Evidence Submission - Online safety and online harms

HATE SPEECH AND THE ONLINE SAFETY BILL: ENSURING CONSISTENCY WITH CORE INTERNATIONAL HUMAN RIGHTS INSTRUMENTS

Dr Talita Dias*

I thank the Digital, Culture, Media and Sport Sub-committee on Online Harms and Disinformation for this opportunity to submit my views on the issue of online safety and online harms. I am the Shaw Foundation Research Fellow in Law at Jesus College, University of Oxford, working alongside the Oxford Institute for Ethics Law and Armed Conflict (ELAC), a research institute dedicated to strengthening the rule of law across international borders. My research focuses on the application of international law to digital technologies, and I am one of the authors of the Oxford Statements on International Law Protections in Cyberspace. My latest article, published in the European Journal of International Law, looks at states' due diligence obligations in the cyber context. The present submission draws on my ongoing research into the regulation of online hate speech under international law, which is of great relevance to the upcoming Parliamentary scrutiny of the Online Safety Bill. I am grateful to Sahil Thapa for his research assistance in preparing this piece.

Executive summary

- As it currently stands, the Online Safety Bill is not fully aligned with the United Kingdom's obligations to protect individuals from violence and discrimination arising from certain online hate speech acts, as well as to safeguard users' freedom of expression under core international human rights instruments.
- Omissions regarding the types of *content* falling within the scope of Sections 41, 45 and 46 of the Bill should be addressed to give full effect to those international obligations. Notably, the Bill ought to distinguish between criminal and non-criminal speech acts falling within the category of 'illegal content', clearly define what online hate speech acts are illegal, and further specify the definition of content harmful to adults and children.
- Omissions concerning the types of *measures* that in-scope service providers must adopt to discharge their safety duties under Sections 9-11 and 21-22 of the Bill should be remedied to afford the necessary protection to children and adults against online hate speech whilst giving providers and users sufficient notice of limitations to relevant speech acts. In particular, the Bill ought to lay down and clearly define what restrictive measures in addition to content takedowns providers may or must implement to discharge their safety duties.
- Clear definitions of both speech acts and restrictive measures, in line with the requirements of legitimacy, legality, necessity and proportionality, are the only way to ensure that victims are protected against discrimination, violence and harm, whilst safeguarding users' right to freedom of expression.

*talita.desouzadias@jesus.ox.ac.uk.

1. Introduction: Scope and format of the present submission

As its title suggests, this submission is limited to assessing the extent to which the Online Safety Bill's proposed approach to tackling different forms of online hate speech is consistent with certain core international human rights treaties.[2] This focus is justified by three principal reasons. First and foremost, the United Kingdom (UK) is a party to those treaties.[3] Second, those instruments provide a universal, comprehensive and robust framework to tackle online hate speech whilst protecting freedom of expression, privacy and other fundamental rights. And striking the right balance between online safety and other fundamental human rights remains the most pressing challenge facing Parliamentary scrutiny of the Bill. Third, while many reactions to the Bill have assessed its conformity with the European Convention on Human Rights, the UK Human Rights Act 1998, and fundamental freedoms under the English common law,[4] few have looked at it from the perspective of core human rights treaties.[5]

With this focus in mind, this submission proposes to answer the following questions listed in the Committee's Call for Evidence:

- Is it necessary to have an **explicit definition and process for determining harm** to children and adults in the Online Safety Bill, and what should it be?
- Does the draft Bill focus enough on the **ways** tech companies could be encouraged to consider safety and/or the risk of harm in platform design and the **systems and processes** that they put in place?
- What are the key **omissions** to the draft Bill, such as a general safety duty or powers to deal with urgent security threats, and (how) could they be practically included **without compromising rights such as freedom of expression**?

Given their significant overlap, these questions are addressed together in a two-part analysis of the Bill's key omissions when it comes to tackling online hate speech in line with core human rights treaties. Specifically, Section 2 looks at omissions regarding the definitions of the various types of content that fall within the scope of the Bill, namely, illegal content (Sections 43 and 44) and content that is harmful to children and adults (Sections 45 and 46). Next, Section 3 assesses omissions concerning measures that in-scope service providers are required or permitted to adopt to tackle different types of online hate speech pursuant to their respective safety duties (Sections 9-11 and 21-22).

2. Omissions regarding the definition of illegal and harmful content: An insufficiently granular approach

'Hate speech' *as such* is not a specific legal concept featuring in international human treaties.[6] In common parlance, it has been broadly defined as any expression of hatred, opprobrium, enmity, detestation, or dehumanisation of an individual or group identified by a protected characteristic,[7] i.e. race, colour, sex, language, religion, political or other opinion, national or social origin,

---

[2] See United Nations Human Rights, 'The Core International Human Rights Instruments and their monitoring bodies', accessed 28 August 2021.
[3] United Nations Human Rights, 'Status of Ratification', accessed 28 August 2021.
[4] See, e.g., Caroline Elsom, 'Safety without Censorship: A better way to tackle online harms', Centre for Policy Studies, 27 September 2020, at 22; Index on Censorship, 'Right to Type: How the "Duty of Care" model lacks evidence and will damage free speech', 17 June 2021, at 2 and 14; Open Rights Group, 'Written evidence (FEO0091)', House of Lords Communications and Digital Committee inquiry into Freedom of Expression Online, 15 January 2021, paras 16-17; Timothy Pinto, 'Online Safety Bill – freedom of expression and privacy, journalistic content, and content of democratic importance', 30 July 2021.
[5] See, e.g., Open Rights Group, 'Save Online Speech Coalition Launches', 11 March 2021.
[6] Human Rights Council, A/74/486, para 1.
[7] ARTICLE 19, 'Hate Speech Explained: A Toolkit', 2015, at 9-10.

property, birth or another status.[8] In international human rights law, hate speech is best seen as an umbrella term encompassing a wide variety of speech acts that may have distinct legal implications. Given the variety of such acts and their likely effects, the key challenge of regulating hate speech and other speech acts under international human rights law lies in striking the appropriate balance between freedom of expression and other protected rights or interests, such as the right to security, bodily integrity, non-discrimination, health and reputation.

