Journal of Media Law

ISSN: 1757-7632 (Print) 1757-7640 (Online) Journal homepage: www.tandfonline.com/journals/rjml20

Regulating online harms: an examination of recent developments in the UK and the US through a free speech lens

Eliza Bechtold

To cite this article: Eliza Bechtold (2024) Regulating online harms: an examination of recent developments in the UK and the US through a free speech lens, Journal of Media Law, 16:2, 358-389, DOI: 10.1080/17577632.2024.2395094

To link to this article: https://doi.org/10.1080/17577632.2024.2395094

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 29 Aug 2024.

Submit your article to this journal

Article views: 6579

View related articles

View Crossmark data

Citing articles: 2 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=rjml20

JOURNAL OF MEDIA LAW
2024, VOL. 16, NO. 2, 358–389
https://doi.org/10.1080/17577632.2024.2395094

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Regulating online harms: an examination of recent developments in the UK and the US through a free speech lens

Eliza Bechtold

School of Law, University of Aberdeen, Aberdeen, Scotland

ABSTRACT

This article examines recent development in the regulatory landscape of online harms in the UK and US through a free speech lens. In so doing, it undertakes a comparative analysis of the UK's Online Safety Act 2023 and laws in US states that is grounded in UN guidance advising that approaches to regulating online harms be grounded in human rights. Ultimately, this examination leads to the conclusion that the Online Safety Act and similar efforts at the state level in the US do not reflect human rights based approaches to regulation in this area. While American courts are reigning in state legislation based on the expansive protection of free speech under the First Amendment, the constitutional restraints in the UK are weaker. This raises concerns regarding the adequate protection of the free speech rights of internet users, particularly in the UK.

ARTICLE HISTORY Received 29 February 2024; Accepted 30 July 2024

KEYWORDS Free speech; Online Safety Act 2023; First Amendment; online speech

Introduction

The harm flowing from the use of online platforms is increasingly occupying the attention of legislators across the globe, including in the UK and the US. While the UK's newly enacted Online Safety Act 2023 (OSA) is one of the most ambitious examples of legislative efforts in this area, several US states are also passing laws targeting online harms.[1] These laws reflect a recognition of the potential harm to both children and adults flowing from access to

CONTACT Eliza Bechtold eliza.bechtold@abdn.ac.uk

[1]See, e.g., Texas (HB 1181, Tex Sess Law Serv (Vernon's)); Arkansas (Act 689 of 2023). While bills targeting online harms are pending at the federal level in the US, at the time of writing, there is no federal regulation of online harms in the US. Accordingly, federal legislative efforts lie outside the scope of this article.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

online content as well the increasing power that platforms wield over public discourse in the digital age. They also illustrate diverging conceptions regarding the most serious harms in this context and raise novel questions concerning to what extent aggressive regulatory efforts may threaten the free speech rights of internet users. Indeed, as new regulatory frameworks are introduced directed to protecting internet users from the online harms, there is a risk that such frameworks do not adequately consider the rights of individuals and, in certain cases, may imperil them.[2]

In the context of multifaceted governance issues like platform regulation, human rights offer a well-established collection of rules and a useful normative framework.[3] In 2021, the UN Human Rights Office emphasised the critical importance addressing the challenges from harmful online content by adopting approaches that are *grounded* in human rights.[4] Highlighting that virtually every country to enact legislation tackling online harms has jeopardised human rights, the UN observes that such laws generally suffer from similar problems. These include poor definitions for what constitutes unlawful or harmful content, outsourcing regulatory functions to companies, and an over-emphasis on artificial intelligence, including algorithms.[5] It emphasises that human rights based approaches to regulation should, among other things, focus 'on process, not content' and '[l]ook at *how* content is being amplified or restricted and '[e]nsure actual people – not algorithms – review complex decisions'.[6]

With these considerations in mind, this article undertakes a comparative analysis of the OSA and recent regulatory efforts at the state level in the US through the lens of freedom of expression.[7] As the regulation of online harms is an increasingly popular policy initiative for governments in various jurisdictions,[8] and given the potential transnational free speech implications of regulatory efforts in this area, a comparative analysis is particularly instructive. Such an analysis highlights both similarities and divergences in efforts to address shared challenges within the broader context of the adequate protection of human rights.

