Supplementary Evidence Submission - Online safety and online harms

AMENDING ONLINE SAFETY BILL TO ENSURE CONSISTENCY WITH CORE INTERNATIONAL HUMAN RIGHTS INSTRUMENTS: SPECIFIC RECOMMENDATIONS

Dr Talita Dias

I thank the Digital, Culture, Media and Sport Sub-committee on Online Harms and Disinformation for the opportunity to give oral evidence and submit these supplementary observations on the issue of online safety and online harms.

Executive summary

- To preserve and strengthen a regulatory model based on duties of care or due diligence, amendments to the Online Safety Bill should make it clear that service providers must reserve content takedowns to only the most serious types of illegal and harmful content, guided by the necessity and proportionality of available measures in the circumstances.
- To ensure that the Bill limits speech in a manner consistent with the requirements of legality, necessity and proportionality laid down in Article 19(3) of the International Covenant on Civil and Political Rights, the definitions of illegal and harmful content should be further specified to distinguish between criminal and non-criminal speech acts and require the assessment of their context, speaker, audience and accuracy.
- To address the dissemination of illegal and harmful content at its root, the Bill ought to introduce new duties specifically requiring providers to allow independent audits of their recommendation and content moderation algorithms, as well as to include information on their datasets, efficacy and human rights impact in transparency reports.
- To safeguard the rights to freedom of expression and non-discrimination of all users within the jurisdiction of the United Kingdom (UK), the Bill must do away with special protections for content of democratic importance or journalistic nature; instead it should require platforms to take these circumstances into account when considering the context of the relevant speech act and the necessity and proportionality of available measures.
- To ensure fair and effective redress systems in line with Article 2(3) of the International Covenant on Civil and Political Rights, the Bill must require in-scope providers to give notice of speech limitations and their reasoning to affected users, as well as clarify that judicial remedies under the Bill and other legal instruments remain available on a case-by--case basis.

1. Introduction: Scope of the present supplementary submission

This supplementary evidence submission seeks to inform the Digital, Culture, Media and Sport Sub-committee on Online Harms and Disinformation in making specific proposals to amend the Online Safety Bill, in light of discussions that took place during the Sub-committee hearing of Thursday, 23 September 2021.[1] Specifically, it builds on my previous

[1] Digital, Culture, Media and Sport Sub-committee on Online Harms and Disinformation, Oral Evidence Session, *Parliament TV*, 23 September 2021; Digital, Culture, Media and Sport Sub-Committee on Online Harms and Disinformation, Oral evidence: Online safety and online harms, HC 620, Thursday, 23 September

written and oral submissions to the Sub-committee and provides additional insights as well as concrete steps on how to amend the Online Safety Bill in line with core international human rights treaties,[2] particularly the International Covenant on Civil and Political Rights (ICCPR).[3]

1. Strengthening the Online Safety Bill's Underlying 'Duty of Care' Model

As highlighted in the Online Harms White paper[4] and discussed during the Sub-committee oral evidence session,[5] the Online Safety Bill's proposed regulatory model for online platforms is grounded in the idea of a 'duty of care'. This means that, rather than holding in-scope service providers responsible for hosted content published by others which platforms have knowingly failed to remove or moderate – a regulatory model known as 'intermediary liability' –, the Bill seeks to impose platform responsibility for *systemic* failure to exercise the requisite care or diligence by adopting certain safety measures stipulated in primary and secondary legislation.[6] In short, according to the 'duty of care' model, platforms can only be held liable for their *own conduct* (characterised by their lack of diligence in failing to adopt certain preventive or remedial measures) as opposed to the *conduct of others,* such as illegal speech acts, or specific *results*, such as the failure to remove a specific piece of content.

In my view, adopting a 'duty of care' regulatory model is, *in principle*, a step in the right direction. This is because, if designed and applied properly, this model enables platforms to exert their *best efforts* to moderate illegal or harmful content, as well as to safeguard users' freedom of expression without fear of being held liable for removing or failing to remove *specific* pieces or types of content. Granted, the Bill's proposed duty of care is inspired by general duties of care in tort law as well as specific statutory duties of care for the 'offline environment', such as those of property owners and employers. These may, to a greater or lesser extent, require, incentivise or effectively force duty-bearers to *successfully prevent* some type of *foreseeable harm* or the *risk* thereof.[7] Nevertheless, in the online context, such a duty ought to be understood in its own right as simply imposing an obligation to take *reasonable steps to* prevent, stop and/or redress *known* consequences of online speech acts, without necessarily requiring platforms to remove specific types or pieces of content and thereby 'successfully prevent' their ensuing consequences.[8] Likewise, concerns about the vagueness of the concept of 'online harms' and their foreseeability in the online environment[9] could be addressed by further specifying the definition of harmful content and requiring platforms to assess and disclose the impact of their recommendation algorithms, as proposed in sections 2 and 3 below. In sum, whether or not Parliament decides to label this set of obligations as a 'duty of care', *in substance*, platform duties ought to mirror the concept of

---

2021 (transcript).

[2] See United Nations Human Rights, 'The Core International Human Rights Instruments and their monitoring bodies', accessed 28 August 2021.