In the online environment, this challenge is compounded by the speed, scale and directness with which content is disseminated by individual users on the Internet. On the one hand, information and communications technologies have massively increased opportunities for expressing one's views and receiving information freely, as well as the exercise of other individual freedoms so dependent, such as the rights to freedom of opinion, to participate in democratic processes, and to protest. On the other hand, the pervasiveness of the Internet may also amplify the negative impact of hate speech and other harmful acts, leading to greater hostility, division, and violence in societies. Numerous examples of such impact can be found in the UK and abroad. Suffice it to note the landscape of online hate speech preceding and following the murder of Jo Cox MP by a white supremacist in 2016.[9]

Despite those challenges, and the difficulty of striking the right balance between freedom of expression and protection from harm and discrimination, the core international human rights treaties, particularly the International Covenant on Civil and Political Rights (ICCPR)[10] and the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD),[11] provide a universally accepted, comprehensive and robust legal framework to tackle hate speech online and offline. The cornerstone of this legal framework is Article 19 of the ICCPR, which protects individuals' fundamental rights to the freedoms of opinion and expression, including the right to seek, impart and receive information and ideas of all forms and kinds by any means, whether offline or online. This right is essential in any democratic society, particularly for the protection of vulnerable groups themselves. As such, it covers even the most shocking or offensive forms of expression, such as harsh criticism of governments and religious doctrines, tenets or leaders.[12]

However, as is well-known, freedom of expression under Article 19(2) ICCPR and in other human rights instruments is not absolute. The first of those limitations is provided for in Article 19(3) ICCPR, which *entitles* states to *restrict* freedom of expression by law whenever necessary (and proportionate) to respect the reputations of others or to protect national security, public order, public health or morals. The second limitation to freedom of expression is found in Article 20 ICCPR, which *requires* states to *prohibit* by law any propaganda for war (para 1) and any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence (para 2). This provision embodies the right of individuals to be free from hatred or incitement to certain forms of discrimination, in line with Article 26 ICCPR.[13] As noted by the Human Rights Committee (HRC) in its General Comment No. 11:

---

[8] Article 26 ICCPR.

[9] See e.g., Article 19, 'United Kingdom (England and Wales): Responding to 'hate speech'', Country Report, 2018, at 4 and 7; Ian Cobain, Nazia Parveen and Matthew Taylor, 'The slow-burning hatred that led Thomas Mair to murder Jo Cox', *The Guardian*, 23 November 2016; Katie Forster, 'Jo Cox death: Call for violent threats towards female MPs to be taken more seriously', *The Independent*, 17 June 2016; 'Research finds MP Jo Cox's murder was followed by 50,000 tweets celebrating her death', *Birmingham City University News*, 28 November 2016.

[10] Adopted on 6 December 1966, 999 UNTS 171.

[11] Adopted on 21 December 1965, 660 UNTS 195.

[12] See Human Rights Committee (HRC), 'General comment No. 34 - Article 19: Freedoms of opinion and expression', CCPR/C/GC/34, 12 September 2011, para 48.

[13] See HRC, *Rabbae v The Netherlands* (2017) CCPR/C/117/D/2124/2011, para 10.4; HRC, *Faurisson v France*, CCPR/C/58/D/550/1993 (1996), paras 4 and 10.

> For article 20 to become fully effective there ought to be a law making it clear that propaganda and advocacy as described therein are contrary to public policy and providing for an appropriate sanction in case of violation. The Committee, therefore, believes that States parties which have not yet done so should take the measures necessary to fulfil the obligations contained in article 20, and should themselves refrain from any such propaganda or advocacy.[14]

Article 4 ICERD complements Article 20(2) ICCPR by requiring states parties to 'condemn all propaganda and all organizations which are based on ideas or theories of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form and undertake to adopt immediate and positive measures designed to eradicate all incitement to, or acts of, such discrimination', with due regard to the freedoms of opinion and expression. This includes an obligation to 'declare an *offence* punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin, and also the provision of any assistance to racist activities, including the financing thereof.'[15] According to General Recommendation No. 35 of the Committee on the Elimination of Racial Discrimination (CERD):

> 7. Racist hate speech can take *many forms* and is not confined to explicitly racial remarks. As is the case with discrimination under article 1, speech attacking particular racial or ethnic groups may employ *indirect* language in order to disguise its targets and objectives. In line with their obligations under the Convention, States parties should give due attention to *all manifestations* of racist hate speech and take *effective* measures to combat them. The principles articulated in the present recommendation apply to racist hate speech, whether emanating from individuals or groups, in *whatever forms* it manifests itself, orally or in print, or disseminated through electronic media, *including the Internet and social networking sites*, as well as *non-verbal* forms of expression such as the display of racist *symbols*, *images* and *behaviour* at public gatherings, including sporting events.[16]

As aptly noted by the United Nations (UN) Special Rapporteur for Freedom of Expression, in light of the growing list of individual or group characteristics protected under international human rights law, prohibited speech under Article 20 ICCPR ought to be expanded to include incitement to discrimination, hostility or violence not only on the basis of race, colour, nationality or religion but also sex, sexual orientation, gender identity or intersex status, language, political or other opinion, social origin, property, birth or other status, including indigenous origin or identity, disability, migrant or refugee status.[17]

This means that, under international human rights law, states must follow a tiered or structured approach to the regulation of *any* content – including hate speech – whereby speech acts must be treated differently depending on the legal category under which they fall. These distinct legal categories are 1) **prohibited speech** (i.e., Articles 20 ICCPR and 4 ICERD); 2) **limited speech** (Article 19(3) ICCPR); and 3) **protected or free speech** (Article 19(2) ICCPR). In addition, *any* prohibition or limitation to freedom of expression, whether by criminal, civil or administrative means, must follow the general requirements listed in Article 19(3) ICCPR, namely, it must be established by law, for a legitimate purpose (including the purposes stated in Articles 20 ICCPR and 4 ICERD), and be necessary to achieve such a legitimate purpose.[18]

---

[14] HRC, 'General Comment No. 11: Prohibition of propaganda for war and inciting national, racial or religious hatred (Art. 20) : . 29/07/1983. CCPR General Comment No. 11. (General Comments)', 1983.
[15] Emphasis added.
[16] CERD, 'General recommendation No. 35: Combating racist hate speech', CERD/C/GC/35, 26 September 2013 (emphasis added).
[17] United Nations General Assembly, 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression', A/74/486, 9 October 2019, para 9.
[18] HRC, General Comment nº 34 (n 12), paras 50-52, CERD, General Recommendation No. 35 (n 16), paras 4, 19-20; A/74/486 (n 17), paras 12 and 16.

