scripted

Volume 19, Issue 1, February 2022

[Redacted]: This Article Categorised [Harmful] by the Government

Edina Harbinja and Mark R. Leiser***

DOI: 10.2966/scrip.190122.88

Abstract

In April 2019, the UK Government's DCMS released its White Paper for 'Online Harms', which would establish in law a new duty of care towards users by platforms to be overseen by an independent regulator. Our earlier research outlines how we got to this point, sets out what the White Paper proposes, and criticises its key aspects. Our objections and criticism remain applicable to the UK Government's Online Safety Bill. The Parliament is now scrutinising the Bill. The House of Lords Report sparked some optimism that the scrutiny could address critical concerns around free speech in particular. The Draft Online Safety Bill Joint Committee Report, however, suggest otherwise. This paper returns to key arguments as to why risk-based regulation and duty of care are not appropriate for policing content and expression online. We focus on the human rights implications of the Bill, in particular, the provider duties to 'handle' legal but harmful content. Here, we reemphasise the vague conceptualisation and nature of this harm, as well as the inadequate duties attached to it. We argue that the independence of OFCOM cannot be guaranteed.

Keywords
Free expression, online harms, platform regulation, duty of care, platforms, Online Safety Bill

* Senior lecturer in media/privacy law, Aston Law School University, Birmingham, United Kingdom, e.harbinja@aston.ac.uk.
** Assistant Professor in Law and Digital Technologies, eLaw, Leiden University, Leiden, The Netherlands, m.r.leiser@law.leidenuniv.nl.

1 Introduction

During the early days of April of 2020, five telecom towers were torched in an alleged arson attack after a conspiracy theory took hold online that the 5G technology installed on the towers was responsible for a variety of health issues, environmental damage, and accountable for the spread of Covid-19.[1] Although no one was hurt, officials argued that these telecoms were crucial communications infrastructure and responsible for emergency service connections, thus putting lives at stake. The conspiracy theory that 5G towers are responsible for the spread of coronavirus was driven mainly by online posts by secretive groups. Actors spread *harmful* but not *illegal* disinformation that manifested into real-world illegality.

The unpredictable nature of the incident requires an examination of risk and preventative measures associated with online harm. Further investigation should be done to assess trends of movement towards new models of regulation for large social media platforms.[2] The pandemic, and subsequent declaration of

---

1 Cecilia Kang, "Fake News Onslaught Targets Pizzeria as Nest of Child-Trafficking" (*The New York Times*, 21 November 2016), available at https://www.nytimes.com/2016/11/21/technology/fact-check-this-pizzeria-is-not-a-child-trafficking-site.html (accessed 31 January 2022).

2 European Commission, "Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Online Platforms and the Digital Single Market Opportunities and Challenges for Europe", pp. 4-5, available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52016DC0288&from=EN (accessed 31 January 2022); European Parliament, Directorate General For Internal Policies Policy Department, "A: Economic and Scientific Policy, Providers Liability: From the eCommerce Directive to the Future", available at http://www.europarl.europa.eu/RegData/etudes/IDAN/2017/614179/IPOL_IDA(2017)614179_EN.pdf (accessed 31 January 2022); see also Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law" (*Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression*, April 2019), available at https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf (accessed 31 January 2022).

a public health emergency,[3] plus the increasing scrutiny of platform power,[4] and the United Kingdom's alarming drive toward specific statutory duties of care to protect users from illegal and harmful content,[5] all raise the following question:

> Is risk-based regulation, through the imposition of a statutory duty of care on large social media platforms, an appropriate model for regulating user-generated content?

A consistent finding from research on harmful online behaviours in the US is that they tend to concentrate among a few individuals. Whether examining the amplification of misinformation, consumption of radical content, or posting of hate speech, a small group of individuals typically accounts for the majority of the behaviour.[6] Therefore, the second part of this article addresses the following question:

---

3 WHO, "Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)" (*World Health Organization*, 31 January 2020), available at https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov) (accessed 31 January 2022).

4 European Commission, "The Digital Markets Act: ensuring fair and open digital markets" available at https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en (accessed 31 January 2022); see also European Commission, Proposal for a Regulation on contestable and fair markets in the digital sector (Digital Markets Act), COM/2020/842 final, available at https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608116887159&uri=COM%3A2020%3A842%3AFIN (accessed 31 January 2022); Robin Mansell, "Platforms of power", (2015) 43(1) *Intermedia* 20-24; Orla Lynskey, "Regulating Platform Power", *LSE Law, Society and Economy Working Papers* 1/2017, available at http://eprints.lse.ac.uk/73404/1/WPS2017-01_Lynskey.pdf (accessed 31 January 2022).

5 For our analysis of the precursor to the Online Harms Bill, see our earlier work on the White Paper in Mark Leiser and Edina Harbinja, "Content not available: Why the United Kingdom's Proposal For A 'Package Of Platform Safety Measures' Will Harm Free Speech" (2020) *Technology and Regulation* 78-90.

6 David Lazer et al., "Meaningful measures of human society in the twenty-first century" (2021) 595 *Nature* 189–196.

> Is the Online Safety Bill's risk-based approach an appropriate response to the behaviour of a small number of users?

As the authors have previously discussed in 'Content Not Available', "a duty of care normally carries with it a three-stage test of foreseeability, proximity, and policy."[7] Absent any indication to the contrary, and platforms could be liable for conspiracy theories like 5G towers spreading coronavirus. A shift in the default liability position could mean the end of the 'no general monitoring obligation'[8] and a significant chilling effect on free speech and expression as platforms seek to limit their potential liability for user-generated content. We set this article as follows: Part one provides an update to the UK's position on platform regulation by analysing the 'Online Harms Bill' working its way through the British Parliament. Part two examines the appropriateness of risk-based content moderation as a means of platform regulation. We briefly explore the proposal's implications through the European Convention of Human Rights (ECHR) lens.

2 The Online Safety Disaster – A Captured Regulator Protecting Offended from Indirect Harm

Compared to the proposal set out in its predecessor, the Online Harms White

---

7 *Ibid*., at n. 51, citing *Caparo Industries v. Dickman* [1990] 2 AC 605 (HL).

8 Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce'), OJ L 178, 17.7.2000, Art. 15; given effect in the United Kingdom by Regs. 17, 18, and 19 of the E-Commerce (EC Directive) Regulations 2002, SI 2002/2013.

