result in content removals that reflect 'biased or discriminatory concepts.'"[22]

Recommendation 12: We do not believe that a "duty of care" is an appropriate regulatory model to tackle online harms. That being said, assuming that no other model than a "duty of care" is being considered by the government, the Online Harms Bill should explicitly make clear that the "duty of care" under the Bill is not comparable with, and should not be understood or interpreted in a like manner as, any other existing duty of care, whether found in other legislation or the common law.

Recommendation 13: The Online Harms Bill should explicitly state that compliance with the duty of care does not require, and should not be interpreted by the regulatory body or any court as generally requiring:

- Companies within scope to filter content at the point of upload, generation or sharing;
- Companies within scope either to generally or proactively monitor content; or
- Companies within scope to use artificial intelligence or other forms of automated decision-making.

Recommendation 14: The Online Harms Bill should explicitly state that if codes of practice are to include any measures set out in **Recommendation 13**, such measures are only required in relation to forms of harmful content or activity which are illegal and where the filtering or proactive identification is of copies of content which have already been identified by a human as illegal.

We repeat **Recommendation 8**: If "harms with a less clear definition" are not to be addressed through a separate regulatory framework, distinct from that proposed in the White Paper, then the Online Harms Bill should explicitly state that the duty of care does not require online platforms to remove, restrict or moderate such forms of "harmful content or activity" and that the duty of care can be complied with through other actions (a "modified duty of care").

Codes of practice

We recognise the value that codes of practice could play in assisting online platforms to understand how they can comply with their duty of care. While we note that the codes of practice would not strictly be binding, the White Paper states that "[t]here will be a strong expectation that companies follow the guidance set out in these codes. If they choose not to do so, companies will have to explain and justify to the regulator how their alternative approach will effectively deliver the same or greater level of impact" (p. 43). As such, it is safe to assume that, in practice, most companies, particularly small and medium-sized companies, will simply follow the codes of practice rather than develop alternative approaches.

[22] *Ibid*.

Rather than suggest the outcomes that are to be achieved, the codes of practice proposed by the White Paper appear to be highly prescriptive, setting out "the systems, procedures, technologies and investment, including in staffing, training and support of human moderators, that companies need to adopt" (p. 42). We are particularly concerned over many of the suggestions of the types of requirements that will be included in codes of practice in Chapter 7 of the White Paper, such as:

- Setting the terms of service and content moderation policies that online platforms should use;
- Requiring online platforms to make certain forms of content from being made available to users at the point of upload;
- Mandating particular processes for moderating content;
- Mandating particular forms of technology to identify, flag, block or remove certain forms of content;
- Mandating particular timeframes for the removal of certain forms of content; and
- Mandating that online platforms modify their search results when there is a risk that they might lead to certain forms of content appearing.

A public body which made demands of these sorts from online platforms would risk being in breach of international, European and the HRA 1998 given the risks to the rights to freedom of expression and privacy that stem from their application. In relation to the imposition of particular time limits, for example, Recommendation CM/Rec(2018)2 states that legal requirements should not be "designed in a manner that incentivises the take-down of legal content, for example due to inappropriately short timeframes".[23] Responding to the comparable NetzDG in Germany, which imposes time limits of 24 hours and 7 days for online platforms to make decisions regarding particular forms of content, the UN Special Rapporteur highlighted the fact that:

> "The short deadlines (...) could lead social networks to over-regulate expression - in particular, to delete legitimate expression, not susceptible to restriction under human rights law, as a precaution to avoid penalties. Such pre-cautionary censorship, would interfere with the right to seek, receive and impart information of all kinds on the internet."[24]

Indeed, since the introduction of NetzDG, 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.[25]

More recently, in response to the Australian Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019, which simply requires online platforms to remove certain forms of content "expeditiously", the UN Special Rapporteur (along with the UN Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism) made clear that:

> "Short timelines pose highly negative implications to the practical realization of protection for freedom of expression and interlinked rights in real time. We are concerned that accelerated time lines will not allow Internet platforms sufficient time to examine requests in detail, and may

[23] See above, note 4, Para 1.3.7.
[24] UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, UN Doc. OL DEU 1/2017, 1 June 2017.
[25] 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/germanyhate-speech-netzdg-facebook-youtube-google-twitterfree-speech.

in practice mean that providers will consistently produce an abundance of caution, for concern of financial fines and other consequences."[26]

It is therefore vital that, before any codes of practice are made binding, or if they are revised, they are published in draft with a full consultation process which includes an assessment of the potential risks to freedom of expression.

Recommendation 15: The Online Harms Bill should explicitly require the regulatory body to publish any codes of practice in draft form and to consult upon them before a final code of practice is adopted.

Recommendation 16: In addition to any general duty upon the regulatory body to protect and respect the rights to freedom of expression and privacy, the Online Harms Bill should further explicitly require the regulatory body to undertake a human rights impact assessment, which includes consideration of potential impacts upon these rights, when developing or revising any code of practice. This human rights impact assessment should be published alongside the draft code of practice.

Recommendation 17: The Online Harms Bill should require all draft codes of practice should be scrutinised by the Joint Committee on Human Rights before any final code of practice is adopted by the regulatory body. Section 120(6) of the Coroners and Justice Act 2009 provides a precedent for this form of scrutiny, providing that draft sentencing guidelines developed by the Sentencing Council must be scrutinised by the Justice Committee of the House of Commons.

Transparency reports

The same principles which underpin permissible restrictions on freedom of expression apply online as they do offline. This means that restrictions on that right, including the removal, restriction or moderation 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 online platforms to decide on whether content is illegal shift judicial and quasi-judicial functions to those platforms or their nominees. We have highlighted our concerns over such proposals above. However, the White Paper makes no proposals to guarantee that there would be mechanisms for transparency or accountability, as there are when equivalent decisions are made by public authorities.

When it comes to the "harms with a less clear definition", there are similar concerns over whether companies are well-placed and able to make determinations as to what content is harmful if no clear, precise definitions are provided. The sheer scale of content means that human reviews are unlikely to be feasible, and we have highlighted above how automated processes are poor at making decisions at identifying particular kinds of content. As with illegal content being removed, there would not necessarily be any mechanisms for transparency or accountability.

Given this, we believe there is a particularly critical role for transparency over online platforms decisionmaking, both in relation to "harmful content or activity", as well as their own content moderation policies. As the White Paper notes, it is not always clear at present how platforms make decisions about what content to remove, the standards and processes

[26] See above, note 18.

that are employed, those involved in the process, and how the quality of decisionmaking is ensured (p. 37).

Mandatory transparency reporting requirements would encourage online 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 and moderation. Good practice could be more easily identified and adopted by other platforms. 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 and more quickly.