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

[4] Online Harms White Paper, April 2019, at 7-9, 41-49.

[5] Oral evidence: Online safety and online harms, HC 620 (n 1), at page 2, Q1.

[6] See generally Daphne Keller, 'Systemic Duties of Care and Intermediary Liability', *Stanford Law School, Centre for Internet and Society*, 28 May 2020.

[7] See Lorna Woods and William Perrin, 'Online harm reduction – a statutory duty of care and regulator', Carnegie Trust UK, April 2019, at 5, 25 and 26. For a critique of transposing this model to the context of 'online harms', see Graham Smith, 'Intermediary Liability And Responsibilities Post-Brexit', *Greenhouse*, 4 September 2020 and

[8] Lorna Woods and William Perrin (n 7), at 17; Mark Leiser and Edina Harbinj, 'CONTENT NOT AVAILABLE: Why The United Kingdom's Proposal For A "Package Of Platform Safety Measures" Will Harm Free Speech', September 2020, 78-90.

[9] Graham Smith (n 8); Mark Leiser and Edina Harbinj (n 8), at 80-81.

human rights due diligence, according to which *all* online and offline businesses have a social responsibility to exercise *reasonable diligence or care* in addressing their human rights impact.[10]

This model has the potential to strike an appropriate balance between the right to freedom of expression and the protection against non-discrimination and violence, in line with Articles 19, 20 and 26 of the ICCPR and other core human rights instruments. In contrast, by removing platform immunities with respect to user content, the intermediary liability model may easily force platforms to err on the side of removing content, thereby unduly restricting users' freedom of expression.[11] This is especially so if the power to give notice and require content takedowns comes not from a court order following due process but from an executive body exercising its own discretion.

The duty of care model is rightly reflected in *some of the limits* imposed on OFCOM's enforcement powers to issue 'use of technology warning notices' with respect to terrorist content and child sexual exploitation and abuse, as well as 'provisional notices of enforcement action' and subsequent confirmation decisions with respect to all other platform duties. Specifically, through a 'use of technology warning notice', OFCOM may require a user-to-user or search service provider to use accredited technology or human moderators to identify and swiftly remove terrorist or child sexual abuse or exploitation content (Sections 63-65). Nevertheless, such notices may *only* be issued if OFCOM has reasonable grounds to believe that the relevant provider has failed to comply with its safety duties with respect to illegal content in a *systemic* manner, rather than on a case-by-case basis, as evinced by the prevalence and persistent presence of terrorist or child sexual abuse or exploitation content (see Sections 63(2)-(3), 64(5), 65(5) of the Bill).

Furthermore, whilst the Bill empowers OFCOM to give notice of enforcement action (Sections 80-82) and confirm their decision to enforce a penalty (Section 83) if providers breach any of their statutory duties, including safety duties with respect to illegal and harmful content (Sections 9-11 and 21-22), it also stipulates in Section 83(11) that:

> A confirmation decision *may not* impose a requirement—
>
> (a) in the case of a user-to-user service, to use technology to identify *a particular kind of content* present on the service *with a view to taking down such content*;
>
> (b) in the case of a search service, to use technology to identify *a particular kind of content* in search results *with a view to such content no longer appearing in search results*.[12]

In essence, this provision seeks to avoid an intermediary liability model by precluding the regulator from issuing enforcement notices and decisions that require providers to remove particular types or pieces of content.

However, elsewhere in the Bill, the vagueness and breadth of OFCOM's enforcement powers, coupled with the imposition of equally vague and broad safety duties with respect to illegal and harmful content, may have the inadvertent effect of introducing an intermediary liability regime through the backdoor. This may happen because, although the regulator cannot specifically order the use of technology to remove particular kinds of content, no

---

[10] 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; 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, paras 40-45.

[11] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (n 10), para 30.

[12] Emphasis added.

additional restrictions exist on what types of platform behaviour may be held in breach of safety duties with respect to illegal and harmful content (Sections 9-11, 21and 22) and thus trigger the prohibitively high fines listed in Section 85 (4), i.e. £18 million or 10% of the provider's qualifying worldwide revenue, whichever is higher.

Thus, for illegal content *other than* terrorism and child sexual exploitation and abuse, such as stirring up racial hatred, OFCOM may take enforcement action if it considers that a provider has failed to put in place sufficient *human* moderators to swiftly take down types or pieces of illegal content, when alerted of their presence on the platform, in violation of Section 9(3)(d). Likewise, for content that is harmful to children, such as pressure to conform to certain stereotypes,[13] the regulator is empowered to take enforcement action if it considers that a provider has somehow failed to prevent its dissemination on the platform, in line with Section 10(3)(a). More worryingly, for the incredibly wide category of 'content that is harmful to adults', OFCOM may take enforcement action if considers that a service provider is not enforcing its terms of service or community standards *consistently*, in line with Section 11(3)(b). If the provider's own terms of service say that *legal but harmful* content must be *taken down* (which is the case of Facebook and Instagram's community standards/guidelines), the regulator may well fine the provider for failing to enforce such content takedown rules against certain users.