[2]See Emma Bond and Andy Phippen, *Safeguarding Adults Online: Perspectives on Rights to Participation* (BUP 2022) 34.

[3]See Yasmin Afina and others, 'Towards a global approach to digital platform regulation' (January 2024) *Chatham House Research Paper* <https://www.chathamhouse.org/2024/01/towards-global-approach-digital-platform-regulation/04-establishing-global-frameworks> accessed 13 February 2024.

[4]'Press Briefing, Online Content Moderation and Internet Shutdowns' (*UN Human Rights Office*, 14 July 2021) <https://www.ohchr.org/sites/default/files/Documents/Press/Press_briefing_140721.pdf> accessed 1 February 2024.

[5]ibid.

[6]ibid.

[7]For the purposes of this article, freedom of expression and freedom of speech are used interchangeably. While the regulation of online harms also raises significant privacy concerns, this lies outside the scope of this article.

[8]The UN notes that approximately 40 new laws relating to social media were adopted worldwide between 2019 and 2021, with 30 under consideration in 2021. See (n 4).

This analysis proceeds in five parts. The first part contextualises the topic by offering a broad sketch of the most relevant protections afforded to online speech in the UK and the US. The second part analyses the OSA with respect to some of the most significant free speech concerns arising from the new regulatory framework. Next, recent developments at the state level in the US are examined, including successful First Amendment challenges to online harms legislation. This is followed by an examination of the increasing popularity of age-verification as a regulatory tool in the UK and the US, as well as the concerns this raises regarding the adequate protection of freedom of expression on the internet. The final part offers some preliminary observations, including that the OSA and recent US legislative efforts contravene UN guidance directed to ensuring regulatory approaches are firmly grounded in human rights and, in so doing, jeopardise the free speech rights of internet users.

Existing frameworks for the protection of online speech in the UK and US

The OSA and recent regulatory efforts at the state level in the US provide an opportunity for useful comparative analysis, with some similar approaches in relation to the regulation of content within different free speech frameworks while offering, at times, divergent views of the harm that justifies government intervention. An understanding of the fundamental differences in the free speech frameworks in the UK and the US elucidates why legislative efforts that raise little constitutional concern in the UK are subjected to rigorous scrutiny – and often fail as a result – in the US. They also aid in appreciating the challenges in striking the balance between freedom and regulation in this area in the US, where the protection of speech often predominates over legitimate government interests in ameliorating online (and offline) harms. This section sketches the current protections for free speech, particularly with respect to online content, in these jurisdictions.

Broad latitude for regulating online speech in the UK

While the right to free speech did not historically enjoy distinct constitutional status in the UK, English courts have long recognised that such a right exists in the common law.[9] Its contemporary approach is informed

[9]See, e.g., *Attorney-General v Guardian Newspapers Ltd (No 2)* [1990] 1 AC 109, at 283–284 (in which Lord Goff opined that there was, in principle, no difference between English law and Article 10 with respect to the right to freedom of expression; *R v Secretary of State for the Home Department*; *Derbyshire County Council v Times Newspapers Ltd* [1993] AC 534. See also Eric Barendt, 'Freedom of Expression in the United Kingdom Under the Human Rights Act 1998' (2009) 84 Ind L J 851.