As such, we welcome the proposals in the White Paper for mandatory transparency reporting. Indeed, we believe that this should be the most critical part of any regulatory proposals ultimately put forward, and should be the focus and starting point of any regulatory enforcement. We particularly welcome the move away from quantitative transparency reporting which is of little value and risks incentivising the removal of ever more pieces of content. The focus on qualitative transparency reporting will prove much more helpful and effective in identifying areas where improvement is needed.

To really drive transparency and accountability, and therefore build trust, it is critical that platforms are required to provide sufficient detail when it comes to their own actions. The following should be part of any mandatory transparency reporting:

- Details on how the platform develops its terms of service which touch upon content moderation, including if and how these are revised, and how external stakeholders are involved in any development and revision processes;
- Details on how the platform enforces its terms of service which touch upon content moderation, including the publication of any internal enforcement policies or guidance, and details on the number of moderators and how they are trained and supported;
- Details on how the platform makes decisions over the legality of content which is reported, where it is not prohibited by its own terms of service;
- Details of any demands or requests that the platform receives from law enforcement agencies, courts or other public bodies for the removal or moderation of content and the platform's responses;
- Details on any use of automated processes such as artificial intelligence to identify or moderate content, to what extent human moderation is involved in such circumstances, and what safeguards are in place to prevent inappropriate identification or moderation;
- Details on any use of automated processes such as artificial intelligence to filter or curate the content that an individual user sees, or the order in which they see it;
- Details on any processes in place which ensure users are informed when content they have posted or shared is moderated in any way, how they can challenge that decision, and how reviews are considered; and
- Any further information on how the platform fulfils its responsibility under the UN Guiding Principles to respect human rights, including the right to freedom of expression.

We further consider that the importance of mandatory transparency reporting means that any transparency reporting templates should be developed by the regulator body in collaboration with other relevant stakeholders, including civil society.

Recommendation 18: The Online Harms Bill should explicitly set out the issues which must be included in any mandatory transparency reporting templates and these should include those issues set out above.

Recommendation 19: The Online Harms Bill should also provide the regulator with discretion to consider any further issues, to ask further questions to some or all platforms, as well as to determine the precise format and wording contained within the template.

Recommendation 20: The Online Harms Bill should explicitly state that any mandatory transparency reporting templates should be published by the regulator in draft form and be subject to consultation before a final template is adopted.

Companies within scope

Chapter 4 of the White Paper states that "the regulatory framework will apply to companies that provide services or tools that allow, enable or facilitate users to share or discover user-generated content, or interact with each other online". Our analysis of the scope of companies who will be subject to the duty of care is set out in our response to questions 5 to 7 of the consultation below. In summary, we do not consider that there is a sufficiently strong evidence base for the scope of companies to be as broad as it is, and that many types of online platforms will be captured despite there being no evidence of harm being caused or facilitated thereby. In addition, we have particular concerns around online platforms which provide private communication services being within the scope of the regulatory framework.

3. Enforcement

A regulatory body

The existence of a regulatory body, proposed in Chapter 5 of the White Paper, does not, in and of itself, pose any particular concerns in relation to the right to freedom of expression and privacy. However, risks may well be created, other than those set out elsewhere in our response, stemming from how the regulatory body fulfils its functions, particularly in relation to the content of the codes of practice it develops, and its approach towards enforcement. The Online Harms Bill should therefore include appropriate provisions in relation to how the regulatory body will operate and exercise its functions, and which mitigate such risks should therefore be included. Under no circumstances should the government be able to direct the regulatory body with regard to the development or enforcement of its codes of practice, as is proposed in the White Paper (p. 43).

Recommendation 21: The Online Harms Bill should provide for the involvement of the Equality and Human Rights Commission in the work of the new regulatory body, and in particular the enforcement of its duties and functions, in order to determine impacts upon freedom of expression and privacy.

Recommendation 22: The regulatory body should 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.

Recommendation 23: In addition to the above, the Online Harms Bill should explicitly state that protecting and respecting the rights to freedom of expression and privacy is one of the regulatory body's statutory duties.

Recommendation 24: The regulatory body should be fully independent from government and political direction from government in all aspects, including the development and enforcement of its codes of practice.

Recommendation 25: The regulatory body should have a clear research and evidence-gathering function, and this should inform all of its work in developing codes of practice and

undertaking its enforcement powers. This function could be mirrored on the Food Standards Agency's function under section 8 the Food Standards Act 1999 which provides that:

- The FSA has the function of "obtaining, compiling and keeping under review information about matters connected with food safety and other interests of consumers in relation to food";
- This function includes "monitoring developments in science, technology and other fields of knowledge" relating to the above, and "carrying out, commissioning or co-ordinating research" on them; and
- The FSA should carry out that function "with a view to ensuring that the Agency has sufficient information to enable it to take informed decisions and to carry out its other functions effectively".

A similar function exists in relation to OfCom under sections 14 to 16 of the Communications Act 2003.

Recommendation 26: The government should consider further means by which relevant expertise, including on human rights, informs and reviews the work of the regulatory body. In addition to a statutory function to undertake research to inform its work, the Online Harms Bill could, for example, require the regulatory body to establish a standing advisory committee on human rights to inform and review the work of the regulatory body. A comparable requirement exists in relation to OfCom which, under section 21 of the Communications Act 2003, is required to establish an advisory committee on elderly and disabled persons.

Recommendation 27: The Online Harms Bill should require the regulatory body to report annually on the exercise of its functions. This annual report should include an assessment of how the body has complied with its duty to protecting and respecting the rights to freedom of expression and privacy (**Recommendation 23**).

Sanctions

While the proposals in relation to the scope of harms and the regulatory model themselves pose very real risks to the rights to freedom of expression in and of themselves, the proposals for enforcement proposed in Chapter 6 of the White Paper amplify these risks, particularly through the specific sanctions.

Heavy or disproportionate sanctions will skew incentives and exacerbate the risks to freedom of expression outlined above. If an online platform is making decisions as to whether to remove, restrict or otherwise moderate 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 do so rather than risk sanction. The heavier the potential sanction, the greater the incentive. 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 sanctions are likely to lead to the restriction of lawful content and to have a chilling effect on the right to freedom of expression."[27]

We consider that three of the proposed "core powers" of the regulator are proportionate, namely (i) serving a notice to a company that is alleged to have breached standards, and setting a timeframe to respond with an action plan to rectify the issue; (ii) requiring additional information from the company regarding the alleged breach; and (iii) publishing public notices about the proven failure of the company to comply with standards.

[27] See above, note 4, Para 1.3.6.