Paper,[9] the Draft Online Safety Bill (from now on, the Bill),[10] published in May 2021, introduces significant new features.[11] Notably, the Bill has dropped an unhelpful suggestion, originating from the Carnegie UK Trust, of an overarching duty of care for platforms (service providers):[12] Instead, the Bill introduces several specific duties of care:

(1) Duty of care related to illegal CSEA (child sexual exploitation and abuse) and terrorist-related content;
(2) Duty of care related to other illegal content;
(3) Duty of care related to 'harmful but legal content';
(4) Category 1 Service Provider's duty of care (journalistic and democratic content);
(5) Duty to have regard to the importance of free speech and privacy;
(6) Duties about reporting and redress;
(7) Record-keeping and review duties.

In light of other work on the imposition of any overarching duty of care,[13] these

---

9 UK Government, "Online Harms White Paper: Full Government Response to the Consultation", available at https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response (accessed 31 January 2022).

10 DCMS, Draft Online Safety Bill, C 405, (hereinafter 'Draft Online Safety Bill') available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf (accessed 31 January 2022).

11 We analysed in an earlier paper, Leiser and Harbinja, *supra* n. 5.

12 William Perrin and Lorna Woods, "Reducing harm in social media through a duty of care" (*Carnegie UK Trust*, 8 May 2018), available at https://www.carnegieuktrust.org.uk/blog/reducing-harm-social-media-duty-care/ (accessed 31 January 2022).

13 Leiser and Harbinja, *supra* n. 5; Graham Smith, "Harm Version 3.0: the draft Online Safety Bill" (*Cyberleagle Blog*, 16 May 2021), available at https://www.cyberleagle.com/2021/05/harm-version-30-draft-online-safety-bill.html (accessed 31 January 2022); "Take care with that social media duty of care" (*Cyberleagle Blog*, 19 October 2018) https://cyberleagle.com/2018/10/take-care-with-that-social-media-duty.html (accessed 31 January 2022).

more specific duties carry remarkably similar connotations and still represent an inadequate analogy with the duty of care in the offline world (health and safety law or the 'wet floor paradigm'). Worryingly, the Joint Online Safety Bill Committee's scrutiny of the Bill revealed that Parliament is still considering reintroducing an overarching duty of care. If the number of questions to witnesses about an overriding duty of care is anything to go by, the imposition is still one of the options under consideration by Parliament.[14] As a result of the reasons given by the Office of the Parliamentary Counsel, the Secretary of State considers this option unlikely. She explained it to the Committee quite bluntly: "If it does not pass the lawyers in Parliament, we cannot do it."[15] We will not repeat our analysis of the imposition of a duty of care for online speech.[16] Instead, this article goes a step further, focusing on the specifics of the more problematic proposals in the Bill. Crucially, our analysis includes the duties of Category 1 Service Providers, the category of 'legal but harmful content' and confusing, vague types of journalistic and democratic content. All of these largely justify references to the Bill as a 'Censors' Charter'.[17]

Absent from the Online Harms White Paper, 'Category 1' Service Providers appears to be a new addition to the scope of the Online Safety Bill. In addition to common duties for all providers, these service providers have two

---

[14] Draft Online Safety Bill (Joint Committee), All events, 9 September–8 November 2021, available at https://committees.parliament.uk/committee/534/draft-online-safety-bill-joint-committee/events/all/ (accessed 31 January 2022).

[15] Draft Online Safety Bill (Joint Committee), "Uncorrected oral evidence: Consideration of Government's draft Online Safety Bill", 4 November 2021, p. 40, available at https://committees.parliament.uk/oralevidence/2949/pdf/ (accessed 31 January 2022).

[16] Leiser and Harbinja, *supra* n. 5.

[17] David Davis MP and others, "Government's Online Safety Bill will be 'catastrophic for ordinary people's freedom of speech'" (*Index on Censorship*, 23 June 2021) available at https://www.indexoncensorship.org/2021/06/governments-online-safety-bill-will-be-catastrophic-for-ordinary-peoples-freedom-of-speech-says-david-davis-mp/ (accessed 31 January 2022).

distinct kinds of duties: duties to protect the content of 'democratic importance' and duties from safeguarding 'journalistic content'. Content included in the duty to preserve content that is 'of democratic importance' is broadly defined in clause 13 of the Bill as content intended *to contribute to democratic political debate in the UK or a part of it*. Essentially, this is a duty not to remove a limited number of undefined categories of speech in the public interest. The definition of this content is extensive and overlaps with the duty to protect 'journalistic content', set out in clause 14 of the Bill.

Of vital importance is the needed clarity over whether political, 'democratic' speech should be distinguished from other crucial forms of free speech and to what extent one can establish a clear separation of the two. For example, is speech related to health and vaccines political and of democratic importance? Is speech about religion or questioning religious views of democratic importance, such that blasphemy is protected? In their report earlier this year, the House of Lords Communications and Digital Committee recommended that this category be reconsidered and broadened to include debates about 'social change'.[18] The Government responded to this recommendation explaining that "this definition does not afford protection to content designed to undermine democratic processes, such as harmful disinformation designed to damage the integrity of elections."[19] Thus, the Government has only provided one example of excluded speech, but what about

---

[18] House of Lords, Communications and Digital Committee, "Free for all? Freedom of expression in the digital age", HL paper 54, 22 July 2021, para. 80, available at https://committees.parliament.uk/publications/6878/documents/72529/default/ (accessed 31 January 2022).

[19] Department for Digital, Culture, Media and Sport, "Government response to the House of Lords Communications Committee's report on Freedom of Expression in the Digital Age", October 2021, para. 7, available at https://committees.parliament.uk/publications/7704/documents/80449/default/ (accessed 31 January 2022).

other types of speech that could be 'harmful' and thus unprotected as democratic content?[20] Worryingly, the Bill appears to leave the determination of whether topics like these are of democratic importance to the regulator, OFCOM, and service providers.[21] The Bill's lack of compliance with human rights law is further discussed in the final section.

In terms of journalistic content, the Bill introduces a broad definition of 'journalistic content'.[22] The definition seems to cover user content 'generated for journalism', e.g. citizen journalists, as long as there is a UK link between the content and the author. The Government's press release noted that the content of 'citizen journalists' will have the same protections as 'professional journalists'.[23] For content and the UK link, clause 14(9) reads:

> (a) United Kingdom users of the service form one of the target markets for the content (or the only target market), or
> (b) the content is or is likely to be of interest to a significant number of United Kingdom users.

---

[20] Similar to this type of 'harm' is a new category proposed by Carnegie UK in their "Revised Online Safety Bill", i.e. "(d) harm to democratic debate or to the integrity and probity of the electoral process.", Carnegie UK, "Revised Draft Online Safety Bill", (November 2021), c. 3A, available at https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2021/11/10120738/Carnegie-UK-Revised-Online-Safety-Bill-November-2021.pdf (accessed 31 January 2022).