Necessity, in this context, has been understood as a two-part test requiring states to assess not only whether the limitation is the *least restrictive* means to fulfil the legitimate aim(s) but also whether the restriction to free speech is *proportionate* to the right or interest it aims to uphold.[19] In other words, the types of limitations to speech acts, e.g. criminalisation, prohibition, de-prioritisation, tagging, partial redaction, etc., must be carefully calibrated to the importance of the legitimate aim(s) protected, such as non-discrimination and public health.[20]

As a result, when prohibiting and limiting speech domestically, states must carefully distinguish between a) speech acts constituting **criminal offences**; b) content that is not criminal but is **prohibited** and thus sanctioned **by civil or administrative law**; and c) expressions that neither give rise to criminal or civil sanctions and are thus **unsanctioned**, however repugnant they may be.[21] Although the categories of prohibited, limited, and protected speech need not squarely correspond to criminal acts, civil wrongs, and unsanctioned speech, respectively, proportionality requires that criminal punishment be reserved to only the most serious forms of hate speech. These include instances of incitement to hatred constituting 'the most severe and deeply felt form of opprobrium', taking into account the context, the speaker, any intent, the content, form and extent of the speech act, as well as its likelihood of harm.[22]

In sum, international human rights law *requires* states to declare as **offences** punishable by law only **the most serious types of incitement to hostility, violence, or discrimination** on the basis of race, religion, nationality, and other internationally protected characteristics, as well as the dissemination of ideas of **racial superiority or hatred**. Less serious types of incitement to hostility, violence, or discrimination on those same grounds *must* be **prohibited** by less severe means, such as **civil or administrative sanctions**. Other types of hate speech and other forms of harmful expression *may* be limited to respect the rights or reputations of others or to protect national security, public order, public health, or morals. Yet any such limitation, whether by criminal, civil or administrative sanctions, must be provided for in clear, accessible, and foreseeable laws, as well as necessary and proportionate to achieve its stated aim. Whatever speech acts fall outside the scope of such limitations must be **protected and thus remain unsanctioned**.

In its current state, the Online Safety Bill does not entirely mirror those international legal standards, or at least fails to do so in a sufficiently clear, accessible and foreseeable manner. As explained in the following sections, those omissions are likely due to an **excessive emphasis on imposing a general duty of care on service providers** with respect to illegal or harmful content, **to the detriment of defining the actual *conduct* which is the object of such a duty**. In the same vein, the Bill focuses too much on sanctions to be applied by the regulator on *platforms themselves*, as opposed to the measures that may or must be applied by platforms to sanction or protect *individual users*. Simply put, the Bill imposes on service providers a form of intermediary liability without clearly defining the underlying user-generated speech acts that may give rise thereto.

a. Prohibited Speech

When it comes to prohibited speech, the Bill's definition of 'illegal content' in Section 41 not only conflates criminal, civil, and administrative wrongs but also fails to spell these out, as required by

---

[19] HRC, General Comment nº 34 (n 12), paras 22, 33; A/74/486 (n 17), paras 6(c), 51; United Nations Human Rights Council, 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression', A/HRC/38/35, 6 April 2018, paras 7, 28, 44-45.

[20] A/74/486 (n 17), para 51.

[21] Human Rights Council, 'Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred', Appendix, Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, A/HRC/22/17/Add.4, 11 January 2013, paras 12 and 20.

[22] A/HRC/22/17/Add.4, *ibid*, para 29.

Articles 19(3) and 20 ICCPR. And by simply relying on existing laws, the Bill suffers from the same gaps in protection found therein. These include failing to prohibit racist groups, in line with Article 4(b) ICERD and to punish incitement to hostility, violence, or discrimination on grounds *other than* race, religion and sexual orientation, in accordance with Article 20 ICCPR, read together with Article 26 ICCPR (on the right to non-discrimination), Article 2 of the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW)[23] (on the duty to protect women from all forms of discrimination), and Article 4(e) of the Convention on the Rights of Persons with Disabilities (CRPD)[24] (on the obligation to take all appropriate measures to eliminate discrimination on the basis of disability by any person, organization or private enterprise).

In more detail, when defining '**illegal content**', the Bill simply refers to a '**relevant offence**' (Section 41(2)), which means either: i) an existing 'terrorism offence' as specified in Schedule 2 (Section 42); ii) an existing child sexual exploitation or abuse offence, as specified in Schedule 3 (Section 43); iii) an 'offence' to be specified or described by the Secretary of State in secondary legislation (described as 'priority illegal content' and further regulated in Section 44), or iv) another offence of which the victim or intended victim is an individual (or individuals). The problem lies in the latter two categories: 'offences' to be determined by the Secretary of State and other (existing) offences.

On the one hand, it is unclear whether the 'offences' to be specified or defined in secondary legislation by the Secretary of State must amount to existing crimes, civil or administrative wrongs, or whether the Secretary of State is effectively empowered to define new such crimes or wrongs.

If the former, **the Bill must clearly indicate which among those existing 'offences' may fall within the scope of 'priority illegal content'**, such as the criminal offences of stirring up hatred on racial, religious or sexual grounds, laid down in Parts III and Part 3A of the Public Order Act, or the offence of improper use of public electronic communications network, listed in section 127 of the Communications Act 2003. Likewise, if the meaning of 'offences' goes beyond strictly criminal acts punishable by law, then the Bill must clearly distinguish between content subject to criminal, civil and administrative sanctions, in accordance with the requirements of legality, necessity and proportionality in Article 19(3) ICCPR.