With respect to the fourth "core power", issuing civil fines for proven failures in clearly defined circumstances, it is difficult to assess the proportionality of such a measure without further information. Without any information on what the size of these fines could be, or what the "clearly defined circumstances" are that will lead to their being issued, there are particular risks raised by this proposal. The comparable NetzDG in Germany, for example, allows for fines of up to 5 million euro for non-compliance, and that this has been determined by the UN Special Rapporteur potentially representing an undue interference with the right to freedom of expression on the basis that high fines "raise proportionality concerns, and may prompt social networks to remove content that may be lawful".[28]

Recognising that the White Paper is not consulting on the "core powers", it is essential that the power to issue fines is constrained to ensure that it is neither disproportionately used, nor incentivises online platforms to remove content which may be protected by the right to freedom of expression.

Recommendation 28: The Online Harms Bill should explicitly provide that the regulatory body's power to issue a civil fine cannot be exercised unless and until the other three "core powers" have been exercised and there has been a failure to comply with them. This would ensure that online platforms have sufficient opportunities to be aware of concerns and to respond to them before a risk of a fine materialises.

Recommendation 29: The Online Harms Bill should explicitly provide that the regulator may not issue a civil fine of an amount with would be disproportionate taking into account the size and resources of the online platform, and the level of harm or potential harm caused as a result of the online platform's non-compliance with its legal obligations.

We repeat **Recommendation 6**: The Online Harms Bill should also explicitly provide that in relation to any potential enforcement proceedings or action taken under the Bill, whether by the regulator or another body, an online platform is able to argue that, with respect to the alleged non-compliance, it was reasonably acting in accordance with its requirement to take sufficient measures to safeguard the right to freedom of expression. Where such an argument is raised by an online platform, the regulator or other body enforcing the Bill should be required to reconsider its enforcement proceedings or action, and withdraw them if the online platform was in fact reasonably acting in accordance with its requirement.

We have even greater concerns in relation to the further powers under consideration, namely the imposition of civil or criminal liability on senior managers of online platforms, compelling ISPs to block websites, and disrupting business activities. We do not consider that there are safeguards that can be put in place to mitigate the risks to freedom of expression that would stem from these being potential sanctions an online platform could face.

We would note that the first of these echoes provisions which were also contained in the Australian Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019. These provisions were also criticised by the UN Special Rapporteur and the UN Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, particularly given the imposition of time limits on decisionmaking:

> "[T]he time and effort required to make such nuanced assessments of context and preserve protected exercises of freedom of expression are at odds with the proposed obligation on service providers to 'expeditiously' remove content. Given these conflicting considerations, the threat of criminal sanctions is likely to tip the scales in favor of disproportionate

[28] See above, note 24.

restrictions on freedom of expression, which may undermine rather than protect the public interest."[29]

Recommendation 30: The Online Harms Bill should not contain any of the further means of enforcement proposed in the White Paper, namely the imposition of civil or criminal liability on senior managers of online platforms, compelling ISPs to block websites, and disrupting business activities.

4. Safeguards for freedom of expression and privacy

As noted above, the White Paper makes repeated reference to the UK's commitment to freedom of expression and privacy, evidencing a recognition by the government that there are potential impacts and risks to these rights which stem from the proposals. Indeed, our analysis, as set out above, highlights some very real and serious risks which require substantial mitigation. Given these risks, and despite the government's commitment to freedom of expression and privacy, the White Paper fails to acknowledgement or analyse of the risks posed in any meaningful way, or how they would be mitigated through effective, practical safeguards.

Indeed, there are only two points where the commitment to freedom of expression and privacy is explicitly reflected through specific proposals. The first is an obligation on the regulatory body "to protect users' rights online, particularly rights to privacy and freedom of expression" (p. 44). By itself, however, it is difficult to see how this requirement is not undermined by the remainder of the proposals relating to the regulatory body and its enforcement powers, which, as we highlight above, strongly incentivise the removal of legal and legitimate content.

The second is an inclusion of a requirement in the transparency reports for companies to set out "measures and safeguards in place to uphold and protect fundamental rights" (p. 56) but there are no obligations on companies to develop and implement such measures and safeguards in the first place, simply to report on whether or not they have any. Nor are there any proposals for the regulatory body to take any action where a company does not have any measures in place to uphold and protect human rights.

As such, these two proposals, while welcome, do not go far enough to ensure meaningful protection of the right to freedom of expression and privacy given the very real risks which are created by the model and which we have set out above. The recommendations we make in this consultation response, if fully incorporated and implemented, would go some way to mitigating these risks.

5. Cumulative impact

Although each of the specific aspects of the proposals analysed above poses risks to the rights to freedom of expression and privacy, their cumulative impact significantly amplifies those risks. If an online platform has a duty of care to prevent a particular form of harm to its users, but that harm is not clearly defined, then the risk of heavy sanctions creates a strong incentive to ensure that all content which is potentially harmful is removed, even if this means removing content which is not actually harmful. There is no incentive to make careful, nuanced decisions as to whether the content is actually illegal or harmful, given that there is only a risk of sanction through a failure to remove, and not over-removal. There is a further incentive to restrict the availability of potentially harmful content in the first place, rather than wait for it to be reported by users. This means filtering content, invariably using artificial intelligence or

[29] See above, note 21.

automated decisionmaking, which as we have highlighted above, will inevitably filter out content which is neither illegal nor harmful.

In order to address the risks created by the cumulative impact of the proposals in the Online Harms White Paper, there is a further, essential safeguard that should be included in the Online Harms Bill, and that is to maintain the existing intermediary liability framework set out in the e-Commerce Directive.

Recommendation 31: The Online Harms Bill should put Articles 17 to 19 of the Electronic Commerce (EC Directive) Regulations 2002 into primary legislation. This would, as is recommended by Recommendation CM/Rec(2018)2, make clear that online platforms 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.

Further, it is difficult to fully assess the ultimate impact upon the rights to freedom of expression and privacy as a result of many of the issues highlighted above being subject to consultation and further refinement. As such, it will only be when the actual legislation is presented that a fully informed analysis can be undertaken. Given the novelty of this policy area, and the importance of proceeding with caution, the proposed legislation should be published in draft and subjected to pre-legislative scrutiny before a final Bill is put before Parliament.

Recommendation 32: The Online Harms Bill should be published in draft and subjected to pre-legislative scrutiny before a final version of the Bill is presented to Parliament.

The consultation questions posed in the Online Harms White Paper

Question 1: This government has committed to annual transparency reporting. Beyond the measures set out in this White Paper, should the government do more to build a culture of transparency, trust and accountability across industry and, if so, what?

We welcome the proposals in the White Paper for mandatory transparency reporting and refer to our comments above under "Transparency reports" as to the further steps that should be taken to enhance transparency, trust and accountability.