[21] "While platforms will have some discretion about what these policies are, they will need to balance the importance of protecting democratic content with their safety duties, and will still be able to remove content that is prohibited by their terms of service. For example, *platforms will need to consider whether the public interest in seeing some types of content outweighs the potential harm it could cause, or vice versa*." (emphasis added), *ibid.*, para. 9. Platforms are surely not in a position to consider public interest of any sort, let alone that specific to the UK's democratic processes and debates.

[22] Draft Online Safety Bill, cls. 13 and 14.

[23] Gov.uk Press Release, "Landmark laws to keep children safe, stop racial hate and protect democracy online published" (12 May 2021), available at https://www.gov.uk/government/news/landmark-laws-to-keep-children-safe-stop-racial-hate-and-protect-democracy-online-published (accessed 31 January 2022).

For content creators, the clause requires a person to 'be' in the UK or for the entity formed under UK law or any part of the UK. This clause means a foreign correspondent working as a freelance journalist might not get protection when writing from abroad but does get protection as soon as they are back in the UK. Furthermore, a platform may arbitrarily interpret the 'significant number' of users. For instance, a platform may consider a couple hundred users 'significant', while another requires thousands before this threshold is met. One can easily see the difficulty with this definition if one considers a journalist based in the Republic of Ireland writing about the situation in Northern Ireland. Given the special status of Northern Ireland and historical difficulties in this area, one could argue that this content is likely to be of interest to a significant number of UK users. However, if the journalist is not based in the UK, would their content be protected under the Bill? This outcome remains unclear.

The Category 1 Service Provider will be required to "make a dedicated and expedited complaints procedure available to a person who considers the content journalistic content". In practice this means that it will be difficult for platforms to ascertain if a user post on social media should be deemed as journalistic and the user considered a citizen journalist. Suppose a platform does not view the user or the content as journalistic; in that case, the user will not have an option to challenge a take-down because the redress option is exclusively available for journalistic content and not for the content of democratic importance. It is unclear as to why this distinction exists. For example, a user posts on social media that they were attacked in the toilets by 'a man in a dress'; this post could be harmful to the trans community and the platform. However, is this not a citizen journalist self-reporting an incident that happened personally? Is a woman not feeling safe in public not part of a conversation of democratic importance? Countless examples show why this distinction is problematic. Arguably, both journalistic speech and speech of democratic

importance are essential in a modern democratic, diverse society and deserve a user redress mechanism. In their critique of Germany's approach to content moderation as set out in the NetzDG, the Human Rights Committee recently reemphasised their criticism about the lack of a system for user appeal and the lack of "judicial oversight and access to redress in cases where the nature of online material is disputed."[24] The Bill likely prioritises Journalist content due to fierce lobbying from the traditional media industry, which sought assurance that their content and activity will not be policed under the Bill.[25]

Furthermore, the full extent of the scope of Category 1 Service Providers is unclear. The Bill provides that the Secretary of State will make regulations to specify conditions for 'Category 1' services, based on the number of users and service functionalities.[26] Thus, the Secretary of State will need to consult OFCOM before making regulations. In their press release, the Government did hint that 'Category 1' will include large platforms and social media, but the exact scope has yet to be determined.[27] The requirement to consult is just one of the arbitrary decisions that the Secretary of State will have the power to make. The second controversial proposal in the Bill is not distinguishing between 'illegal' and 'legal but harmful' content.

As it stands, 'legal but harmful' content will be a category of regulated content and subject of the service provider's duty of care and risk assessments.[28]

---

[24] Human Rights Committee, "Concluding observations on the seventh periodic report of Germany", 11 November 2021, paras. 46–47, available at https://tbinternet.ohchr.org/Treaties/CCPR/Shared%20Documents/DEU/CCPR_C_DEU_CO_7_47161_E.pdf (accessed 31 January 2022).

[25] Draft Online Safety Bill (Joint Committee), "Written evidence submitted by DMG Media (OSB0133)", 14 October 2021, available at https://committees.parliament.uk/writtenevidence/39297/pdf/ (accessed 31 January 2022).

[26] Chapter 59 and sch. 4 of the Draft Online Safety Bill.

[27] Gov.uk Press Release, *supra* n. 23.

[28] The Draft Online Safety Bill, c. 46.

Our previous article discussed the concept of harm and the vague nature of many types of harm set out in the Online Harms White Paper.[29] The Bill has not addressed these concerns, and has amplified them in many ways. Harmful content will need to be 'dealt with' when "[t]he provider…has reasonable grounds to believe that the nature of the content is such that there is a material risk of the content having, or indirectly having, a significant adverse physical or psychological impact on a child (adult)."[30] The first problem is that the harm within the scope of the Bill can be 'indirect'. It is unclear what this indirect harm includes and amounts to a vast category that adversely affects human rights.[31] The lack of certainty and requirements of evidence that prove causation for these forms of harm is problematic, especially if we consider it from the perspective of the judicial tests for lawful limitations of human rights.[32] For instance, would disclosing that there is no Santa Claus, which can be psychologically damaging for young children, be harmful under the Bill?

Moreover, much of what defines harmful content is already illegal; for example, terrorist-related content, content related to child abuse, extreme pornography etc.[33] Many of the harms that the Government's White Paper identified as 'legal harms' (e.g. disinformation, trolling, or intimidation) could potentially be within the remit of the protection awarded by Article 10 of the ECHR (the right to freedom of expression). Offensive content may be harmful to some but not rise to the threshold of illegality and may be protected speech.[34] The

---

29 Leiser and Harbinja, *supra* n. 5, pp. 82-84.

30 The Draft Online Safety Bill, c. 46.

31 Graham Smith, "Harm Version 3.0: the draft Online Safety Bill" (*Cyberleagle Blog*, 16 May 2021) available at https://www.cyberleagle.com/2021/05/harm-version-30-draft-online-safety-bill.html (accessed 31 January 2022).

32 *Axel Springer AG v Germany*, 39954/08 [2012] ECHR 227 (7 February 2012).

33 Leiser and Harbinja, *supra* n. 5, pp. 82-84.

34 *Ibid*., p. 84.

new legislation should apply only to illegal content and leave out the controversial concept of legal but harmful content. If the content is indeed detrimental, legislation needs to outlaw it. This is likely to be done soon for aspects of disinformation, introducing a new offence, i.e. "sending knowingly false communications".[35] Otherwise, all other speech should remain within the domain of lawful free speech. We will elaborate on the overarching problem of causation in the following section in more detail.