If it is the case that the Secretary of State is entitled to create *new* criminal offences, civil or administrative wrongs beyond existing laws when enacting secondary legislation pursuant to Section 44 of the Bill, **then there is a clear legality issue**. This is because, in accordance with Article 19(3) ICCPR, any limitation (criminal, civil or administrative) to otherwise free speech must be **provided by law**, following the necessary Parliamentary scrutiny, in a way that is clear, foreseeable and accessible. More fundamentally, Article 15 ICCPR requires *criminal* offences to be made punishable *by law*, as opposed to secondary legislation. In short, it is for Parliament to decide which speech acts may amount to criminal offences, civil or administrative wrongs, not the Secretary of State.

On the other hand, the Bill's reference to 'other offences' of which the victim is an individual in Section 41(4)(d) lacks the same clarity found in the definitions of relevant terrorist and child abuse offences, which have been spelt out in Schedules 2 and 3. Again, it is unclear which existing offences other than terrorism and child abuse fall within the scope of Section 41(4)(d) of the Bill and whether these are limited to crimes *strictu sensu* or include civil and administrative wrongs as well. If 'offences' include criminal, civil and administrative wrongs, these must be clearly *defined* and *distinguished* in line with Article 19(3) ICCPR. This is to provide sufficient notice to both platforms and individual users about the severity of different types of online content as well as the different types of measures they may be subject to, according to their degree of seriousness. Even

[23] Adopted on 18 December 1979, 1249 UNTS 13.
[24] Adopted on 24 January 2007, A/RES/61/106.

if criminal offences and civil or administrative wrongs are to be found in existing laws and regulations, *new restrictions* are being imposed by the Bill to a subset of those existing criminal or wrongful acts in the online environment. Accordingly, the specific conduct which is subject to those new restrictions must be spelt out, whether or not they fall within the broader scope of existing laws.

In England, Wales and Scotland, hate speech crimes are dealt with in Parts III and 3A of the Public Order Act 1986, which criminalises 'acts intended or likely to stir up hatred' on the basis of race, religion or sexual orientation. These are i) the use of words or behaviour or display of written material; ii) publishing or distributing written material; iii) the public performance of a play; iv) distributing, showing or playing a recording; v) broadcasting or including a programme in a cable programme service; and vi) possessing racially inflammatory material, where a) the material is threatening, abusive or insulting, *and* b) either there is an intention to thereby stir up racial hatred or, having regard to all the circumstances, racial hatred is likely to be stirred up thereby. Such offences are punished by imprisonment of up to seven years and/or a fine. Similar offences are found in Section 3 of Northern Ireland's Public Order Act 1987, as amended in 2001 and 2004 to include, alongside stirring racial hatred, incitement on the basis of disability, religious belief and sexual orientation Other types of hate speech acts, including incitement on other grounds and racist propaganda short-of-incitement, may fall within the scope of the broader offences defined in Section 1 of the Malicious Communications Act 1988 ('sending letters etc. with intent to cause distress or anxiety') and Section 127 of the Communications Act 2003 ('improper use of public electronic communications network').

As with the terrorism and child abuse offences, **Schedules should clearly *specify* which other existing 'offences' are relevant for Section 41(4)(d) of the Bill, and clearly *distinguish* between criminal, civil and administrative wrongs**. To bring further clarity and specificity to the definition of those offences and *how* they apply in the online environment, **Schedules should list examples of particularly concerning or recurrent online content** falling with relevant definitions, such as incitement to violence, discrimination or hostility against racial and religious groups, women, persons with disability and members of the LGBTQ+ community. **Clear definitions**, in line with the requirements of **legitimacy, legality, necessity and proportionality**, are the only way to ensure that victims are protected against discrimination, violence and harm, whilst safeguarding users' right to freedom of expression. Incidents involving **prohibited speech** should also be characterised as '**urgent security threats**' given their proximity to actual violence, hostility or discrimination. This, together with the clear identification of corresponding measures as described in Section 3 below, should be the first step in dealing with those types of threats.

Including concrete examples would allow in-scope service providers and users to know with greater certainty that posts such as those inviting users to 'punish a [N-word]' with different forms of violence, following England's defeat in the Euro 2020 final,[25] amount to prohibited speech and the offence of stirring up racial hatred, thus falling neatly within the Bill's category of 'illegal content'. Likewise, the Bill should clarify that instances of online harassment and calls for violence against women, now widespread in the UK,[26] amount to criminal offences and should be dealt with the most stringent measures to tackle illegal content. Examples include online trolling, doxing, mobbing, sextortion, and revenge porn.[27] Particular targets of such forms of gender-based violence are female human rights defenders, journalists, politicians, bloggers, members of the LGBTQ+

[25] Tweet by David Lammy, 12 July 2021.
[26] Amnesty International, 'Online abuse of women widespread in UK', accessed on 28 August 2021.
[27] United Nations Human Rights Council, 'Report of the Special Rapporteur on violence against women, its causes and consequences on online violence against women and girls from a human rights perspective', A/HRC/38/47, 18 June 2018, paras 39-42. See also Amnesty International, 'Toxic Twitter - Women's Experiences of Violence and Abuse on Twitter – Chapter 3', accessed 28 August 2021.

community and disabled women, who should, accordingly, be granted corresponding levels of protection by online platforms.[28] Speech acts against persons with disabilities which are criminal or otherwise illegal, such as posts encouraging others to commit violence against disabled persons, should also be clarified to ensure that they are effectively dealt with by platforms, as well as reported to and prosecuted by relevant authorities.[29]

At the same time, the Bill has missed an opportunity to bring existing hate speech laws into full conformity with Article 4(b) ICERD, which requires states to prohibit by law racist groups or organisations. Such rules have been lacking in the UK since 1987.[30] Yet activities of racist groups have become even more frequent and organised in the online environment, as shown by the recent posts of white supremacist and neo-Nazi groups on Telegram following the Euro 2020 final.[31]

It is worth noting that the Bill's failure to clearly define prohibited types of speech is symptomatic of the UK's longstanding 'hands-off' approach to tackling hate speech, i.e. its failure to adopt the necessary legislation defining the types of content which may or must be limited by different media services. As such, neither victims of hate speech are sufficiently protected, nor are users and media outlets put sufficiently on notice as to what types of otherwise free speech may be censored.