We repeat **Recommendation 18**: The Online Harms Bill should explicitly set out the issues which must be included in any mandatory transparency reporting templates and these should include those issues set out above, namely:

- Details on how the company develops its terms of service which touch upon content moderation, including if and how these are revised, and how external stakeholders are involved in any development and revision processes;
- Details on how the company enforces its terms of service which touch upon content moderation, including the publication of any internal enforcement policies or guidance, and details on the number of moderators and how they are trained and supported;
- Details on how the company makes decisions over the legality of content which is reported, where it is not prohibited by its own terms of service;
- Details of any demands or requests that the company receives from law enforcement agencies, courts or other public bodies for the removal or moderation of content and the company's responses;

- Details on any use of automated processes such as artificial intelligence to identify or moderate content, to what extent human moderation is involved in such circumstances, and what safeguards are in place to prevent inappropriate identification or moderation;
- Details on any use of automated processes such as artificial intelligence to filter or curate the content that an individual user sees, or the order in which they see it;
- Details on any processes in place which ensure users are informed when content they have posted or shared is moderated in any way, how they can challenge that decision, and how reviews are considered; and
- Any further information on how the company fulfils its responsibility under the UN Guiding Principles to respect the right to freedom of expression.

We repeat **Recommendation 19**: The Online Harms Bill should also provide the regulator with discretion to consider any further issues, to ask further questions to some or all companies, as well as to determine the precise format and wording contained within the template.

We repeat **Recommendation 20**: The Online Harms Bill should explicitly state that any mandatory transparency reporting templates should be published by the regulator in draft form and be subject to consultation before a final template is adopted.

Question 2: Should designated bodies be able to bring 'super complaints' to the regulator in specific and clearly evidenced circumstances?

No. The White Paper provides few details on what these "super complaints" would look like. In the absence of any clarity, we are unable to support such a proposal. We would instead encourage the regulator, working with companies, to develop guidance on how users can raise concerns with online platforms over the existence or moderation of a particular piece of content.

Question 3: What, if any, other measures should the government consider for users who wish to raise concerns about specific pieces of harmful content or activity, and/or breaches of the duty of care?

Although the White Paper is clear that users should have access to a complaints function regarding "specific pieces of harmful content or activity, or wider concerns that the company has breached its duty of care", there is no equivalent requirement for a complaints function when an online platform has removed or moderated content or activity inappropriate, or adversely impacted a user's right to freedom of expression.

This one-way approach regarding users' ability to make complaints runs counter to other parts of the White Paper. On p. 70, for example, the White Paper states that the regulator is likely to develop a code of practice which sets out the processes that companies should have in place to appeal the removal of content, or other responses, in order to protect users' rights online. And, as noted above, the White Paper also makes repeated reference to the need to ensure protection of the right to freedom of expression. This one-way approach is also inconsistent with the UK's obligations under the UNGPs, Principle 3 of which requires, among other things, states to "enforce laws that are aimed at, or have the effect of, requiring business enterprises to respect human rights". Enabling users to challenge the actions of online platforms that amounted to adverse impacts upon their right to freedom of expression would strongly incentivise platforms better to respect that right through their content moderation policies and enforcement.

As such, users should have the ability to challenge decisions to remove, restrict or moderate content if they consider that the removal, restriction or moderation amounted to an adverse interference with their right to freedom of expression.

Recommendation 33: The Online Harms Bill should require online platforms to provide opportunities for individuals to challenge content moderation decisions which they consider to amount to a restriction on their right to freedom of expression. Such opportunities should be equivalent to those provided to individuals to raise concerns about specific pieces of harmful content or activity, and/or breaches of the duty of care. These opportunities should include, at a minimum, (i) requiring affected users to be informed by the platform of content that has been flagged for removal, restriction or moderation; (ii) requiring an opportunity for that user to be able to input into the moderation process; and (iii) requiring platforms to introduce independent appeal mechanisms for affected users to challenge decisions.

Question 4: What role should Parliament play in scrutinising the work of the regulator, including the development of codes of practice?

As we note above, we do not believe that the regulator should be able to determine the forms of "harmful content and activity" which online platforms will be expected to take action to prevent. Given that this role is, in essence, one which determines which forms of expression are and are not permitted online, this is a role that must be reserved for a democratically elected legislature. However, we believe that Parliament has two further roles to play in relation to the work of the regulatory body, which relate to the approval of the codes of practice and scrutiny of the regulatory body's work.

We repeat **Recommendation 17**: The Online Harms Bill should require all draft codes of practice should be scrutinised by the Joint Committee on Human Rights before any final code of practice is adopted by the regulatory body. Section 120(6) of the Coroners and Justice Act 2009 provides a precedent for this form of scrutiny, providing that draft sentencing guidelines developed by the Sentencing Council must be scrutinised by the Justice Committee of the House of Commons.

Recommendation 34: Given the significant impact of the regulatory body upon the right to freedom of expression, the Chief Executive and other relevant members of staff at the regulatory body should be required to report annually before the Joint Committee on Human Rights, as are the Chief Executives of many other public bodies before other parliamentary committees.

Question 5: Are proposals for the online platforms and services in scope of the regulatory framework a suitable basis for an effective and proportionate approach?

No. We recognise the difficulty in providing a clear definition of the online platforms and services in scope of the regulatory framework, however the breadth of the current definition draws in a number of types of companies without there being any evidence of harm being caused or facilitated by their services. As currently worded, the definition would include not only social media platforms - which are clearly the target of the proposals - but websites which allow users to leave comments or reviews, the websites of newspapers which have comments functions, encrypted cloud storage service providers, private messaging applications, web hosting service providers and website developers.

We have not seen any evidence, and nor is any provided in the White Paper, that all of these types of companies are either causing harm or facilitating harm via their online platforms and services. Their inclusion in the scope of the proposals is therefore neither proportionate, nor

will it help ensure effectiveness, and time and resources of both the regulator and such companies will be spent unnecessarily on enforcing, and comply with, the duty of care.

The government should revisit its definition of companies which are within scope, to narrow it as far as possible to companies where evidence exists of their causing or facilitating harm via their online platforms and services. Such a definition should be set out in the Online Harms Bill, and amended only if further evidence appears that further types of companies are causing or facilitating harm.

Recommendation 35: The Online Harms Bill should define the types of companies within scope narrowly, so as to include only those types of companies where there is clear evidence of their having caused or facilitated harm via their online platforms and services.

Recommendation 36: In order to allow some flexibility, the Online Harms Bill should also provide the power to the government to develop secondary legislation which sets out whether particular companies, or types of companies, are within or outside of the scope of that definition. Such secondary legislation should be subject to the affirmative procedure. The Online Harms Bill should also require the government to consult beforehand on the companies, or particular types of companies, that it is considering including in secondary legislation.