What effects does this odd concept, as set out in the Bill, have? Service providers will need to be able to specify in their terms of service how harmful content "is to be dealt with by the service",[36] after having determined if the content is harmful (directly or indirectly) to children or adults ("service has reasonable grounds to believe that the nature of the content is such that there is a material risk of harm…"). Section 46(3) of the Draft Bill defines legal but harmful content as "content having, or indirectly having, a significant adverse physical or psychological impact on an adult of ordinary sensibilities". The standard used for this assessment is a risk of harm to an adult or child of 'ordinary sensibilities'. This legal standard is inadequate and does not correspond to the well-established standard of a 'reasonable person' in tort law.

---

[35] Law Commission, "Modernising Communications Offences: A final report", Law Com No 399, 20 July 2021, pp. 77–94, available at https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2021/07/Modernising-Communications-Offences-2021-Law-Com-No-399.pdf (accessed 31 January 2022); according to the recommendation, a defendant would be liable if they knowingly send or post a communication that they know to be false and they intend to cause **non-trivial emotional, psychological, or physical harm to the likely audience**, without a reasonable excuse. This would address issues related to the low bar in s. 127 the Communications Act 2003, i.e. "annoyance, inconvenience or needless anxiety", raising the threshold for the offence to causing non-trivial harm. For a sounds critique of this proposal, see Graham Smith, "Licence to chill" (*Cyberleagle Blog*, 22 November 2021), available at https://www.cyberleagle.com/2021/11/licence-to-chill.html (accessed 31 January 2022).

[36] Clause 11 of the Draft Online Safety Bill.

It leaves open questions about whether those 'easily offended' fall within this category and what level or sensibility is considered ordinary.[37]

Furthermore, UK courts have not developed this test and service providers and regulators lack any exact points of reference. The Bill takes the concept from the tort of abuse of private information; however, in that context, it operates in an entirely different manner. The person with ordinary sensibilities is the person whose data is disclosed, not the person who receives the information.[38] The person who receives the information may be an individual (or a member of a particular group) with a low level of 'sensibility' and find, for instance, blasphemous, profane, violent, sexualised, or satirical content offensive and harmful to their mental health and peace of mind. Can we order companies to censor this type of speech just to avoid any risk of offence to those who present as statistical infrequencies? Looking at tests applied by the UK and European courts, the answer is no.[39]

A consistent finding from research on harmful online behaviours in the US is that the bad behaviour concentrates among a small set of individuals. Whether examining the amplification of misinformation, consumption of radical content or posting hate speech, a small set of individuals typically accounts for most of the behaviour.[40] Despite this statistical infrequency, the social

---

37 Graham Smith, "On the trail of the Person of Ordinary Sensibilities" (*Cyberleagle Blog*, 28 June 2021), available at https://www.cyberleagle.com/2021/06/on-trail-of-person-of-ordinary.html (accessed 31 January 2022).

38 *Ibid*., and *Campbell v MGN Ltd* [2004] UKHL 22.

39 See e.g. *Handyside*, *R v Scottow* [2020] EWHC 3421 (Admin), per Bean LJ: "I do not consider that under s. 127(2)(c) there is an offence of posting annoying tweets", *Hayden v Dickenson* [2020] EWHC 3291 (QB) [40-44], "The Court's assessment of the harmful tendency of the statements complained of must always be objective, and not swayed by the subjective feelings of the claimant."

40 Bertie Vidgen, Helen Margetts, Alex Harris, "How much online abuse is there? A systematic review of evidence for the UK", Policy Briefing – Full Report, Public Policy Programme, Hate Speech: Measures and Counter Measures, The Alan Turing Institute, available at https://www.turing.ac.uk/sites/default/files/2019-

consequences of such behaviour can be substantial, especially for groups targeted by these individuals. The observation that human behaviour often follows a heavy-tailed distribution (i.e. the Pareto principle) is well established and applies to many behaviours. However, studies have usually focused on measuring platform-specific prevalence in the context of harmful online behaviours. This emphasis leaves gaps in our understanding of the overarching mechanisms driving these behavioural patterns, assessing their social impact, and designing proper, scientifically-informed and evidence-led regulatory interventions.

There are numerous challenges with designating OFCOM as the online safety regulator, tasked with making important decisions about systems and processes used to 'deal' with content and speech.[41] We indicated differences between sectors they currently regulate, historical rationales for regulating different regimes, problems with capacity and independence.[42] The Bill goes a step further in the wrong direction and endows extensive regulatory powers to the Secretary of State for Digital, Culture, Media and Sport (DCMS). One of these powers is the power to amend Schedule 1 (exempt services), and either add new services to the list of exemptions or remove some of those already exempt, based on assessing the risk of harm to the individuals.[43] This power gives the Government minister quite a lot of discretion, which, if misused, could lead to policing private messaging and communications channels such as Signal or Telegram. Other powers include the power to designate illegal and harmful

11/online_abuse_prevalence_full_24.11.2019_-_formatted_0.pdf (accessed 31 January 2022); Antigoni-Maria Founta et al., "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior" (2018) *ICWSM* 1–11.

41 Leiser and Harbinja, *supra* n. 5, pp. 86-87.

42 *Ibid.*

43 Sch. 1 of the Draft Online Safety Bill.

content, the power to change codes created by OFCOM, as well as the ability to set OFCOM's strategic priorities.[44] These powers give the Government minister impermissible discretion to compromise regulatory independence and escape parliamentary oversight. Parliament will likely revise these powers before the Online Safety Bills (OSB) is passed.[45] Surprisingly, however, OFCOM executives did not think that such a broad interference by the Executive was problematic. On the contrary, they applauded this approach and found nothing about it unusual.[46]

Even more concerningly, the Bill mentions human/digital rights quite tangentially. It vaguely mandates "duties about rights to freedom of expression and privacy."[47] This section in the Bill is very underdeveloped, and it seems disjointed from the rest of the proposal. Elsewhere it is categorised as "just a necessary add-on."[48] It is unclear how this section feeds into the rest of the Bill and how regulatory enforcement will focus on this duty. The regulator, OFCOM, is not clear on its scope either. Yet, it claims it is still confident that the Government will adequately resource it to tackle this essential, yet misplaced task.[49]

In summary, the OSB does not follow the White Paper's recommendations. Instead, it replaces the general duty of care with more specific

---

44 *Ibid*, c. 109.

45 See Draft Online Safety Bill (Joint Committee), "Formal meeting (oral evidence session): Draft Online Safety Bill", 4 November 2021, available at https://committees.parliament.uk/event/5673/formal-meeting-oral-evidence-session/ (accessed 31 January 2022).