by the European Convention on Human Rights (ECHR), incorporated into domestic law by the Human Right Act 1998. Article 10 of the ECHR protects freedom of expression. In adjudicating on Article 10 cases, the European Court of Human Rights (ECtHR) applies a balancing test that weighs the competing interests of the speaker with the stated aims of the government.[10] Specifically, the ECtHR makes a determination as to whether an interference with an individual's Article 10 rights was necessary in a democratic society for achievement of one of the 'legitimate aims' identified in paragraph 2.[11] This involves a determination of whether the interference was prescribed by law,[12] pursues one of Paragraph 2's 'legitimate aims',[13] and whether the interference was proportionate to the legitimate aim pursued and corresponded to a pressing social need.[14] Thus, an interference with an individual's right to freedom of expression only violates Article 10 if it fails to satisfy the requirements of Paragraph 2, that is, if the government's identified 'legitimate aim' does not prevail over the free speech interests of the speaker.

While the ECtHR interprets Article 10 as applying to information or ideas that offend, shock or disturb, it subjects such freedom to many restrictions.[15] It holds that not all speech is worthy of inclusion in public debate and the State may and should act as the arbiter in determinations of such worthiness. By way of example, Member States may restrict speech that does 'not contribute to any form of public debate capable of furthering progress in human affairs'.[16] Similarly, the Council of Europe expressly eschews 'absolute liberalism' in the realm of speech regulation in favour of an approach predicated on the notion that not all ideas are deserving of circulation and that the right to express one's ideas may be outweighed by competing societal interests.[17]

Thus, while the ECtHR holds that an individual taking part in a public debate on a matter of general concern is permitted a degree of exaggeration or even provocation, that is, 'to make somewhat immoderate statements', it imposes significant limitations on what types of ideas may be permitted in public discourse as well as the *ways* in which such ideas may be expressed by imposing duties and responsibilities on the exercise of Article 10

[10]See, e.g., *Wingrove v United Kingdom*, App no 17419/90 (ECtHR, 25 November 1996).
[11]ibid.
[12]That is, whether the regulation was sufficiently clear and precise to enable a citizen to regulate his conduct in a way that is compatible with the law. See *Sunday Times v United Kingdom* App no 6538/74 (ECtHR, 26 April 1979).
[13]These include public safety, the protection of health or morals, and the protection of the reputation or rights of others.
[14]See, e.g., *Vejdeland v Sweden,* App no 1813/07 (ECtHR, 9 February 2012).
[15]See, e.g., *Mamère v France*, App no 12697/03 (ECtHR, 7 February 2007).
[16]*Otto-Preminger-Institut v Austria*, App no 13470/87 (ECtHR, 20 September 1994) para 49.
[17]See 'Report on the Relationship between Freedom of Expression and Freedom of Religion: the Issue of Regulation and Prosecution of Blasphemy, Religious Insult and Incitement to Religious Hatred' (*Venice Commission*, 23 October 2008) <https://www.venice.coe.int/webforms/documents/default.aspx?pdffile=CDL-AD(2008)026-e> accessed 10 May 2024.

rights.[18] This stems from a conceptualisation of free speech as a 'dialogue and a spirit of compromise, necessarily entailing various concessions on the part of individuals or groups of individuals, which are justified in order to maintain and promote the ideals and values of a democratic society'.[19] This requires the subordination of individual interests to those of the group in certain cases to ensure 'the fair and proper treatment of people from minorities and avoid any abuse of a dominant position'.[20] This explains why the ECtHR does not regard content-based proscriptions on expression as raising free speech concerns and regards proscriptions on particular categories of expression, such as hate speech, as justified interferences with Article 10 rights.[21]

While the ECtHR has yet to opine directly on regulations like those found in the OSA, which place compulsory duties on online platforms, its case law offers guidance with respect to the regulation of online speech. In *Delfi v Estonia* – its first case addressing the applicability of Article 10 to online speech – the ECtHR held that internet news portals could be liable for the speech of third parties due to paragraph 2's 'duties and responsibilities' when 'they provide for economic purposes a platform for user-generated comments on previously published content and some users … engage in clearly unlawful speech, which infringe the rights of others'.[22] The court emphasised in *Delfi* that the case concerned a large internet news portal that published its own news articles and invited reader comments, as opposed to 'other fora on the Internet where third-party comments can be disseminated'.[23] It further opined that in light of 'the particular nature of the Internet', the duties and responsibilities on news portals under Article 10 may differ from those of a traditional publisher as regards third-party content.[24] These may include the imposition of liability for failing to take measures to remove 'clearly unlawful comments without delay' even absent notice from the alleged victim or third parties.[25]