For online platforms, the easiest and cheapest solution is to simply take down all kinds of 'problematic' content, especially by using automated moderation technology, instead of having a human moderator carefully assess what kind(s) of action or measure(s) every piece of content actually deserves.[14] This is particularly the case of providers with a worldwide presence or operating in multiple countries with distinct regulatory frameworks. 'Reconciling' or finding the minimum common denominator across these distinct rules often leads to wide categories of 'offensive content' and broad content takedown rules in platform standards.[15] In this context, to ensure that any limitations to speech comply with the requirements of Article 19(3) of the ICCPR (legality, legitimacy, necessity and proportionality), legislation must require platforms to adopt clear standards that *carefully calibrate* limiting measures to the seriousness of the speech act and the importance of the right or interest to be safeguarded in a non-discriminatory manner.

The Bill does require platforms to a) take proportionate steps to mitigate and effectively manage the risks of harm to adults and children (Sections 9(2), 10(2), 21(2) and 22(2); b) put in place proportionate systems and processes designed to minimise the presence, dissemination and exposure to illegal content (Sections 9(3) and 21(3)), as well as to protect children from harmful content and prevent their exposure thereto (Sections 10(3) and 22(3)); c) adopt clear and accessible standards and apply them consistently (Sections 9(5), 10(4)-(5), 11(2)-(3), 21(4)-(5) and 22(4)-(5)).

But what it does *not* require is that providers ***reserve*** *content takedowns* to situations where these measures are **both *necessary* and *proportionate*** to tackle illegal and harmful types of content, considering, among other things, the seriousness of the speech act, its context, author, and audience. Instead, it *assumes that swift content takedowns are always*

[13] See Georgia Wells, Jeff Horwitz and Deepa Seetharaman, 'Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show', *The Wall Street Journal*, 14 September 2021.

[14] See Alexandre De Streel et al, 'Online Platforms' Moderation of Illegal Content Online: Law, Practices and Options for Reform', *European Parliament*, June 2020, at 40-41, 51-52; Spandana Singh, 'Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content', *New America Open Technology Institute*, 15 July 2019, at 5 and 35.

[15] Daphne Keller, 'Broad Consequences of a Systemic Duty of Care for Platforms', *Stanford Law School, Centre for Internet and Society*, 1 June 2020.

proportionate for all kinds of *illegal content* hosted on user-to-user services, irrespective of their degree of seriousness, context, etc., and leaves platforms free to use such measures to tackle all forms of illegal content (Section 9(3)(d)). Likewise, it leaves user-to-user service providers entirely free to choose which measures to adopt for tackling content that is harmful to children and adults, including binary leave-or-takedown content moderation policies, so long as they clearly and accessibly state these policies in their terms of service and apply them consistently (Sections 10(3)-(5), 11(2)-(3)). The same is true for search service providers with respect to both illegal content and content harmful to children (Sections 21(3)-(5) and 22(3)-(5)). It is for those reasons that many fear that, if the Bill is adopted in its current state, it will lead to platforms erring on the side of censorship[16] and defeating the very purpose for what it was put in place: the imposition of a duty to exercise due care or diligence rather than an intermediary liability regulatory model that forces platforms to take down content upon notice.

As highlighted in my original submission,[17] the root of this inconsistency lies in the Bill's failure to a) clearly define the different types of user-generated speech acts (apart from terrorism and child sexual abuse and exploitation) falling within in-scope providers' safety duties, and b) clearly specify the types of measures, *other than content takedowns*, which service providers may need to adopt with respect to different categories of content (i.e., prohibited versus limited, criminal versus non-criminal), in line with Article 19(3) of the ICCPR. More fundamentally, as discussed in section 3 below, **the Bill places excessive emphasis on *remedial* measures to tackle user-generated content**, such as content moderation and redress mechanisms, rather than measures to **address platforms' actual role in the *dissemination* of such content**: their curation and amplification through platform recommendation algorithms.[18]

To remedy those omissions and ensure that the Bill effectively puts in place a duty of care/due diligence regulatory model, as opposed to an intermediary liability (or notice-and-takedown) framework, I suggest:

a. **Redrafting Section 9(3)(d) to require providers to operate systems and processes designed to only take down the *most serious* types of *illegal* content, such as manifest and/or particularly grave criminal offences, in accordance with the requirements of necessity, proportionality and non-discrimination, taking into account, inter alia, the speech act's severity, context, author and audience**. As discussed in my original submission and the Sub-committee hearing,[19] not all types of illegal content have the same level of seriousness and thus require the adoption of such drastic measures as removal/deletion or user expulsion. For instance, while content that, when interpreted in context, clearly stirs up racial or religious hatred[20] should be swiftly taken down,[21] less serious or borderline illegal speech acts such as content posted through unauthorised access to a computer[22] or commercial