46 Draft Online Safety Bill (Joint Committee), "Formal meeting (oral evidence session): Draft Online Safety Bill", 1 November 2021, available at https://committees.parliament.uk/event/5596/formal-meeting-oral-evidence-session/ (accessed 31 January 2022).

47 The Draft Online Safety Bill, c. 12.

48 Edina Harbinja, "The UK's Online Safety Bill: Safe Harmful, Unworkable?" (*Verfassungsblog*, 18 May 2021), available at https://verfassungsblog.de/uk-osb/ (accessed 31 January 2022).

49 Draft Online Safety Bill (Joint Committee), *supra* n. 45.

responsibilities that either (a) place a positive obligation on online services to protect certain forms of content, or (b) protect people from content that does not fall into the protected categories found in (a). This framework regulates content based on its risk. Service providers must conduct various assessments about illegal and harmful user-generated material and decide on the appropriate measures to 'deal' with this content. These will be developed by both the regulator and the regulated, under the considerable influence of the Secretary of State. As it simply sets out a structure for systems and processes, the Government is convinced that the framework is systemic rather than content-focused.[50] This belief, however, is not entirely correct. The proposal is highly content-focused and requires various user content moderation decisions and processes, even for 'content-agnostic' and 'non-content' duties.[51]

We find these specific categories to be even more unsatisfactory as they require categorisation of speech that will add further ambiguities likely to make meaningful content moderation even less successful. The Bill's material scope (illegal to children, harmful to children, illegal to adults, harmful to adults, journalistic content, the content of democratic importance, etc.) will only mean deploying technical solutions to comply. For example, under clause 46, content falls under this subsection if the provider of the service has **reasonable grounds** to believe that the nature of the content is such that there is a material **risk** of the content having, or **indirectly** having, a significant adverse physical or psychological impact.

---

[50] *Ibid*.

[51] Graham Smith, "The draft Online Safety Bill: systemic or content-focused?" (*Cyberleagle Blog*, 1 November 2021), available at https://www.cyberleagle.com/2021/11/the-draft-online-safety-bill-systemic.html (accessed 31 January 2022).

We have moved away, almost entirely, from John Stuart Mill's harm principle.[52] Furthermore, the definition is subject to Government interference.

In the next section, we question the appropriateness of the entire premise of a risk-based approach to the regulation of platforms and user-generated content. We will conceptualise this using various controversial speech and content examples to highlight the challenges in establishing causation in risk-based regulation. Undertaking this exercise will help explore the dangers risk regulation poses to requirements for human rights frameworks, many of which already create obligations for protecting speech using clearly defined legal tests. Over the remainder of the paper, we use various examples of controversial speech to help conceptualise the challenges and dangers of over-regulating harmful content through existing norms of the European Convention. In doing so, we open the door to further analysis of the Bill's compliance with the Convention and other human rights frameworks.

3 Risk versus harms

The Columbine shooting in 1999, the Charleston, South Carolina church shootings, and the Nicholas Cruz shootings at Marjory Stoneman Douglas High School in 2019, all share common characteristics. Each incident involved heavily armed teenagers unloading weapons on unsuspecting victims. After each incident, various actors sought to blame a wide range of cultural artefacts like extreme, 'satanist music', and violent video games (VVGs) for causing mass murder. Nicholas Cruz had a history of behavioural problems, made hateful social media comments about minorities, and was also a keen player of VVGs.[53]

---

[52] Discussed in Eric Barendt, *Freedom of Speech* (2nd edn., OUP, 2005), pp. 7–13.

[53] Colin Campbell, "A brief history of blaming video games for mass murder" (*Polygon*, 10 March 2018), available at https://www.polygon.com/2018/3/10/17101232/a-brief-history-of-video-game-violence-blame (accessed 31 January 2022).

Adam Lanza, the Sandy Hook murderer, was also a fan of video games. His preference? *Dance Dance Revolution*.

There is no suggestion that there is a correlation between dancing video games and mass murder, let alone causation. For many young people, gaming is a way of life, a hobby, and a passion. It is unsurprising to find a *correlation* between mass murder and video games. Yet, according to the American Psychological Association, there is no evidence that violence and murder can be attributed to the playing of VVGs.[54] Furthermore, no longitudinal study has ever found a direct correlation between regular video game usage and criminal violence like mass shootings.[55] Markey, a psychology professor at Villanova University who focuses on video games, found that men who commit brutal violence played VVGs less than the average male. About 20% were interested in VVGs, compared with 70% of the general population.[56] Why do so many 'experts' fall into the trap of inappropriately finding *causation* when there is not even evidence of the existence of a *correlation*? Why do so many politicians look to blame something tangible like playing video games rather than the far more apparent causes of mass shootings like readily accessible firearms, loopholes in background checks, the lack of accessible mental health services and healthcare?

As the *act* of mass murder was wrongly attributed to the *risk* of aggression and violence associated with playing video games, the history of assessing the risk of playing VVGs highlights the challenges in regulating content that may

---

54 American Psychological Association, "APA Resolution on Violent Video Games", February 2020 Revision to the 2015 Resolution, available at https://www.apa.org/about/policy/resolution-violent-video-games.pdf (accessed 31 January 2022).

55 There are also highly contestable links between indirect harms and extreme music, pornography, and horror movies.

56 Patrick Markey and Christopher Ferguson, *Moral combat: Why the war on violent video games is wrong* (Dallas, TX: BenBella Books, 2017).

contribute to the risk of harm. There are similar issues with applying a duty of care to content moderation and online safety. As discussed in the previous section, the obligation sets a low standard for removal of content: "the provider…has reasonable grounds to believe that the nature of the content is such that there is a material risk of the content having, or indirectly having, a significant adverse physical or psychological impact on a child", [57] or adult.

This obligation on a private actor will be judged subjectively with the platform's beliefs only mitigated by applying a reasonableness test that there is a material risk of content *directly* or *indirectly* having a significant adverse physical or psychological impact on a user. If game-streaming services like Twitch and video-sharing services that host recordings of gaming performances like YouTube are presented with evidence that, *prima facie,* VVGs are associated with a material risk of increased aggression among users that have behavioural issues like Nicholas Cruz and Adam Lanza. This raises the question: should they remove content under the threat of liability for failure to comply with a specific duty of care as a result of an increased material risk that access to this content might cause indirect harm?