In more detail, on several occasions, CERD has expressed concern over and called upon the UK to reconsider its restrictive interpretation of Article 4 ICERD,[32] according to which 'further legislative measures in the fields covered by sub-paragraphs (a), (b) and (c) of that article [are required] *only in so far as [a state party] may consider*'.[33] Concern has been especially borne out by 'statements by some public officials and media reports',[34] as well as 'the continuing virulent statements in the media that may adversely affect racial harmony and increase racial discrimination in the State party'.[35] Accordingly, CERD recommended that the UK 'adopt comprehensive measures to combat racist hate speech and xenophobic political discourse, including on the Internet', and to 'take effective measures to combat racist media coverage, taking into account the Committee's general recommendation No. 35 (2013) on combating racist hate speech.'[36]

Also 'concerned about the prevalence in the media and on the Internet of racist and xenophobic expressions that may amount to incitement to discrimination, hostility or violence', the UN Human Rights Committee stated that the UK:

---

[28] A/HRC/38/47, *ibid*, paras 25-29; International Center for Journalists, 'Online Attacks on Women Journalists Leading to 'Real World' Violence, New Research Shows', 25 November 2020.
[29] UK Parliament, 'Online abuse and the experience of disabled people', 22 January 2019, para 33.
[30] UN General Assembly, 'Report of the Committee on the Elimination of Racial Discrimination', A/42/18, 1987, para 703; UN General Assembly, 'Report of the Committee on the Elimination of Racial Discrimination', A/46/18, 1992, para 189.
[31] David Gilbert, 'England Players Suffer Racist Abuse and Threats on Neo-Nazi Telegram Channels', *Vice News*, 12 July 202.
[32] See, e.g., A/46/18 (n 30), para 189; CERD, 'Consideration of reports submitted by States parties under article 9 of the Convention: Concluding observations of the Committee on the Elimination of Racial Discrimination - United Kingdom of Great Britain and Northern Ireland', 14 September 2011, CERD/C/GBR/CO/18-20, para 11; CERD, 'Concluding observations on the combined twenty-first to twenty-third periodic reports of the United Kingdom of Great Britain and Northern Ireland', 3 October 2016, CERD/C/GBR/CO/21-23, paras 15 and 17.
[33] United Nations Treaty Collection, 'International Convention on the Elimination of All Forms of Racial Discrimination - Declarations and Reservations', accessed 28 August 2021.
[34] CERD, 'Consideration of reports submitted by States parties under article 9 of the Convention: concluding observations of the Committee on the Elimination of Racial Discrimination : United Kingdom of Great Britain and Northern Ireland', 10 December 2003, CERD/C/63/CO/11, para 12.
[35] CERD/C/GBR/CO/18-20 (n 32), para 11.
[36] CERD/C/GBR/CO/21-23 (n 32), para 16(d)-(e)

> should strengthen its efforts to prevent and eradicate all acts of racism and xenophobia, including in the mass media and on the Internet, in accordance with articles 19 and 20 of the Covenant and the Committee's general comment No. 34 (2011) on freedoms of opinion and expression.[37]

Following the recommendations from both human rights bodies, UK reported[38] the adoption of a hate crime action plan – the so-called *Action against Hate*.[39] Yet the action plan only applies to hate *crime* as defined in *existing* legislation. Non-criminal forms of prohibited hate speech continue to be undefined and unlimited by civil or administrative law in England.

b. Limited speech

The Bill's definition of 'content that is harmful' to children or adults (Sections 45 and 46) does not fully meet certain requirements for limiting speech under Article 19(3) ICCPR, namely, **legality, necessity, and proportionality**. This is so to the extent that the identification of such content is delegated to secondary legislation (Sections 45(2)(b)(i)(ii) and 46(2)(b)(i), defining 'priority harmful content') or service providers themselves, using a broad and vague definition of harm that takes into account the physical or psychological wellbeing of victims (Sections 45(2)(b)(iii)(3)-(9) and 46(2)(b)(ii)(3)-(11)). Likewise, however laudable and legitimate the stated aims of such definition may be, such as the protection of the rights and reputations of others, the Bill or its Explanatory Notes fail to explain i) why it is necessary to limit *legal but harmful* content by imposing a *general* duty of care on service providers; and ii) what exact *measures* may or must be *proportionately* adopted by such providers to address different types of harmful content.

As explained earlier, states are entitled to limit speech insofar as those limitations are provided by **clear, accessible and foreseeable laws**, and are **necessary and proportionate** to achieve their aims, in accordance with Article 19(3) ICCPR. Simply delegating the task of identifying limited speech acts (as 'priority harmful content') to the Secretary of State falls short of the legality requirement. For one thing, limitations to otherwise free speech must be laid down in laws enacted by Parliament, not the executive. For another, there are no guarantees that the Secretary of State's regulations will define limited content falling under the category of 'priority harmful content' in a clear, accessible, and foreseeable manner.

In the same vein, the definition of *other* types of legal but harmful content beyond those identified by the Secretary of State in secondary legislation is not only excessively broad but also vague. This definition revolves around the concept of *harm*, defined as the 'material risk of the content having, or indirectly having, a significant adverse physical or psychological impact on a child or adult of ordinary sensibilities, taking into account any of their known characteristics or group membership' (Sections 45(2)(iii)(3)-(8) and 46(2)(ii)(3)-(7), respectively). Notably, this definition includes content *indirectly* causing a risk of harm, i.e., causing the individual targeted to do, say or act in a way that would lead to serious physical or phycological impact or increases the likelihood of such an impact, such as acts of instigation or encouragement. The Bill then requires in-scope service providers to determine which types of content they have *reasonable grounds* to believe fall under this definition, taking into account the content's visibility as well as the ease and speed with which it may be disseminated (Sections 45(5) and 46(5)). In short, judgement calls as to whether harmful content which is not listed in secondary legislation may be considered harmful to children or adults are to be made by service providers, according to their own understanding of 'adverse physical or phycological impact'.