Recommendation 37: If the definition of the types of companies within scope in the Act, once passed, is to be amended, this should be done ideally via primary legislation. Alternatively, but less satisfactorily, this could be done via secondary legislation subject to the affirmative procedure. If secondary legislation is to be used to amend the definition, the Bill should require the government to consult beforehand on the new definition that it is considering.

Question 6: In developing a definition for private communications, what criteria should be considered?

As we note above, we are concerned that the White Paper is proposing that private channels be subject to all or some aspects of the proposed regulatory framework. For the reasons set out in our response to question 7 below, private channels should be entirely out of scope of the proposed regulatory framework.

Ultimately, the most important criterion in determining whether communications are private or not is whether or not, within those communications, the platform or the participants have taken steps to make those communications inaccessible to non-participants, including the online platform which providers those services. This would include communications between two individuals and communications between a particular, specified group of specified individuals. If the communications are inaccessible to non-participants or to the online platform which provides those services, then they should be considered private communications and out of scope of this regulatory framework.

Further, if the communications are protected by end-to-end encryption, then this would also render them inaccessible to anyone other than those who are communicating, including the communication service providers Despite the suggestion in the White Paper otherwise (p. 50), it should not matter whether those communications are between two people or a larger group. If the communications channel is protected by end-to-end encryption, then it should be considered private and out of scope of this regulatory framework.

Recommendation 38: The Online Harms Bill should explicitly provide that, regardless of its definition of companies within scope, the regulatory framework does not apply with respect to (i) companies who only provide services which are protected by end-to-end encryption,

and (ii) companies who provide a range of services, which includes at least one which is protected by end-to-end encryption, in respect of those services which are protected by end-to-end encryption.

Question 7: Which channels or forums that can be considered private should be in scope of the regulatory framework?

None. Private communications should be entirely out of scope of this regulatory framework. Such channels and forums use, almost universally, end-to-end encryption, which means those who develop and provide such services are almost entirely unable to filter or monitor, or otherwise access or moderate content which is generated or shared using them. The application of the proposed regulatory framework would simply not be workable in any way unless those channels and forums ceased to use end-to-end encryption, which would amount to an unjustifiable restriction on the right to privacy. As such, any regulatory framework which applied to private channels and forums would be of such a different nature to the regulatory framework being proposed in the White Paper, that it should be considered entirely distinctly.

Question 7a: What specific requirements might be appropriate to apply to private channels and forums in order to tackle online harms?

There is scope for some very limited application of a regulatory framework on private channels or forums. However, this should be limited to requiring them (i) to enable users to report illegal and harmful content, and (ii) provide transparency reports on how they enable users to do so.

Question 8: What further steps could be taken to ensure the regulator will act in a targeted and proportionate manner?

We welcome the commitment in the White Paper that the regulatory body will take a "proportionate approach" (p. 55). However, the fact that so many other aspects of the proposals - such as the scope of harms included, the scope of companies subject to the duty of care, the requirements that will be necessary to fulfil in order to comply with the duty of care, and the potential sanctions - are all so wide undermines this commitment. While a requirement for the regulatory body to act proportionately is good, there will be fewer risks of a disproportionate approach if those other aspects of the proposals are more tightly and narrowly constrained, as we recommend throughout this response. In addition, the Online Harms Bill should explicitly set out further controls on the regulatory body in order to mitigate risks against disproportionate action.

Recommendation 39: The Online Harms Bill should explicitly require the regulatory body to undertake an impact assessment for each code of practice it publishes. This impact assessment should be published alongside a draft version of the code of practice. It should set out an evidence-based assumption of the potential costs to companies of compliance with the code of practice as well as the potential benefits.

Recommendation 40: The Online Harms Bill should explicitly require the regulatory body to develop a regulatory action policy or similar guidance, which contains details on how the regulatory body will exercise its functions in a proportionate manner. The Online Harms Bill should explicitly provide that the regulatory body cannot undertake any regulatory enforcement until its final policy has been published.

This requirement could be mirrored on the Information Commissioner's Office's (ICO) requirement under sections 160 and 161 the Data Protection 2018 which provides that:

- The ICO must produce and publish guidance about how it proposes to exercise its enforcement functions, including the factors that it will take into account when deciding whether and how to exercise them; and
- The first version of the guidance must be submitted to the Secretary of State and laid before Parliament.

A similar, albeit more narrow, requirement exists in relation to OfCom where section 392 of the Communications Act 2003 provides that:

- OfCom must prepare and publish a statement containing the guidelines they propose to follow in determining the amount of penalties imposed by them under provisions contained in the Act; and
- Before publishing a statement or revised statement, OfCom must consult both the Secretary of State, and any other persons as they consider appropriate, about the guidelines they are proposing to include in the statement.

Recommendation 41: In addition to the above, the Online Harms Bill should require the regulatory body to publish its proposed regulatory action policy in draft form and for it to be open to consultation prior to submission to the Secretary of State.

We repeat **Recommendation 28**: The Online Harms Bill should explicitly provide that the regulatory body's power to issue a civil fine cannot be exercised unless and until the other three "core powers" have been exercised and there has been a failure to comply with them. This would ensure that online platforms have sufficient opportunities to be aware of concerns and to respond to them before a risk of a fine materialises.

We repeat **Recommendation 29**: The Online Harms Bill should explicitly provide that the regulator may not issue a civil fine of an amount with would be disproportionate taking into account the size and resources of the online platform, and the level of harm or potential harm caused as a result of the online platform's non-compliance with its legal obligations.

Question 9: What, if any, advice or support could the regulator provide to businesses, particularly start-ups and SMEs, comply with the regulatory framework?

We have no particular comments on what advice or support the regulatory body could provide to businesses to comply with the regulatory framework.

Question 10: Should an online harms regulator be: (i) a new public body, or (ii) an existing public body?

We believe that the particular expertise required by a regulatory body, as well as the focus that it should direct towards the specific online platforms and activities requires (i) the creation of a new public body.

Question 11: A new or existing regulator is intended to be cost neutral: on what basis should any funding contributions from industry be determined?

We have no particular comments on what basis any funding contributions from industry should be determined.

Question 12: Should the regulator be empowered to i) disrupt business activities, or ii) undertake ISP blocking, or iii) implement a regime for senior management liability? What, if any, further powers should be available to the regulator?

No, we do not believe that the regulator should be empowered to i) disrupt business activities, ii) undertake ISP blocking, or iii) implement a regime for senior management liability. Nor should it have any further powers. We refer to our position set out above under "Enforcement" and particularly our comments on the proposal for the regulatory body to be able to issue civil fines.

We repeat **Recommendation 28**: The Online Harms Bill should explicitly provide that the regulatory body's power to issue a civil fine cannot be exercised unless and until the other three "core powers" have been exercised and there has been a failure to comply with them. This would ensure that online platforms have sufficient opportunities to be aware of concerns and to respond to them before a risk of a fine materialises.