[16] E.g., Index on Censorship, 'Right to Type: How the "Duty of Care" model lacks evidence and will damage free speech', 17 June 2021, at 4 and 11.
[17] Talita Dias, 'Hate Speech and the Online Safety Bill: Ensuring Consistency with Core International Human Rights Instruments', Evidence Submission on Online safety and online harms, Digital, Culture, Media and Sport Sub-committee on Online Harms and Disinformation, September 2021, at 6 and 13.
[18] See Knight First Amendment Institute at Columbia University, 'Submission to Facebook Oversight Board (2/11/2021)', 11 February 2021, para 2; Yaël Eisenstat, 'Section 230 Revisited: Web Freedom vs Accountability', *Cornell Tech Critical Reflections*, 14 May 2020, at 4.
[19] Talita Dias (n 17), at 7; Oral evidence: Online safety and online harms, HC 620 (n 1), at page 11, Q18.
[20] See Parts III and 3A of the Public Order Act 1980, Article 20 of the ICCPR
[21] See Articles 4 and 5 of the International Convention on the Elimination of All Forms of Racial Discrimination (adopted on 21 December 1965, 660 UNTS 195).

advertisements potentially amounting to fraud by false representation,[23] may be labelled as such and/or de-prioritised, at least until a judicial decision confirms its illegality.

b. **Redrafting Sections 9(5), 10(5) and 11(2)-(3)** to require providers to not only ensure that their terms of service are clear, accessible and applied consistently but also **adopt a variety of measures, beyond mere content takedowns**, which may be **necessary and proportionate** to the seriousness of the illegal speech act in question, taking into account, inter alia, **the speech act's gravity, context, author and audience**.

c. In addition to requiring that any measure to tackle illegal or harmful content must be necessary, proportionate and applied in a non-discriminatory manner, Parliament may wish to include a **non-exhaustive list of such measures in Sections 9(5), 10(5) and 10(2)-(3).** Examples include labelling, tagging, redacting, deprioritising content and promoting counter speech, as well as the **measures already listed in Sections 13(7) and 14(10) for democratic and journalistic content**, i.e. 'giving a warning to a user, or suspending or banning a user from using a service, or in any way restricting a user's ability to use a service.' As explained in section 4 below, such a catalogue of measures – *not just content takedowns* – should be available as a response to *any* type of illegal or harmful content, not just content of democratic important and journalistic nature, to the extent necessary and proportionate. **Clarifying that applicable measures *must not* be restricted to content takedowns** and **listing examples** of such measures would go a long way to ensuring that the new restrictions on free speech introduced by the Bill are clear, accessible, necessary and proportionate. At the same time, this addition would **limit the power of OFCOM to enforce measures that may effectively incentivise or force companies to take down content**, inadvertently or not.

d. **Amending Section 83(11)** to stipulate more comprehensively that **a confirmation decision by OFCOM *may not* impose a requirement to take measures *amounting or leading to* the removal of *a particular piece or kind of content*** by user-to-user or search service providers. This would widen the scope of the current restriction to include all OFCOM confirmation decisions that have the *actual effect* of requiring or forcing platforms to take down specific types or pieces of content, whether by automated tools or human moderation.

2. Specifying the Definitions of Illegal and Harmful Content

As argued in my original written submission and oral evidence,[24] the definition of illegal content other than terrorist and child sexual exploitation and abuse conflates speech acts of different degrees of seriousness, such as prohibited speech under Article 20(2) of the ICCPR (advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence) and limited speech under Article 19(3) of the ICCPR (content that *may* be limited to respect the reputations of others or to protect national security, public order, public health or morals),[25] as well as criminal and non-criminal illegal speech acts.[26]

---

[22] Section 1 of the Computer Misuse Act 1990.
[23] Section 2 of the Fraud Act 2006.
[24] Talita Dias (n 17), at 6-7; Oral evidence: Online safety and online harms, HC 620 (n 1), at pages 2-3, 8, 11, Q1, Q2, Q11, Q18.
[25] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (n 10), paras 8-24.
[26] Human Rights Council, 'Report of the United Nations High Commissioner for Human Rights on the expert

Moreover, even illegal speech acts, including nudity and terrorist content,[27] can only be assessed and identified *in context*.[28] For instance, a user may be referring to or quoting known terrorist expressions or instances of incitement to violence to condemn these.[29] In the same vein, in certain environments, the use of words that are normally associated with terrorism or racial incitement, such as names of mosques, may not have a terrorist or discriminatory connotation. [30] Conversely, words that seem innocuous out of context, such as 'monkey',[31] 'worm' or 'cancer',[32] may amount to illegal content when used in certain contexts, such as expressions of racial superiority and incitement to violence. In sum, context is key: it is what separates illegal content from its denunciation or protected speech.

Likewise, I have posited that the definition of content harmful to children and adults is far too vague,[33] as it only takes into account the risk of direct or indirect adverse physical or psychological impact *on the victim* (Section 46(3)), and the extent of the content's *online dissemination* (Section 46(5)). Yet whether or not a certain type or piece of content is harmful or dangerous to a particular victim or a certain community and therefore deserving of limitation depends on **at least four other factors** that are *not* mentioned in the Bill. These are, first and foremost, the **context** of the speech act (for language is inherently contextual); the **position and intentionality of the speaker** (the more powerful and intent the speaker, the higher its risk of having an impact on the audience or victim); the **audience's susceptibility** to being influenced or affected by the speech act (which is particularly important to assess the content's risk of causing indirect harm to the victim, such as by inciting others to commit violence or discrimination); and the **content's accuracy** (false information may be particularly harmful to the health and reputation of others as well as public order, all of which are legitimate aims for limiting speech under Article 19(3) of the ICCPR).[34]

To illustrate the point, consider that a certain illegal or harmful speech act, such as a racial slur, a sexist comment or false information about a disease or medical treatment, has been posted by a high-profile politician with millions of keen followers and an intent to instigate violence or discrimination, in a social context marked by racial and political division and inequality. The exact same content, if published by an ordinary individual who is reckless about the consequences of their actions, in a more stable social context marked by resilient audiences, would be less harmful or risky for actual or potential victims. And no significant harm would likely arise if the same content were quoted or referenced as part of a protest or denunciation effort; quite the opposite: the quote would be raising awareness of the problem.