This might sound ludicrous in today's climate. Still, post-Columbine, the prevailing narrative among journalists and politicians alike circled claims that VVGs did contribute to aggression and violence in young males. Over 300 scientific reports were undertaken into studying the effects of video games on aggression and violence in the aftermath of the Columbine incident, with politicians uniting over the development of a rating system for video games. It allowed both sides of the political divide to be seen to be doing something about mass shootings beyond the obvious. Yet time, and science, have corrected the

[57] The Draft Online Safety Bill, cls. 45(3) and 46(3).

erroneous belief that VVGs are linked to aggression and violence in young males. Had the 'Online Safety Bill' been introduced in the early 2000s, we may be looking at a very different world for gamers and content services that host gaming platforms. More recently, allegations circled that the internet was causing depression in adolescents. In 2015, it was reported that heavy web use harms a child's mental health.[58] Yet research has conclusively stated that internet usage has not caused any downturn in children's mental health; alternatively, parents and teachers believe that the internet is the cause at nearly twice the rate of children themselves.[59]

It is no surprise that risk is at the heart of techno-regulation. According to Gellert, the GDPR was designed around the concept of 'compliance risk.'[60] In simple terms, failure to comply with the regulatory requirements of the GDPR increases the risk of harm to data subjects' fundamental right to the protection of personal data. Gellert refers to Article 35 of the GDPR:

> Where a type of processing (. . .) is likely to result in a *high risk* to the rights and freedoms of natural persons, the controller shall, before the processing, *assess the impact* (. . .) on the protection of personal data.

For Gellert, risk management requires an examination of an event (lack of compliance) and its consequences (risks to DS as a consequence). But the OSB

---

58 Sonia Livingstone, "If we can't prove the internet makes children unhappy, we shouldn't lay the blame at its door" (*LSE Blog*, 21 March 2016), available at https://blogs.lse.ac.uk/parenting4digitalfuture/2016/03/21/if-we-cant-prove-the-internet-makes-children-unhappy-we-shouldnt-lay-the-blame-at-its-door/ (accessed 31 January 2022).

59 Rachel Rosen, "The Perfect Generation: Is the Internet Undermining Young People's Mental Health?" (*Parent Zone*, 17 March 2016), available at https://parentzone.org.uk/article/report-perfect-generation-internet-undermining-young-people%E2%80%99s-mental-health (accessed 31 January 2022).

60 Raphaël Gellert, "Understanding the notion of risk in the General Data Protection Regulation" (2018) 34(2) *Computer Law & Security Review* 279-288.

takes a different approach. It does not require a specific event, and the threshold to determine the effects is purely hypothetical. Consider section 46:

> Content is within this subsection if the **provider of the service has reasonable grounds to believe** that the nature of the content is such that there is a **material risk of the content having, or indirectly having, a significant adverse physical or psychological impact**.

Unlike article 35 of the GDPR, which considers risk to fundamental right of data protection, the Bill does require platforms to only assess the risk of the impact of a person.

Risk regulation requires risk analysis and effective risk management through reduction, control, response, and mitigation. The latter also requires explicit recognition that some speech will be harmful but does not pose any risk to the overwhelming majority of users. Risk calculations should only come on top of fulfilled compliance with fundamental rights obligations like article 10 of the ECHR, article 11 of the EU Charter, article 19 of the ICCPR, particularly the right to receive information. Compliance is not possible without a significant number of technical measures *and* removing the no general monitoring obligation for service providers.[61]

The OSB does not offer any protection for historical content that does not contribute to a democratic debate or journalistic content. Thus, content that may not be considered harmful at the time of posting might be removed retroactively upon evidence that the content may be detrimental to today's users. As the OSB offers no protection for creative content that neither political nor journalistic

[61] Art. 15, e-Commerce Directive, *supra* n. 8.

content, cultural artefacts may be threatened by powerful lobby groups that press regulators to remove additional content categories.

4 The shifting of cultural norms

From 1994-2004, one of the most successful programs in the NBC network's catalogue was the *Friends* sitcom. One of its more infamous storylines involved the actress Kathleen Turner playing the role of Chandler Bing's father, a transgender woman. The trans community has accused the TV programme of blatant transphobia, including using phrases like 'hermaphrodite', lines like "Don't you have a little too much penis to be wearing a dress like that?" alongside blatant misgendering, forced feminisation tropes, and blatant homo/transphobic content. Increasing awareness of trans rights alongside 'cancel culture' strategies could put political pressure on platforms to take a stance against historical content that reinforces historical biases and stereotypes. This concern is not without precedent. In May 2018, #MeToo pressure groups called for a boycott of hip-hop artist R Kelly's music. Spotify removed the catalogues of both R Kelly and rapper XXXTentacion. The former's catalogue was removed after a *documentary* made allegations of child sex trafficking. Although he was eventually convicted in 2021 and his catalogue eventually reinstated, his tracks are still not included in Spotify's curated lists or promoted tracks.[62] There is a certain oddity bordering on the hypocrisy that a Conservative Government actively legislating against 'cancel culture' has created a statutory duty to remove material that could pose a risk of making children racists.

---

[62] Laura Snapes, "R Kelly: Time's Up campaign against me is 'attempted lynching of a black man'" (*The Guardian*, 1 May 2018), available at https://www.theguardian.com/music/2018/may/01/r-kelly-times-up-muterkelly (accessed 31 January 2022).

The transphobia in *Friends* and the R Kelly case highlight how social and political norms can shift over time, and with these shifts come changes in what speech is considered acceptable. Moreover, it is often speech itself that effectuates these shifts. For these reasons, we should be uncomfortable when the Government claims the power to censor or punish private speech, even when it is speech that we find abhorrent. In our form of parliamentary democracy, the Government can (with few exceptions) say what it wants. Still, this power should be used humbly and with the recognition that the current majority will not hold power forever. We should be deeply sceptical of Government power to punish speakers for speech that lacks concrete evidence that is harmful.

Although this speech is not directly attached to specific harm, and these were measures initiated by private actors without any pressure from state actors, consider the present climate for 'cancelling' those whose speech we do not like. The controversy over these symbols teaches important lessons to all concerned about scepticism of Government power and the importance of free speech. Social and political norms shift over time, and with these shifts come changes in what speech is considered acceptable. Moreover, it is often speech itself that effectuates these shifts. For these reasons, we should be uncomfortable when the Government claims the power to censor or punish private speech, even when it is speech that we find abhorrent.