We repeat **Recommendation 29**: The Online Harms Bill should explicitly provide that the regulator may not issue a civil fine of an amount with would be disproportionate taking into account the size and resources of the online platform, and the level of harm or potential harm caused as a result of the online platform's non-compliance with its legal obligations.

We repeat **Recommendation 6**: The Online Harms Bill should also explicitly provide that in relation to any potential enforcement proceedings or action taken under the Bill, whether by the regulator or another body, an online platform is able to argue that, with respect to the alleged non-compliance, it was reasonably acting in accordance with its requirement to take sufficient measures to safeguard the right to freedom of expression. Where such an argument is raised by an online platform, the regulator or other body enforcing the Bill should be required to reconsider its enforcement proceedings or action, and withdraw them if the online platform was in fact reasonably acting in accordance with its requirement.

Question 13: Should the regulator have the power to require a company based outside the UK and EEA to appoint a nominated representative in the UK or EEA in certain circumstances?

We have no particular comments on whether the regulatory body should have the power to require a company based outside the UK and EEA to appoint a nominated representative in the UK or EEA in certain circumstances.

Question 14: In addition to judicial review should there be a statutory mechanism for companies to appeal against a decision of the regulator, as exists in relation to Ofcom under sections 192-196 of the Communications Act 2003?

Yes. The costs and other resources involved in bringing a claim for judicial review are so significant that this will not provide a realistic and practical opportunity for many of the companies within scope, particularly small and medium-sized companies, to appeal decisions of the regulatory body. A simpler, cheaper and more accessible mechanism is therefore essential.

Question 14a: If your answer to question 14 is 'yes', in what circumstances should companies be able to use this statutory mechanism?

As with sections 192-196 of the Communications Act 2003, a company should be able to use the statutory mechanism whenever it considers that a decision of the regulator is based on an error of fact or was wrong in law or both. Given the particular risks to freedom of expression and privacy which stem from the proposals, a company should also be able to use the statutory mechanism whenever it considers that a decision of the regulator is unlawful by virtue of section 6(1) of the Human Rights Act 1998 (i.e. that it is incompatible with one or more rights contained within the European Convention on Human Rights).

Recommendation 42: The Online Harms Bill should explicitly provide that a company is able to use a statutory mechanism to appeal against a decision of the regulatory body wherever it considers that the decision was based on an error of fact, was wrong in law, was unlawful by virtue of section 6(1) of the Human Rights Act 1998, or a combination of the three.

Recommendation 43: The Online Harms Bill should ensure that a "decision" of the regulator includes, at a minimum, (i) a decision to publish a particular code of practice and (ii) a decision to undertake a particular action of enforcement.

Question 14b: If your answer to question 14 is 'yes', should the appeal be decided on the basis of the principles that would be applied on an application for judicial review or on the merits of the case?

The appeal should be decided on the basis of the merits of the case.

Question 15: What are the greatest opportunities and barriers for (i) innovation and (ii) adoption of safety technologies by UK organisations, and what role should government play in addressing these?

We have no particular comments on what the greatest opportunities and barriers for (i) innovation and (ii) adoption of safety technologies by UK organisations are, and what role should government play in addressing these.

Question 16: What, if any, are the most significant areas in which organisations need practical guidance to build products that are safe by design?

We have no comments on what the most significant areas in which organisations need practical guidance to build products that are safe by design are.

Question 17: Should the government be doing more to help people manage their own and their children's online safety and, if so, what?

We have no comments on any further efforts the government should be making to help people manage their own and their children's online safety.

Question 18: What, if any, role should the regulator have in relation to education and awareness activity?

We welcome recognition of the importance of education and awareness activities, as well as increasing the level of digital and media literacy among the population. These are critically important to tackling the harms identified in the White Paper, and it is disappointing that these measures were relegated to the end of the White Paper, rather than being at the heart of it.

We welcome the proposal for the regulatory body to have a responsibility to promote online media literacy. We believe that the regulatory body has an important role to play in highlighting what existing efforts are being made by online platforms to promote digital literacy, and to complement these efforts by promoting coordination among online platforms and promoting digital literacy itself where gaps exist.

As part of this, we note the proposal for the regulatory body to have the power "to require companies to report on their education and awareness raising activity" (p. 93). We have no concerns with such a power and see the benefit of such a power in order for the regulatory body to be able to identify existing efforts, promote coordination, and identify gaps where it

can promote digital literacy itself. We note, however, that the same section of the Online Harms Whites Paper also states that the government is "consulting on appropriate powers for the regulator in this area". We would not support any powers that went beyond requiring information from online platforms, such as the power to require online platforms to undertake particular forms of education and awareness raising activities with sanctions for non-compliance.

Recommendation 44: The Online Harms Bill should explicitly provide for the regulatory body to have the power to require information from online platforms on their education and awareness raising activities. The Bill should not, however, provide for any powers that go beyond this, such as the power to require online platforms to undertake particular forms of education and awareness raising activities with sanctions for non-compliance.

Annex: Recommendations

Recommendation 1: To ensure that the requirements of legal clarity and precision are met, the Online Harms Bill should not include those forms of harmful content and activity which are criminal offences highlighted by the Law Commission as needing review. Such forms of content and activity should only be included in the Bill (or, if it has passed into law, added to the Act) once the Law Commission has completed its review of communications offences and the relevant criminal offences subsequently amended in line with any recommendations.

Recommendation 2: The Online Harms Bill should explicitly provide that all online platforms must take sufficient measures to safeguard the right to freedom of expression when complying with the duty of care and any other obligations under the Bill, such as complying with codes of practice. This would help ensure an equivalent level of protection to that provided by section 6 of the HRA 1998.

Recommendation 3: Before any statutory duty of care or other legal obligations come into force, the Online Harms Bill should require the regulatory body to produce detailed guidance to all online platforms on how to fulfil any obligations in a way which will fully protect the right to freedom of expression. Online platforms should be given sufficient time, once such guidance has been published, to adopt the necessary policies and processes which will allow them to do so, before any duty of care or other legal obligations will start to apply. This period of time should be no less than six months.

Recommendation 4: The guidance should be sufficiently detailed and provide specific guidance in relation to each of the particular forms of harms listed in the Online Harms Bill, and be tailored for different sizes of online platforms, and different types of online platforms. The guidance should be developed in collaboration with other stakeholders with relevant expertise, including civil society organisations and the Equality and Human Rights Commission. The guidance should be published in draft form and open to consultation, with sufficient time provided for feedback to be received and incorporated.

Recommendation 5: The Online Harms Bill should make clear that any online platform within scope has, at any time, the right to require further guidance from the regulatory body where the platform reasonably believes that fulfilment of the duty of care or any other legal obligations would undermine their ability to safeguard the right to freedom of expression.