---

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; Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (n 11), paras 14, 18 and 24.

[27] Spandana Singh (n 15), at 18 and 22.

[28] Dangerous Speech Project, 'Dangerous Speech: A Practical Guide', 2020, at 20-21.

[29] Spandana Singh (n 15), at 13.

[30] See Justin Scheck, Newley Purnell and Jeff Horwitz, 'Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response Is Weak, Documents Show', The Wall Street Journal, 16 September 2021 (reporting that '[w]hen violence broke out between Israel and Palestinians […] [Facebook] erroneously suppressed Arabic-language regional news sources and activists, and began removing posts that included the name "Al Aqsa," an important Jerusalem mosque that was a focus of the conflict. Al Aqsa is also used in the name of the Al Aqsa Martyrs' Brigade, which the U.S. has designated as a terrorist organization.')

[31] Jeremy Burge, 'How the Monkey Emoji is Racist', *Emojipedia*, 12 July 2021.

[32] Dangerous Speech Project (n 28), at 13-14.

[33] Talita Dias (n 17), at 10-11.

[34] Dangerous Speech Project (n 28), at 7, 19-23.

I also urge Parliament to make it explicit that content that is harmful to adults must amount to content that, whether in a physical or non-physical way, undermines the aims that justify a speech limitation under Article 19 of the ICCPR, namely a) the rights or reputations of others; b) national security or public order; and c) public health or morals. Notably, 'the rights or reputations of others' include an individual's right to participate in the conduct of public affairs and vote (Article 25(a)-(b) of the ICCPR), which means that content such as electoral interference or disinformation and voter suppression could be deemed harmful and thus limited in a necessary and proportionate way, commensurate with the degree of seriousness of the speech act, when considered in context. I believe **reframing the list of 'online harms'** in this way not only addresses the Bill's omission of important phenomena, such as disinformation, but does so in a way that is fully in line with the ICCPR's requirements for limiting speech, especially legality, legitimacy, necessity and proportionality.

In this light, I suggest three sets of amendments to the Bill:

a. **Section 41(9) should be redrafted to specify that:**
    i. **'Illegal content' is limited to *existing criminal* offences under UK law**, not regulatory offences and other civil or administrative wrongs, such as breaches of gambling and consumer protection laws.
    ii. **Even criminal speech acts amounting to illegal content must be ascertained in context**.

b. **Section 41(5)(d) should be redrafted to stipulate that criminal offences** *other than* terrorism and child sexual exploitation and abuse failing within the scope of Section 41 are those **specified in a new Schedule 3** to be introduced by Parliament. The new Schedule should list **all or at least the most relevant criminal offences** in the laws of England and Wales, Scotland and Northern Ireland **that come within the scope of the Bill**, such as Parts III and Part 3A of the Public Order Act 1986, Sections 125-127 of the Communications Act 2003, Sections 1-3A of the Computer Misuse Act 1990, and Sections 1-11 of Fraud Act 2006.

c. **Section 45(5) should be redrafted to include as relevant factors for assessing whether content is harmful to children**, *in addition to* the content's impact on the victim (Section 45(5)), user exposure (Section 45(5)(a)), and dissemination of content (Section 45(5)(b)):
    i. The **context** in which the content is published, including the social and historical environment;
    ii. The position and mental state of the **speaker**, including any **intention to cause harm or spread disinformation**;
    iii. The **susceptibility of the audience** to accept, act upon or be otherwise influenced by the content;
    iv. The content's **accuracy**.

d. **In Section 46(5),** 'content carrying a material risk of directly or indirectly having significant adverse physical or psychological impact on an adult of ordinary sensibilities' should be **reframed as content that undermines or risks undermining, in a physical or non-physical manner, the legitimate aims for which speech may be limited under Article 19(3) of the ICCPR, namely:**
    i. **the rights or reputations of others;**

ii. **national security or public order; and**

iii. **public health or morals**.

e. **Section 46(5) should include as relevant factors for assessing whether content is harmful to adults** *not only* the physical or non-physical impact of the content on victims (Section 45(3)-(4)), the content's user exposure (Section 46(5)(a)) and its dissemination rate (Section 46(5)(b)) but also:

i. The **context** in which the content is published, including the social and historical environment;

ii. The position and mental state of the **speaker**, including any **intention to cause harm or spread disinformation**;

iii. The **susceptibility of the audience** to accept, act upon or be otherwise influenced by the content;

iv. The content's **accuracy**.