A risk-based approach to content moderation risks, the actual politicisation of enforcement, and the 'risk managers problem', such as the backlash after school shootings in America, demonstrate the effects of experiencing traumatic events and eliciting passionate responses. The associated public pressure usually results in grandstanding politicians and special interest groups' active involvement in regulatory reform, without evidence that the *harm* is effectively managed by targeting the *risk*. Thus, the effectiveness of risk regulation is often undone by the corrupting of the regime by politicised

stakeholders and dependent regulators (as noted above). What *risk* is Spotify mitigating by removing artists' catalogues from their service? What harm are they preventing? The regulator is susceptible to both political interference and pressure from special interest and civil society groups that disproportionately target specific harms that manifest themselves on social media platforms. The real risk under the proposed model is to the integrity of the regulator. Not only is OFCOM required to adapt to shifts in preferences and objectives, but they also risk misaligning the risks to society with political dangers to the regulator.

One can argue that risk has two meanings – linguistic and technical. Linguistically, the risk is a future, possible danger, i.e., as "an eventual danger was foreseen to some extent." In a technical sense, the risk is two-fold. Its constitutive elements exist in a symbiotic relationship: forecasting future events (both negative and positive) and making decisions. A decision-making strategy based on the prediction of future events. One can argue that any decision relating to risk requires objective facts and a subjective view about the desirability of what is to be gained or lost by the decision. But in a fundamental rights framework, one cannot allow the State's subjective view about the desirability of speech to guide decisions about managing the risk associated with the speech. Doing so opens the door to tyranny.

It is clear from the approach taken in the OSB that the precautionary principle is at the heart of the specific duties of care found within. How do we manage risk when users generate the content and the harm? The precautionary principle permits intervention when there are many indications of seriousness but incomplete knowledge about the scope and severity of adverse impacts. Interference is allowed when there are partly unknown cause-effect relations between applications and societal effects (e.g., 'systemic risks'), and the nature of seriousness can (only) be estimated in qualitative terms (not enough knowledge for quantitative risk assessments). Adopting a risk-based perspective as a

framework of governance requires introducing internal risk management systems to govern the activities of users and their content. Yet, the OSB gives little guidance on how the adoption of risk-based tools by platforms should look. This omission risks the development of a variety of ad hoc tools to comply with a statutory obligation at the expense of a fundamental right.

Risk regulation sets standards based on assessments of risk of activity to health, safety, well-being, fundamental rights, economics, etc. As it shifts the legal obligations associated with the precautionary principle onto private actors (thus becoming the prevailing directive for content-related decision-making), risk regulation is increasingly appealing as a balanced and proportionate means of transfers to powers of regulation from the State to the private sector.[63] Levels of risk regulation can be tightened to enhance regulatory performance and facilitate change. It is neither 'free-standing', technical nor mechanical. It does the latter by 'constructing' challenges; for example, constructing abstract claims that digital technologies infringe human dignity or "harms the child's autonomy by manipulating their limited knowledge." But interferences with fundamental rights do not operate holistically. They require preciseness to pursue a legitimate aim while satisfying the quality of law tests.[64]

---

[63] Christopher Hood, Henry Rothstein, and Robert Baldwin, *The Government of Risk: Understanding Risk Regulation Regimes* (OUP, 2001), p. 4.

[64] Accessibility: *Sunday Times v UK* 6538/74 [1979] ECHR 1 (26 April 1979), §47: "the law must be sufficiently clear in its terms to give individuals an adequate indication as to the circumstances in which and the conditions on which the authorities are empowered to resort to any such measures"; On Foreseeability, see *Sunday Times* judgement 1979, §49: "Citizen must be able-if need be with appropriate advice- to foresee, to a degree that is reasonable in the circumstances, the consequences which a given action may entail". A test developed recently in the jurisprudence of the European Court of Human Rights has resolved to discuss whether a measure is compatible with the rule of law. This rule explores the set of safeguards around a given measure. The rule exists as an additional consideration against the certainty of a law, so if there is any wavering around the accessibility and foreseeability of a given measure's compatibility with the rule of law can be considered to bring about a decision. This test began in cases relating to secret surveillance where accessibility and foreseeability of a law would be

5 Convention compliant?

With few exceptions, article 10 of the European Convention provides that "Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers." In (very) broad terms, article 10 prohibits States from regulating speech based on its contents. Generally speaking, this prevents states from censoring expression or punishing people for what they say. The Convention rarely permits Government officials to prevent individuals from speaking or penalise them for expressing their views.

Initially, the article only imposed negative obligations on States ("...without interference by public authority..."). Still, over the past 20 years, the ECHR has developed a body of case law that says effective exercise of certain freedoms does not depend merely on the State's duty not to interfere but may require positive obligations of protection even in the sphere of relations between private individuals.[65] In deciding whether a positive obligation under article 10 exists, regard must be had to the kind of expression rights at stake: their capability to contribute to public debates, the nature and scope of restrictions on

limited in other circumstances (*Rotaru v Romania* [2000] ECHR 192; *Liberty and Others v United Kingdom,* Application no. 58243/00 *and Sallinen and Others v Finland*, Application no. 50882/99), but it has become part of considerations in other cases (see *Gillan and Quinton v United Kingdom*, Application no. 4158/05, *Stefanov v Bulgaria,* Application no. 73284/13, and *Prezhdarovi v Bulgaria,* Application no. 8429/05).

65 Under Article 2, see *McCann and Others v the United Kingdom*, 27 September 1995, Series A no. 324 at §161, and *Osman v. the United Kingdom*, 28 October 1998, Reports of Judgments and Decisions 1998-VIII, §§ 115-117); Article 3 (see *Assenov and Others v Bulgaria*, 28 October 1998, Reports of Judgments and Decisions 1998-VIII, §102), Under Article 8 (see, amongst others, *Gaskin v the United Kingdom*, 7 July 1989, Series A no. 160, §§ 42-49) and Article 11 (see *Plattform "Ärzte für das Leben" v Austria*, 21 June 1988, Series A no. 139, § 32).

expression rights, the ability of alternative venues for expression, and the weight of countervailing rights of others or the public.[66]

The Court has arguably conceded that a positive obligation arises for the State to protect the right to freedom of expression by ensuring a reasonable opportunity to exercise a right of reply.[67] Moreover, the Court has stressed that States are required to create a favourable environment for participation in public debate by all the persons concerned, enabling them to express their opinions and ideas without fear.[68] The Court has also made it clear that it will offer some speech beyond political debate a heightened degree of protection. This protection extends to all public interest matters, stating there is "no warrant" for distinguishing between the two.[69] Furthermore, the Court seems to recognize that article 10 places a positive obligation on States to ensure a right to receive information: "Not only do the media have the task of imparting such information and ideas: the public also has a right to receive them."[70]

Freedom of speech is eternally radical because a person's right to speak their mind and challenge prevailing orthodoxies, no matter how controversial, is what Justice Oliver Wendell Holmes referred to as "freedom for the thought that we hate"— is so rare in human history. Acknowledgements that speech is powerful are often predicates to censoring it. For instance, Feldman-Barrett argues that certain kinds of speech are so powerful that they should be considered violence because they can lead to chronic stress, which can do serious physiological damage.[71] Acknowledgement of how powerful speech can serve as

---

66 *Appleby and Others v The United Kingdom*, Application no. 44306/98, ECHR 2003-VI, §§ 42-43 and 47-49.