Recommendation 6: The Online Harms Bill should also explicitly provide that in relation to any potential enforcement proceedings or action taken under the Bill, whether by the regulator or another body, an online platform is able to argue that, with respect to the alleged non-compliance, it was reasonably acting in accordance with its duty to take sufficient measures to safeguard the right to freedom of expression. Where such an argument is raised by an online platform, the regulator or other body enforcing the Bill should be required to reconsider its enforcement proceedings or action, and withdraw them if the online platform was in fact reasonably acting in accordance with its requirement.

Recommendation 7: Unless the government proposes that the "harms with a less clear definition" be prohibited generally and clearly defined in the Online Harms Bill or other legislation (a proposal which we would not support), then, ideally, the Online Harms Bill should not contain any such forms of "harmful content or activity". They should instead be dealt with by a separate regulatory framework and other measures, distinct from those proposed in the White Paper.

Recommendation 8: If "harms with a less clear definition" are not to be addressed through a separate regulatory framework, distinct from that proposed in the White Paper, then the Online Harms Bill should explicitly state that the duty of care does not require online

platforms to remove, restrict or moderate such forms of "harmful content or activity" and that the duty of care can be complied with through other actions (a "modified duty of care").

Recommendation 9: All forms of "harmful content or activity" that are to be addressed through the duty of care should be specified in the Online Harms Bill itself. They should also all be clearly and precisely defined in the Bill itself, or the Bill should make reference to other pieces of legislation which set out clear and precise definitions.

Recommendation 10: The regulatory body should not be given any power to introduce further forms of "harmful content or activity" or to require companies within scope to take any action in relation to any forms of "harmful content or activity" which are not specified in the Bill. The regulatory body could, of course, be given the power to make recommendations to the government that certain forms of "harmful content or activity" should be added to the Act, once passed.

Recommendation 11: If further forms of "harmful content or activity" are to be added to the Act, once passed, this should be done ideally via primary legislation. Alternatively, but less satisfactorily, this could be done via secondary legislation subject to the affirmative procedure. If secondary legislation is to be used to amend the list of forms of "harmful content or activity", the Online Harms Bill should require the government to consult beforehand on the particular forms that it is considering including. The Bill should explicitly state, however, that if further forms of "harmful content or activity" are added that are not already prohibited generally, whether through criminal law or otherwise, then the modified duty of care (**Recommendation 8**) will apply.

Recommendation 12: We do not believe that a "duty of care" is an appropriate regulatory model to tackle online harms. That being said, assuming that no other model than a "duty of care" is being considered by the government, the Online Harms Bill should explicitly make clear that the "duty of care" under the Bill is not comparable with, and should not be understood or interpreted in a like manner as, any other existing duty of care, whether found in other legislation or the common law.

Recommendation 13: The Online Harms Bill should explicitly state that compliance with the duty of care does not require, and should not be interpreted by the regulatory body or any court as generally requiring:

- Companies within scope to filter content at the point of upload, generation or sharing;
- Companies within scope either to generally or proactively monitor content; or
- Companies within scope to use artificial intelligence or other forms of automated decision-making.

Recommendation 14: The Online Harms Bill should explicitly state that if codes of practice are to include any measures set out in **Recommendation 13**, such measures are only required in relation to forms of harmful content or activity which are illegal and where the filtering or proactive identification is of copies of content which have already been identified by a human as illegal.

Recommendation 15: The Online Harms Bill should explicitly require the regulatory body to publish any codes of practice in draft form and to consult upon them before a final code of practice is adopted.

Recommendation 16: In addition to any general duty upon the regulatory body to protect and respect the rights to freedom of expression and privacy, the Online Harms Bill should further explicitly require the regulatory body to undertake a human rights impact assessment, which includes consideration of potential impacts upon these rights, when developing or

revising any code of practice. This human rights impact assessment should be published alongside the draft code of practice.

Recommendation 17: The Online Harms Bill should require all draft codes of practice should be scrutinised by the Joint Committee on Human Rights before any final code of practice is adopted by the regulatory body. Section 120(6) of the Coroners and Justice Act 2009 provides a precedent for this form of scrutiny, providing that draft sentencing guidelines developed by the Sentencing Council must be scrutinised by the Justice Committee of the House of Commons.

Recommendation 18: The Online Harms Bill should explicitly set out the issues which must be included in any mandatory transparency reporting templates and these should include those issues set out above, namely:

- Details on how the company develops its terms of service which touch upon content moderation, including if and how these are revised, and how external stakeholders are involved in any development and revision processes;
- Details on how the company enforces its terms of service which touch upon content moderation, including the publication of any internal enforcement policies or guidance, and details on the number of moderators and how they are trained and supported;
- Details on how the company makes decisions over the legality of content which is reported, where it is not prohibited by its own terms of service;
- Details of any demands or requests that the company receives from law enforcement agencies, courts or other public bodies for the removal or moderation of content and the company's responses;
- Details on any use of automated processes such as artificial intelligence to identify or moderate content, to what extent human moderation is involved in such circumstances, and what safeguards are in place to prevent inappropriate identification or moderation;
- Details on any use of automated processes such as artificial intelligence to filter or curate the content that an individual user sees, or the order in which they see it;
- Details on any processes in place which ensure users are informed when content they have posted or shared is moderated in any way, how they can challenge that decision, and how reviews are considered; and
- Any further information on how the company fulfils its responsibility under the UN Guiding Principles to respect the right to freedom of expression.

Recommendation 19: The Online Harms Bill should also provide the regulator with discretion to consider any further issues, to ask further questions to some or all platforms, as well as to determine the precise format and wording contained within the template.

Recommendation 20: The Online Harms Bill should explicitly state that any mandatory transparency reporting templates should be published by the regulator in draft form and be subject to consultation before a final template is adopted.

Recommendation 21: The Online Harms Bill should provide for the involvement of the Equality and Human Rights Commission in the work of the new regulatory body, and in particular the enforcement of its duties and functions, in order to determine impacts upon freedom of expression and privacy.

Recommendation 22: The regulatory body should 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.

Recommendation 23: In addition to the above, the Online Harms Bill should explicitly state that protecting and respecting the rights to freedom of expression and privacy is one of the regulatory body's statutory duties.

Recommendation 24: The regulatory body should be fully independent from government and political direction from government in all aspects, including the development and enforcement of its codes of practice.