3. Addressing the Role of Platform Algorithms in the Dissemination of Illegal and Harmful Content

As discussed during the Sub-Committee oral evidence session,[35] online harms are not simply caused by user-generated content. Platform recommendation or optimisation algorithms play a significant role in *who* sees *what* content, *how*, and *under what circumstances*.[36] These are machine-learning algorithms[37] that are programmed to figure out for themselves whichever content generates more engagement.[38] Crucially, offensive, hateful, divisive and sensationalist content are prone to virality.[39] This means that platform recommendation algorithms, if made to boost engagement as they normally are, will inevitably increase the visibility of illegal and harmful content, by feeding users with more such content and prioritising it over other types of content.[40] For instance, the Wall Street Journal recently revealed that Facebook was aware that changes to its optimisation algorithm to place a heavier weight on reshared material 'made the angry voices louder', and led to '[m]isinformation, toxicity, and violent content [being] inordinately prevalent among reshares', as well as 'unhealthy side effects on important slices of public content, such as politics and news'.[41] In this context, content moderation can only be a remedial, palliative fix, just like a bush and dustpan cannot effectively clean the dirt continuously spread by a fan.[42]

---

[35] Oral evidence: Online safety and online harms, HC 620 (n 1), at pages 7 and 8, Q9 and Q11.

[36] Knight First Amendment Institute at Columbia University (n 18), para 2.

[37] Pavel Kordík, 'Machine Learning for Recommender systems — Part 1 (algorithms, evaluation and cold start)', *Medium*, 3 June 2018.

[38] 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; 'Facebook Whistleblower Frances Haugen: The 60 Minutes Interview', *YouTube*, 4 October 2021, timestamp 05:18.

[39] Ronald Deibert, 'The Road to Digital Unfreedom: Three Painful Truths About Social Media', 30 (2019) *Journal of Democracy* 25-39; Matthew Shaer, 'What Emotion Goes Viral the Fastest? On Twitter and Facebook, which spreads quickest: joy, sadness or disgust?', *Smithsonian Magazine*, April 2014; 'Facebook Whistleblower Frances Haugen: The 60 Minutes Interview' (n 38), timestamps 05:28 and 07:36.

[40] Yaël Eisentat (n 18), at 4.

[41] Keach Hagey and Jeff Horwitz, 'Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead', *The Wall Street Journal*, 15 September 2021.

And while recommendation algorithms excel at generating user engagement, content moderation ones are not very effective at identifying illegal and harmful content beyond plain nudity.[43] Examples include terrorist content, disinformation and hate speech.[44] This is because, as explained earlier, context is essential to assess whether such types of content are indeed illegal or harmful. Yet machine-learning content moderation algorithms cannot grasp the nuances of context in human language, images, audio or video.[45]

In this light, to address the *dissemination* of illegal and harmful content at its root, the Bill needs to regulate the design and operation of both platform recommendation and content moderation algorithms.[46] In the words of Facebook whistle-blower Frances Haugen, arguing for an independent government agency that would employ experts to audit the impact of social media:

> Today, Facebook shapes our perception of the world by choosing the information we see. Even those who don't use Facebook are impacted by the majority who do. A company with such frightening influence over so many people, over their deepest thoughts, feelings and behavior needs real oversight. But Facebook's closed design means it has no real oversight. Only Facebook knows how it personalizes your feed for you.[47] […]
>
> This inability to see in Facebook's actual systems and confirm that they work as communicated is like the Department of Transportation regulating cars by only watching them drive down the highway. […] Facebook should not get a pass on choices it makes to prioritize virality and growth and reactiveness over public safety. [48]

As noted in my previous written submission,[49] the Bill merely contains a general reference to the regulator's role to ensure that online services are designed and assessed 'with a view to protecting United Kingdom users from harm, including with regard to […] algorithms used by the service' (Section 30). To strengthen protection against illegal and harmful content, whilst safeguarding freedom of expression and platform intellectual property rights, the Bill should clearly delineate the ways in which OFCOM must assess the intended effects and actual impact of those algorithms. This could be achieved through the following amendments:

a. In **Chapter 3**, a Section should be introduced to impose on service providers **a duty to allow the regulator or an independent third-party to conduct confidential auditing or vetting of platform recommendation and content moderation algorithms**, to assess how they operate in theory and practice.

b. **Section 49(4)(b)** (on the information OFCOM may require from providers in transparency reports) **should be redrafted to include** not only generic 'information

---

[42] See statement by Facebook whistle-blower Frances Haugen before US Senate consumer protection panel on 5 October 2021 according to which 'Facebook's teams that drive the company's growth often work at cross-purposes with the teams responsible for keeping the platform safe' (John D. McKinnon and Ryan Tracy, 'Facebook Whistleblower's Testimony Builds Momentum for Tougher Tech Laws', *The Wall Street Journal*, 5 October 2021; Frances Haugen, 'Facebook whistleblower hearing', *Sky News*, 5 October 2021, timestamps 32:00, 44:50, opening statement transcript available here; 'Facebook Whistleblower Frances Haugen: The 60 Minutes Interview', *YouTube*, 4 October 2021, timestamp 06:00).