67 *Melnychuk v Ukraine* (dec.), no. 28743/03, ECHR 2005-IX.

68 *Dink v Turkey*, nos. 2668/07, 6102/08, 30079/08, 7072/09 and 7124/09, 14 September 2010, § 137.

69 *Thorgeir Thorgeirson v Iceland*, 1992, § 64.

70 *Sunday Times v United Kingdom*, *supra* n. 64, § 65.

71 Lisa Feldman Barrett, "When is Speech Violence?" (*New York Times*, 14 July 2017),

an argument in favour of restricting it. Under this view, the OSB should protect us from entire categories of speech should be denied because some people are psychologically susceptible to content that others are not.

It is important to remember that those who seek to suppress speech always claim (and often have) pure motives. The McCarthy era was defined by communist witch hunts because "those with power to control the words and fates of others saw the threat [of communism] as real and deadly, always on the verge of rearing its destructive head."[72] Similarly, Chauvin argues that those seeking to restrict hurtful or hateful speech today often do so based on "laudable instincts" to make our public spaces "inclusive for all", particularly "those who have been traditionally under-represented" in those spaces.[73] One can certainly debate whether each of these groups has accurately evaluated the threat posed by the speech they wish to see restricted. However, the point is not whether those who want to restrict speech are right about the dangers they believe speech poses. Instead, in their belief that they are correct, they will take whatever steps deemed necessary to suppress speech they believe is harmful.[74] Additionally, it is worth noting that Governments have long claimed that they are promoting safety and equality by suppressing speech.

The Canadian Supreme Court struck down a broadly worded ban of false "statement, tale or news."[75] The Court ruled that all communications which

available at https://www.nytimes.com/2017/07/14/opinion/sunday/when-is-speech-violence.html (accessed 31 January 2022).

72 Stephen Carter, "We Can Fight for Racial Justice While Tolerating Dissent" (*Bloomberg*, 11 June 2020), available at https://www.bloomberg.com/opinion/articles/2020-06-11/george-floyd-protesters-can-fightracism-while-tolerating-dissent (accessed 31 January 2022).

73 Chauvin, citing Erwin Chemerinsky, "The Challenge of Free Speech on Campus", (2018) 61 *Howard Law Journal* 585, pp. 588–89.

74 Carter, *supra* n. 72; Chauvin, *ibid*., pp. 596–97.

75 *R v Zundel* [1992] 2 S.C.R. 731.

convey or attempt to convey meaning are protected by section 2(b) of the Canadian Charter of Rights and Freedoms, guaranteeing freedom of expression.[76] Courts have also viewed overly broad prohibitions on false and misleading speech unfavourably. Courts tend to look down on systems that amount to prior restraint. The appropriate remedy for harmful speech is more speech, not forced silence. "Any system of prior restraints of expression comes to this Court bearing a heavy presumption against its constitutional validity."[77]

The Bill set forth a regulatory framework whereby a regulator can class entire categories of speech harmful at the whim of the Secretary of State, while imposing specific duties of care on platforms to prevent harms associated with these categories under the threat of sanctions from the regulator. Compliance will require technical measures that amount to a system of prior restraint for content that *may* cause harm.

6 Conclusion

The online 'safety' saga in the UK continues, and the Secretary of State wishes to make the law 'watertight' and timeproof.[78] However, this article demonstrates the little consistency and evidence of reasonable justification for interfering with article 10 rights in the upcoming 'platform law'. Arguably, the Government is trying to score some 'easy' political points by satisfying the child protection and safety camp. Consider Parent Zone's call for the duty of care "to be extended to a duty to take action on both emerging and longstanding functionalities, and *act*

---

[76] *Ibid*., para. 23.

[77] *The New York Times v Sullivan* 403 U.S. 713 (1971) at 714 (quoting *Bantam Books v Sullivan*, 372 U.S. 58,70 (1963)).

[78] Draft Online Safety Bill (Joint Committee), *supra* n. 45.

in users' interests as soon as they become aware they are at risk of harm."[79] But the Bill achieves child protection and online safety at the expense of legal certainty, the rule of law, and fundamental rights. The Bill, in its current form, is likely to contravene established human rights principles. The regulator, with questionable independence, will have to make difficult decisions around the nature of this content, its removal, and subsequent challenges.

The State should not require, nor coerce, intermediaries to remove otherwise protected speech that the Government cannot prohibit directly. Such obligations violate principles of foreseeability, fairness, transparency, and non-discrimination. A duty of care on 'very large networks' risks ominously constraining emergent platforms to major companies, such as Facebook's, baselines of what is permitted and prohibited.

As speech is generative, it is a safe presumption that any list of 'harmful' speech will never contract. The thoughts responsible for the speech do not disappear into a vacuum. This claim begs a second related question: where is prohibited and harmful speech meant to go? How does the Government measure this muted speech and the harms arising from super concentrated groups of users congregating on 'very small' platforms? A social media platform cannot block undesirable speech not allowed on in the first place, nor removed from an otherwise-viable network that never emerged because of pre-emptive baselines. Today's dominant platforms would set standards for speech and impose reporting burdens on yet-to-be platforms. These outcomes would impart permanence to today's largest platforms; thus, defeating the stated purposes of increasing competition and contestability. It is not, as some assert, discombobulation that accompanies free speech that 'threatens democracy', but

---

[79] Parent Zone, "The Online Safety Bill: does the duty of care go far enough?" available at https://parentzone.org.uk/article/online-safety-bill-duty-care (accessed 31 January 2022).

the elitist regulation of that storm. As Federal Communications Commission commissioner Brendan Carr put it, "I think a lot of regimes around the world would welcome the call to shut down political opposition on the grounds that it is harmful speech."[80]

[80] The Federalist Staff, "FCC Commissioner Carr Pushes Back On Zuckerberg's Call To Censor The Internet" (*The Federalist*, 19 April 2019), available at https://thefederalist.com/2019/04/10/fcc-commissioner-carr-pushes-back-zuckerbergs-call-censor-internet/ (accessed 31 January 2022).