---

[37] Human Rights Committee, 'Concluding observations on the seventh periodic report of the United Kingdom of Great Britain and Northern Ireland', CCPR/C/GBR/CO/7, 17 August 2015, para 10.

[38] CERD, 'Concluding observations on the combined twenty-first to twenty-third periodic reports of the United Kingdom of Great Britain and Northern Ireland, Addendum, Information received from the United Kingdom of Great Britain and Northern Ireland on follow-up to the concluding observations', CERD/C/GBR/CO/21-23/Add.1, 17 October 2017, para 6.

[39] UK Government, 'Policy paper: Hate crime action plan 2016 to 2020', 26 July 2016.

To be sure, decisions about what types of content amount to limited speech are inherently difficult given linguistic, cultural, and contextual subtleties. Thus, states enjoy a significant margin of discretion when identifying such content, including hate speech. **However, it is for states, and Parliament more specifically – *not private entities* –, to decide which types of content, online or offline, may be subject to limitation in whatever media outlet**. As others have noted, a broad and vague definition of harmful content, relying solely on a subjective criterion of harm, coupled with a low evidentiary threshold and prohibitively high fines for failure to remove such content, will force platforms to err on the side of censorship when moderating content.[40]

This is why, when making difficult decisions about which types of speech may be subject to limitation, states must *not only* consider the degree of harm or risk caused on potential victims, along with the visibility and dissemination of the relevant content. Rather, to ensure that any decision to limit speech is necessary and proportionate, **several other factors must be taken into account.** These include a) the socio-historical context of the speech act, such as whether the same or similar content was used to spur violence in the past; b) the intention and position of the speaker, i.e., the more powerful or popular the speaker, the greater the likelihood that their speech will influence the audience's attitudes; c) the audience's resilience or susceptibility to act upon or be persuaded by hate speech; and d) the degree of hatred expressed, including the directness or vagueness of the speech act.[41]

Thus, to comply with the requirements of legality, necessity and proportionality under Article 19(3) ICCPR, **the Bill must further limit the definition of speech acts falling within the category of 'legal but harmful' content**. Although it is impossible to list and describe in detail all types of speech falling within this category, the Bill should **lay down additional parameters** for identifying legal but harmful content, such as the ones suggested above, whether this assessment is made by the Secretary of State in secondary legislation or service providers themselves. To ensure that individual users are made sufficiently aware of potential types of harmful content which may be subject to limitation, **the Bill should provide clear examples in a Schedule**.

These examples should include the use of known racial slurs and discriminatory symbols, such as the N-word and, in certain racially-charged contexts, the monkey and banana emojis, which have come to be seen as symbols of racism.[42] Likewise, the Bill should clarify the extent to which expressions of hatred asking individuals belonging to racial, ethnic or national groups to 'get out of [the] country' or 'go back' to a certain country amount to legal but harmful content. Such incidents have included posts targeting Asians[43] and Muslims[44] in the wake of the COVID-19 pandemic, as well as England's football players following the Euro 2020 final.[45] Online expressions of misogyny and sexism, even when legal, have also led to significant physical or psychological harm against women given their vulnerability in many social contexts. Similar types of online hateful rhetoric and harassment have been reported against persons with disabilities in the UK, such as the use of slurs and pejorative memes.[46]

---

[40] Index on Censorship (n 4), at 4 and 11.
[41] Dangerous Speech Project. 'Dangerous Speech: A Practical Guide', 2020; Susan Benesch, 'Dangerous Speech: A Proposal to Prevent Group Violence', 23 February 2013, at 2-6.
[42] Jeremy Burge, 'How the Monkey Emoji is Racist', *Emojipedia*, 12 July 2021.
[43] The Cybersmile Foundation, 'Online Hate Targeting Asian People Spikes as Coronavirus Crisis Deepens', accessed 28 August 2021.
[44] 'COVID-19 sparks online Islamophobia as fake news and racist memes are shared online, new research finds', *Birmingham City University News*, accessed 28 August 2021.
[45] Tweet by 'em', 11 July 2021.
[46] UK Parliament, 'Online abuse and the experience of disabled people', 22 January 2019, para 34-35. See also Leonard Cheshire, 'Online disability hate crimes soar 33%', 11 May 2019; Caleb Spencer, 'Disability hate crime: Rise in reports of online abuse', *BBC News*, 8 October 2020.

The Bill should make clear the extent to which these and other widely reported forms of hateful content are subject to providers' safety duties and any ensuing restrictions on speech. Otherwise, the definition of limited hate speech and the necessary and proportionate measures to constrain it under Article 19(3) ICCPR will be left entirely in the hands of tech companies' community standards or guidelines.

c. Protected speech

Lastly, the Bill does not make it clear that content that is not prohibited nor limited (whether as illegal or legal but harmful conduct) is thus protected, in line with Article 19(2) ICCPR. Instead, provision is simply made for the protection of journalistic and democratic content. This leaves all other types of perfectly legal and thus protected speech which may be offensive to some – such as artistic nudity, satire or harsh criticism of religious doctrines, tenets or leaders – at the mercy of service providers' general (and vaguely defined) duty to protect users' freedom of expression (Section 12(2)(a)). **Yet it is states – *not companies* – that are bound to *protect* individuals' right to seek, receive and impart information and ideas of all kinds, in line with Articles 2(1) and 19(2) ICCPR**. Under existing international law, *corporations* only have *voluntary* responsibilities to respect human rights.[47] Thus, it falls upon states to protect the human rights of those within their jurisdiction by, inter alia, regulating corporate activities that might infringe upon those rights, including social media companies and other Internet service providers.[48] To fulfil this positive or protective duty, states must not only require companies to exercise the necessary degree of care or diligence when providing a service to the public. They must also enforce their *own* positive obligation to exercise due diligence in preventing violations of freedom of expression by corporations and other private entities.[49]

3. Omissions concerning measures to tackle different types of online hate speech: Delegating the UK's own duties to private service providers

The requirements of legitimacy, legality, necessity and proportionality laid down in Article 19(3) ICCPR apply not only to the definition of prohibited and limited speech acts but also to the *measures* that states are required or permitted to adopt to tackle such acts.[50] As mentioned earlier, although states must prohibit incitement to violence, hostility and discrimination, as well as racist groups and propaganda to protect individuals' right to non-discrimination, any such prohibition must still be provided by law. Likewise, compliance with the requirements of necessity and proportionality means that only the most serious types of prohibited or limited speech may be criminalised and thus subject to criminal punishment and other significantly restrictive measures.