[43] Spandana Singh (n 15), at 6, 18-19; Frances Haugen (n 42), timestamps 01:04:00, 01:20:50).

[44] Spandana Singh (n 15), at 18 and 22.

[45] Spandana Singh (n 15), at 6-7, 15-16, 18.

[46] See Knight First Amendment Institute at Columbia University (n 18), para 2.

[47] 'Facebook Whistleblower Frances Haugen Opening Statement Transcript: Senate Hearing on Children & Social Media', *Rev*, 5 October 2021.

[48] John D. McKinnon and Ryan Tracy (n 42). See also Frances Haugen (n 42) on the effects of 'engagement-based ranking' algorithms, at timestamps 00:40:00, 01:01:00, 01:02:46, 03:18:51.

[49] Talita Dias (n 17), at 14.

about how […] content is disseminated by means of a service' but also **specific information about:**

a. **The datasets** on which recommendation and content moderation algorithms are trained;

b. **How** these datasets are **labelled or categorised** (including **who labels** them);

c. The relevant **parameters** for the algorithms' outputs, such as the number of views, likes, comments, shares/reshares, or private user information.[50]

d. Their **success rate** (including both false positives and false negatives, accuracy and reliability indexes);[51]

e. And **their impact on fundamental human rights**, such as non-discrimination, freedom of expression and privacy, including as evinced through internal platform research.[52]

As is well known, machine-learning algorithms are opaque even to their own creators because the code is programmed to continuously develop on the basis of new data.[53] Likewise, many online platforms make use of micro-targeting algorithms harnessing individual data, which means that no one user gets exposed to the same pieces of content.[54] Thus, to ensure greater transparency around the design and operation of these algorithms it is instrumental that the public has access to key information about the data on the basis of which they operate as well as the results they yield.[55]

4. Safeguarding Freedom of Expression for All

This recommendation follows on from a point widely discussed during the Sub-committee oral evidence session:[56] the need for equal and non-discriminatory protection of the right to freedom of expression of *all users and their speech acts*, not just journalistic and democratic content, as presently stated in Sections 13 and 14 of the Bill. As a fundamental human right enshrined in Article 19 of the ICCPR, Article 10 of the European Convention on Human Rights (ECHR), and protected under the UK common law, freedom to receive and impart information and ideas of all kinds belongs to each and every human being within the UK's jurisdiction. Thus, if a certain content is *not prohibited or limited* by law for a legitimate purpose and in a necessary and proportionate manner, *it must be protected*, no matter its source.[57] Likewise, the right to non-discrimination, recognised in Articles 3 and 26 of the ICCPR and Article 1 of Protocol 12 to the ECHR, prohibits the unequal treatment of individuals on the basis of race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or another status. By granting special treatment to content of democratic importance (Section 13) and journalistic content (Section 14) the Bill fails to uphold both these rights.

---

[50] Frances Haugen (n 42), timestamp 02:21:30.
[51] Spandana Singh (n 15), at 13, 17, 19.
[52] See The Santa Clara Principles on Transparency and Accountability in Content Moderation, 2 February 2018, Principle 1; Frances Haugen (n 42), timestamp 01:4149.
[53] Spandana Singh (n 15), at 20; Frances Haugen (n 42), timestamp 01:29:07.
[54] Yaël Eisenstat (n 18), at 4-5; Frances Haugen (n 42), timestamp 34:25.
[55] Frances Haugen (n 42), timestamps 33:40, 01:07:45, 01:41:49, 02:11:40, 02:23:10, 02:26:50)
[56] Oral evidence: Online safety and online harms, HC 620 (n 1), at page 14-15, Q23 and Q25.
[57] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (n 10), para 24.

Granted, whether a certain content is of democratic importance or journalistic nature should be taken into account when considering what measures might be necessary and proportionate to limit speech and safeguard competing rights, such as the health and reputations of others. However, **this does *not* mean reserving special measures or 'privileges'**, such as those listed in Sections 13(7) and 14(10), ***only* for those types of content**. Rather, the democratic importance or journalistic nature of a certain content should be considered: a) as part of its *context* for the purposes of assessing whether the content is indeed illegal or harmful in the first place; and b) as part of the *necessity and proportionality* tests when assessing applicable limiting measures, once the content has been found to be illegal or harmful. As explained in Section 2 above, quite apart from deserving privileged treatment and laxer measures, certain types of democratic or journalistic content should be subject to more stringent responses, *commensurate* to the speaker's prominence, mental state, the susceptibility of their audience to act upon the speech act, and the content's accuracy.

To incorporate this recommendation, Sections 13 and 14 of the Bill should be merged into a single provision that reflects the following:

a. When considering **'context'** for the purposes of ***identifying* illegal content** (Section 41(9)), **content harmful to children** (Section 45(5)), **and content harmful to adults** (Section 46(5)), service providers **must have regard to** the content's **democratic importance** and/or **journalistic nature**.

b. When considering what **measures** under Sections 9, 10, 11, 21 and 22 are **'necessary and proportionate'** to address a certain type of illegal or harmful content, service providers **must have regard to** the content's **democratic importance** and/or **journalistic nature.**

5. Ensuring Effective Redress Systems

As noted in my previous written submission[58] and oral evidence, one important omission of the Bill is the failure to specify that judicial remedies remain available to users and members of the public who have been affected by the dissemination of illegal or harmful content.[59] Access to justice is essential to ensure that victims of illegal or harmful content disseminated offline and online, as well as those affected by measures limiting speech, such as content takedowns, can obtain an effective remedy against unlawful platform action. It is also an important safeguard against the concentration and potential abuse of power by the executive, *in casu*, the Secretary of State and OFCOM.