Recommendation 25: The regulatory body should have a clear research and evidence-gathering function, and this should inform all of its work in developing codes of practice and undertaking its enforcement powers. This function could be mirrored on the Food Standards Agency's function under section 8 the Food Standards Act 1999 which provides that:

- The FSA has the function of "obtaining, compiling and keeping under review information about matters connected with food safety and other interests of consumers in relation to food";
- This function includes "monitoring developments in science, technology and other fields of knowledge" relating to the above, and "carrying out, commissioning or co-ordinating research" on them; and
- The FSA should carry out that function "with a view to ensuring that the Agency has sufficient information to enable it to take informed decisions and to carry out its other functions effectively".

A similar function exists in relation to OfCom under sections 14 to 16 of the Communications Act 2003.

Recommendation 26: The government should consider further means by which relevant expertise, including on human rights, informs and reviews the work of the regulatory body. In addition to a statutory function to undertake research to inform its work, the Online Harms Bill could, for example, require the regulatory body to establish a standing advisory committee on human rights to inform and review the work of the regulatory body. A comparable requirement exists in relation to OfCom which, under section 21 of the Communications Act 2003, is required to establish an advisory committee on elderly and disabled persons.

Recommendation 27: The Online Harms Bill should require the regulatory body to report annually on the exercise of its functions. This annual report should include an assessment of how the body has complied with its duty to protecting and respecting the rights to freedom of expression and privacy (**Recommendation 23**).

Recommendation 28: The Online Harms Bill should explicitly provide that the regulatory body's power to issue a civil fine cannot be exercised unless and until the other three "core powers" have been exercised and there has been a failure to comply with them. This would ensure that online platforms have sufficient opportunities to be aware of concerns and to respond to them before a risk of a fine materialises.

Recommendation 29: The Online Harms Bill should explicitly provide that the regulator may not issue a civil fine of an amount with would be disproportionate taking into account the size and resources of the online platform, and the level of harm or potential harm caused as a result of the online platform's non-compliance with its legal obligations.

Recommendation 30: The Online Harms Bill should not contain any of the further means of enforcement proposed in the White Paper, namely the imposition of civil or criminal liability on senior managers of online platforms, compelling ISPs to block websites, and disrupting business activities.

Recommendation 31: The Online Harms Bill should put Articles 17 to 19 of the Electronic Commerce (EC Directive) Regulations 2002 into primary legislation. This would, as is recommended by Recommendation CM/Rec(2018)2, make clear that online platforms 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.

Recommendation 32: The Online Harms Bill should be published in draft and subjected to pre-legislative scrutiny before a final version of the Bill is presented to Parliament.

Recommendation 33: The Online Harms Bill should require online platforms to provide opportunities for individuals to challenge content moderation decisions which they consider to amount to a restriction on their right to freedom of expression. Such opportunities should be equivalent to those provided to individuals to raise concerns about specific pieces of harmful content or activity, and/or breaches of the duty of care. These opportunities should include, at a minimum, (i) requiring affected users to be informed by the platform of content that has been flagged for removal, restriction or moderation; (ii) requiring an opportunity for that user to be able to input into the moderation process; and (iii) requiring platforms to introduce independent appeal mechanisms for affected users to challenge decisions.

Recommendation 34: Given the significant impact of the regulatory body upon the right to freedom of expression, the Chief Executive and other relevant members of staff at the regulatory body should be required to report annually before the Joint Committee on Human Rights, as are the Chief Executives of many other public bodies before other parliamentary committees.

Recommendation 35: The Online Harms Bill should define the types of companies within scope narrowly, so as to include only those types of companies where there is clear evidence of their having caused or facilitated harm via their online platforms and services.

Recommendation 36: In order to allow some flexibility, the Online Harms Bill should also provide the power to the government to develop secondary legislation which sets out whether particular companies, or types of companies, are within or outside of the scope of that definition. Such secondary legislation should be subject to the affirmative procedure. The Online Harms Bill should also require the government to consult beforehand on the companies, or particular types of companies, that it is considering including in secondary legislation.

Recommendation 37: If the definition of the types of companies within scope in the Act, once passed, is to be amended, this should be done ideally via primary legislation. Alternatively, but less satisfactorily, this could be done via secondary legislation subject to the affirmative procedure. If secondary legislation is to be used to amend the definition, the Bill should require the government to consult beforehand on the new definition that it is considering.

Recommendation 38: The Online Harms Bill should explicitly provide that, regardless of its definition of companies within scope, the regulatory framework does not apply with respect to (i) companies who only provide services which are protected by end-to-end encryption, and (ii) companies who provide a range of services, which includes at least one which is protected by end-to-end encryption, in respect of those services which are protected by end-to-end encryption.

Recommendation 39: The Online Harms Bill should explicitly require the regulatory body to undertake an impact assessment for each code of practice it publishes. This impact assessment should be published alongside a draft version of the code of practice. It should set

out an evidence-based assumption of the potential costs to companies of compliance with the code of practice as well as the potential benefits.

Recommendation 40: The Online Harms Bill should explicitly require the regulatory body to develop a regulatory action policy or similar guidance, which contains details on how the regulatory body will exercise its functions in a proportionate manner. The Online Harms Bill should explicitly provide that the regulatory body cannot undertake any regulatory enforcement until its final policy has been published.

This requirement could be mirrored on the Information Commissioner's Office's (ICO) requirement under sections 160 and 161 the Data Protection 2018 which provides that:

- The ICO must produce and publish guidance about how it proposes to exercise its enforcement functions, including the factors that it will take into account when deciding whether and how to exercise them; and
- The first version of the guidance must be submitted to the Secretary of State and laid before Parliament.

A similar, albeit more narrow, requirement exists in relation to OfCom where section 392 of the Communications Act 2003 provides that:

- OfCom must prepare and publish a statement containing the guidelines they propose to follow in determining the amount of penalties imposed by them under provisions contained in the Act; and
- Before publishing a statement or revised statement, OfCom must consult both the Secretary of State, and any other persons as they consider appropriate, about the guidelines they are proposing to include in the statement.

Recommendation 41: In addition to the above, the Online Harms Bill should require the regulatory body to publish its proposed regulatory action policy in draft form and for it to be open to consultation prior to submission to the Secretary of State.

Recommendation 42: The Online Harms Bill should explicitly provide that a company is able to use a statutory mechanism to appeal against a decision of the regulatory body wherever it considers that the decision was based on an error of fact, was wrong in law, was unlawful by virtue of section 6(1) of the Human Rights Act 1998, or a combination of the three.

Recommendation 43: The Online Harms Bill should ensure that a "decision" of the regulator includes, at a minimum, (i) a decision to publish a particular code of practice and (ii) a decision to undertake a particular action of enforcement.

Recommendation 44: The Online Harms Bill should explicitly provide for the regulatory body to have the power to require information from online platforms on their education and awareness raising activities. The Bill should not, however, provide for any powers that go beyond this, such as the power to require online platforms to undertake particular forms of education and awareness raising activities with sanctions for non-compliance.