The same goes for less serious types of prohibited speech and acts of limited speech, such as expressions of hatred short-of-incitement. This means that any limitations to such speech acts, whether they include civil or administrative sanctions and any sort of content curation or censorship, must be laid down in clear, accessible, and foreseeable laws. While Articles 20 ICCPR and 4 ICERD assume that the prohibition of incitement, racist propaganda and groups is justified by the need to protect victims' rights to non-discrimination, limited speech acts must be specifically justified for a legitimate purpose listed in Article 19(3) ICCPR – to protect the rights or reputations of others, national security, public order, or morals. In the same vein, prohibited and limited speech

---

[47] See UN Human Rights Council, 'Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework', A/HRC/17/31, 16 June 2011, Principle 11.
[48] A/74/486 (n 17), para 41.
[49] Antonio Coco and Talita de Souza Dias, '"Cyber Due Diligence": A Patchwork of Protective Obligations in International Law', *European Journal of International Law* 2021, at 25-30.
[50] A/74/486 (n 17), para 31.

acts may only be subject to necessary and proportionate measures. As such, any limitations to those acts must be calibrated to the seriousness of the relevant content.

Yet the Online Safety Bill fails to meet the legitimacy, legality, necessity and proportionality tests when providing for measures to be implemented by service providers in accordance with their duties of care for different types of content.

Starting with 'illegal content', Sections 9 and 21 of the Bill require in-scope user-to-user services and search engines to 'take proportionate steps to mitigate and effectively manage the risks of harm to individuals.' According to Sections 9(3) and 21(3), these consist of 'proportionate systems and processes designed to '(a) minimise the presence of priority illegal content; (b) minimise the length of time for which priority illegal content is present; (c) minimise the dissemination of priority illegal content. **However, the Bill does not define *what actions* providers must take to 'minimise' the presence, exposure, and dissemination of 'priority illegal content'**. For instance, must providers warn users about the consequences of posting priority illegal content? Must they tweak their recommendation algorithms to minimise engagement with this type of content? May they redact content? We simply don't know. And this affects not only companies, which have little clarity as to what they are legally required (and could be fined for failing) do to but also users, who are left in the dark as to how their speech acts may be limited.

Moreover, for user-to-user services, whenever providers have knowledge of *any* type of *illegal* content, whether criminal or non-criminal, Section 9(3)(d) of the Bill lists *only one* available course of action: swift content takedowns. This approach is hardly proportionate. It lumps together prohibited and limited speech acts, as well as criminal and non-criminal ones, in one and the same category which is subject to the exact same set of measures. Further, neither the Bill nor its Explanatory Notes indicate the exact aim of such content takedowns or justify why they are necessary and proportionate to achieve any such aim. Granted, Section 9(4)-(5) of the Bill does require providers to 'specify in the terms of service how individuals are to be protected from illegal content', in a clear, accessible, and consistent manner. **However, under international law binding on the UK, it is the duty of states – *not private entities* – to ensure that any limitations to speech are clear, accessible, and foreseeable**. Worryingly, content takedowns, if not accompanied by effective measures to preserve relevant evidence, may hinder criminal investigations into speech offences, as well as civil or administrative processes with respect to non-criminal speech acts.

Although the regulation of content that is harmful to children is obviously justified by the need to protect the rights of children, Sections 10 and 22 of the Bill do not clearly indicate what measures providers may or must take to protect those rights. Again, Sections 10(2) and 22(2) of the Bill require user-to-user services and search engines to take proportionate steps to mitigate and effectively manage the risks and impact of harm to children in different age groups. But rather than clearly spelling out what those steps will entail for user-to-user services and their individual users, Section 10(3)-(4) of the Bill merely stipulates a general duty to prevent and protect children from encountering harmful content, as specified in providers' terms of service. In the same vein, Section 22(3) of the Bill simply imposes on search engines a general duty to minimise exposure of children to harmful content via search results, which must be clearly and publicly specified in statements outlining company policies to protect children, as per Section 22(4).

The lack of clarity around applicable measures and content restrictions is compounded for content that is harmful to adults. In this regard, user-to-user service providers are not directed to any particular type of measure or even generally required to adopt proportionate steps to mitigate and manage risks or impact on individuals. According to Section 11(2) of the Bill, such providers are simply required to specify in their terms of service how priority content and other content that is harmful to adults are to be dealt with by the service. This means that, when it comes to content that is harmful to adults, a category that potentially includes speech acts ranging from racial slurs

and misogynistic content to disinformation, the Bill leaves protective measures and limitations to speech entirely at the discretion of platforms. In the absence of a holistic approach to tackling online hate speech, certain measures, such as Instagram's Hidden Words Tool,[51] may have the unfortunate result of compelling victims to self-regulate rather than deterring or educating those responsible.

Two further omissions should be flagged out in respect of the measures covered by the Bill. First, there is no specific provision for algorithm auditing or review, whether by in-scope providers themselves or external bodies. There is simply a general reference to designing and assessing the service 'with a view to protecting United Kingdom users from harm, including with regard to […] algorithms used by the service', as part of the regulator's role to ensure that certain 'online safety objectives' are achieved (Section 30). Yet, it is no secret that platform recommendation algorithms[52] are geared towards engagement[53] and, thus, are in no small part responsible for the dissemination and amplification of hateful content. Thus, the Bill should include measures directly seeking to ensure that companies periodically review and adjust their recommendation algorithms.