Two additional recommendations on this point are warranted. First, to ensure that users affected by measures limiting speech can obtain redress directly from in-scope providers through redress systems (Sections 15 and 24), super-complaints by eligible entities before OFCOM (Sections 106-108) or judicial remedies, the Bill must require both user-to-user and search service providers to give **notice** of any limiting measures applied, along with **the reasons for the limitation**.[60] Without knowledge of restrictions and their basis, it is difficult if not impossible for affected users to have a **fair opportunity** to make an **effective complaint** before service providers, entities representing their interests in OFCOM super-complaints, or courts.

---

[58] Talita Dias (n 17), at 14.
[59] Index on Censorship (n 16), at 3.
[60] Alexandre De Streel et al (n 14), at 42-43, 49-50; See The Santa Clara Principles on Transparency and Accountability in Content Moderation (n 52), Principle 2.

Second, to avoid the introduction of an intermediary liability model through the backdoor in individual cases before courts, the Bill needs to clarify that any **judicial complaints *with respect to/based on* service providers' *duties of care* under Sections 5 to 25 must *not* result in platform liability for user-generated content**. Put differently, any judicial remedies granted *on the basis of the Online Safety Bill* against platform action must not arise from user-generated content *per se* but a breach of a duty of care, i.e., failure to exercise the requisite diligence by putting in place the measures stipulated in the Bill or secondary legislation with respect to that content, whether these measures involve content moderation or the protection of free speech or privacy.

To be sure, courts must be able to assess whether the Bill and ancillary secondary legislation have been correctly applied in individual cases.[61] This is to uphold the right to an effective remedy under Article 2(3)(b) of the ICCPR and avoid the concentration of power in the hands of the executive. As public adjudicators, courts are responsible for interpreting and applying the law in individual cases, even if this means requiring companies to exercise their duties of care with respect to particular types or pieces of content. However, **any individual or collective decision *grounded in the Bill* must not impose intermediary liability** on in-scope providers for those specific types or pieces of user-generated content. Rather, it must hold platforms **responsible for their *own behaviour* or *lack of care***, such as the amplification of illegal or harmful content, the failure to remove it when necessary and proportionate, or the wrongful removal of protected speech.[62] For example, if a user posts content that amounts to the criminal offence of stirring up racial hatred, platforms or their employees should not be liable for this offence but for breaching their own safety duties to implement measures to limit this content in a necessary and proportionate manner under Section 9. Nevertheless, this is ***without prejudice* to platform responsibility under *other applicable laws***, such as tort liability or complicity in criminal offences, provided that their respective elements are present.

To address those concerns, I suggest:

a. Amending **Section 15(3)** to impose on user-to-user service providers a duty to operate a complaints procedure that gives users **notice of any restriction** on their ability to use the service, along with the **reasons** for the restriction;
b. Amending **Section 24(3)** to impose on search service providers a duty to operate a complaints procedure that gives users **notice of any restriction** on their ability to use the service, along with the **reasons** for the restriction;
c. Inserting a provision in **Section 106** to clarify that the right of eligible entities to make super-complaints before OFCOM is **without prejudice** to the right of individuals to **access courts and make judicial complaints** on a case-by-case basis for breaches of user-to-user and search service providers' duties of care laid down in the Bill and other acts or omissions that are unlawful under other applicable laws.

Conclusion

The Online Safety Bill is a step in the right direction when it comes to imposing on online service providers duties of care or due diligence to prevent, limit and redress illegal and harmful content, as opposed to holding them directly liable for such content. However, amendments are essential to uphold the Bill's very regulatory thrust and avoid untended

[61] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (n 11), paras 7, 33, 37.
[62] See Yaël Eisenstat (n 18), at 4-5.

consequences on freedom of expression and other fundamental rights. In this supplementary submission, I have suggested five sets of amendments to the Bill to address key concerns raised in my original submission and the Sub-committee oral submission session. These are: 1) clarifying that service providers have a duty to apply a range of content moderation measures – not just content takedowns – in a way that is necessary and proportionate to the seriousness of the illegal or harmful content and the interest to be protected, along with other relevant factors; 2) further specifying the definition of illegal and harmful content, especially to include a requirement to assess the speech act's context, speaker, audience and accuracy; 3) introducing auditing and transparency duties with respect to platform recommendation and content moderation algorithms; 4) securing the right to freedom of expression of all users and all types of content, not just those of democratic importance or journalistic nature; and 5) ensuring an effective redress system by requiring notice of restrictions and reasons to affected users, whilst safeguarding the rights of all individuals to judicial remedies within and beyond the Bill. With these changes, I believe the Online Safety Bill can bring the UK closer to international human rights instruments, placing it at the forefront of global Internet freedom and safety.