Second, the Bill makes no provision for access to *judicial* remedies by either in-scope providers, content authors or addressees.[54] To be sure, Section 106 does empower 'eligible entities' to make complaints to the regulator (OFCOM) regarding services and/or conduct that presents a material risk of causing significant harm to users, members of the public or groups, significantly adversely affecting freedom of expression, causing significant unwarranted infringements of privacy, or otherwise having a significant adverse impact on users of the services, members of the public, or groups. However, Article 2(3) ICCPR requires states parties to ensure that individuals have access to judicial, administrative or legislative remedies, *and*, at the very least, to develop the possibilities of *judicial* remedy. If the legislator's intent is to provide an *additional* avenue for justice before OFCOM, without prejudice to existing judicial channels, it should have spelt that out.

To remedy those omissions, the following amendments to the Bill are recommended:

a) **Restrictive measures for illegal and harmful content should be specified in Sections 9-11 and 21-22**. For user-to-user services, these should include measures *other than* content takedowns, such as tagging or labelling less serious forms of illegal or harmful content. Similar measures introduced by Twitter,[55] Facebook[56] and other platforms to tackle COVID-19 dis- and misinformation have so far yielded positive results. They can contextualise speech acts without the need for automated or human content takedowns. Likewise, tagging or labelling has the benefit of exposing users to potentially relevant information whilst enabling platforms to take a clear stance against any type of illegal or harmful content.

b) **Requiring providers to ensure that their recommendation algorithms decrease the visibility of prohibited and limited content whilst increasing the visibility of and opportunities for counter speech**.[57] Counter speech is a powerful tool in the fight against

---

[51] Instagram, 'Introducing new tools to protect our community from abuse', 21 April 2021.
[52] Roger Chua, 'A simple way to explain the Recommendation Engine in AI', *Medium*, 26 June 2017; Google Cloud, 'Recommendations AI'.
[53] See Cathy O'Neil, *Weapons of Math Destruction* (Penguin Books, 2016), at 180-185; Access Now, 'Human Rights in the Age of Artificial Intelligence', 8 November 2018, at 16; Yaël Eisenstat, 'Dear Facebook, this is how you're breaking democracy', *TED*, August 2020,; Carole Cadwalladr, 'If you're not terrified about Facebook, you haven't been paying attention', *The Guardian*, 26 July 2020; Cathy O'Neil, 'TikTok's Algorithm Can't Be Trusted', *Bloomberg*, 21 September 2020.
[54] Index on Censorship (n 4), at 3.
[55] Kari Paul and others, 'Twitter targets Covid vaccine misinformation with labels and "strike" system', *The Guardian*, 1 March 2021.
[56] Elizabeth Culliford, 'Facebook to label all posts about COVID-19 vaccines', *Reuters*, 15 March 2021.
[57] A/74/486 (n 17), para 28.

online hate speech: it can empower users themselves to respond to and prevent hateful rhetoric without the need to censor content. This measure could be implemented by, inter alia, introducing 'dislike' buttons, such as those piloted by YouTube,[58] with highly disliked comments hidden from users' views.

c) **Requiring in-scope service providers to scale up their content moderation mechanisms** to cover difficult cases that can neither be conclusively dealt with by companies' automated or human moderators nor added to the caseload of a formal complaints mechanism in accordance with sections 15 and 24 of the Bill. This could be done by, inter alia, allowing especially vulnerable and visible users to nominate moderators for their pages.

d) **Clarifying that providers, affected users and members of the public remain entitled to a judicial remedy** with respect to services and acts falling within the scope of the Bill.

e) **Requiring providers to keep a record and preserve relevant evidence of criminal or otherwise wrongful speech acts, as well as promptly notifying such incidents to the police and other relevant authorities.** Such records should feed official statistics on online hate *speech* crimes in England, which are still lacking in the UK.[59] Prompt notifications are particularly important when dealing with **'urgent security threats'**, which should be notified to the authorities within 24h once time platforms become aware of any such threats. They should also increase the UK's slim record of convictions and prosecutions for 'stirring up hatred' and other hate speech crimes.[60]

Conclusion

As this submission has shown, the Online Safety Bill suffers from significant omissions with respect to the types of content falling within its scope as well as the necessary and proportionate measures that in-scope service providers must put in place to tackle the wide range of online hate speech acts. This means that **the Bill fails to give full effect to several of the UK's obligations under international human rights law to protect individuals from discrimination, violence or harm whilst ensuring that freedom of expression is not undermined**. As it stands, the Bill neither fully achieves its stated aim to protect children and adults from illegal or harmful content online, nor strikes an appropriate balance between their right to be free from violence, harm and discrimination and the rights of users to freedom of expression online. And this is because, rather than taking full responsibility to protect those rights, as states must do under international human rights law, the Bill simply delegates the necessary legislative, executive and judicial functions to profit-driven online platforms. In short, by affording Internet service providers significant discretion to define and sanction what they consider to be illegal and harmful speech acts, the Bill appears to further legitimise private censorship.

To remedy those omissions and align the Bill with the core international human rights treaties, **amendments must be introduced to clearly distinguish between criminal offences, civil and administrative wrongs falling within the scope of 'illegal content' under Section 41**. Likewise, new Schedules must be added to flesh out the extent to which 'offences', criminal or otherwise, defined in existing laws or regulations, amount to the types of online content of concern to the Bill. In the same vein, **the definitions of content harmful to children and adults in Sections 45 and 46 of the Bill must be further specified with parameters other than the harm caused to the victim**, such as the context surrounding the relevant speech act, the language

---

[58] YouTube Help, 'Like or dislike a video', accessed 28 August 2021.
[59] See Freedom of Information Request: Hate crimes relating to social media, *Office for National Statistics*, 23 December 2019.
[60] See, e.g., Crown Prosecution Service, Hate Crime Report 2018–19.

used, the position and intentionality of the speaker. And the Bill should **make clear that speech acts that are neither wrongful nor harmful must be protected, however offensive or critical they may be.**

Article 19(3) ICCPR also requires measures that limit speech to be laid down in clear, accessible and foreseeable laws, and be necessary and proportionate to their legitimate aim. This means that **whatever restrictive measures in-scope providers may or must adopt to discharge their safety duties, they must be clearly spelt out in Sections 9-11 and 21-22**. For user-to-user services, such measures must include **actions other than content takedowns**, seeking to prevent, halt and mitigate the impact of different types of online hate speech according to their level of seriousness.