In 2021, the ECtHR ruled on another case involving online speech, addressing an Article 10 challenge to criminal liability for the offence of incitement to hatred imposed on a politician for failing to remove hateful

[18] See, e.g., *Kuliś and Różycki v Poland*, App no 27209/03 (ECtHR, 6 January 2010) para 39; see also *Mamère v France*, App no 12697/03 (ECtHR, 17 October 2006) para 25.
[19] *Gough v United Kingdom*, App no 49327/11 (ECtHR, 23 March 2015).
[20] ibid para 168 (referencing *Chassagnou and Others v France*, App nos 25088/94, 28331/95 and 28443/95 (ECtHR, 29 April 1999) para 112; Leyla *Şahin v Turkey*, App no 44774/98 (ECtHR, 10 November 2005) para 108, and *Bayatyan v Armenia*, App no 23459/03 (ECtHR, 7 July 2011) para 126).
[21] Provided that any such restrictions are proportionate to a legitimate government aim. See *Jersild v Denmark*, App no 15890/89 (ECtHR, 23 September 1994).
[22] *Delfi v Estonia*, App no 64569/09 (ECtHR, 10 October 2013) para 115.
[23] ibid para 116. See also *Zöchlin v Austria*, App no 4222/18 (ECtHR, 5 September 2023).
[24] ibid para 113.
[25] ibid para 159. See also Dirk Voorhoof, 'Same Standards, Different Tools? The ECtHR and the Protection and Limitations of Freedom of Expression in the Digital Environment' in *Council of Europe, Human Rights Challenges in the Digital Age: Judicial Perspectives* (Council of Europe Publishing, 2020).

comments on his Facebook 'wall'.[26] The court noted that the issue in the case was not the applicant exercising *his* right to freedom of expression but, rather, for his 'lack of vigilance and reaction' to a third party's comments.[27] In finding no violation of Article 10, the court again emphasised the duties and responsibilities under paragraph 2, in this context the applicant's 'status' as 'holder of the "wall" of his Facebook account', which entailed specific obligations (heightened due to his status as a politician).[28] The court also opined that 'there is, without any doubt, a shared liability between the holder of a social media account and the operator of the network'.[29]

The above cases, while not directly addressing the question of liability for social media platforms for user speech, are of value to those seeking to understand what the application of Article 10 to large online platforms like TikTok and Facebook might look like. For example, while the liability in *Sanchez* concerned a Facebook user, not Facebook, the court articulated a belief in 'shared liability' for platforms and users. Additionally, these cases offer insight into the court's broader approach to online speech, including that holding services and individuals liable, both criminally and civilly, for the expression of others does not raise Article 10 problems because those operating and using platforms undertake duties and responsibilities relating thereto.

Finally, it is worth highlighting the relevance of horizontal effect in this area, particularly to the extent it aids in understanding the stark divergences in the approaches of the UK and the US. The principle of horizontality denotes the application of certain fundamental rights to disputes by and between individuals, rather than to disputes between individuals and the state.[30] Assessing the operation of horizontality within constitutional law is an important question of constitutional construction, that is, how fundamental rights enter private relations within a particular legal framework.[31] With respect to the application of the ECHR, it is generally uncontested that ECHR rights generate some level of horizontal effect at the national level and that private parties may adversely affect an individual's enjoyment of at least certain rights.[32] The establishment and development of horizontal effect by the ECtHR is grounded in a theory of positive obligations on the state to protect the enjoyment of fundamental rights in the context of

[26] *Sanchez v France*, App no 45581/15 (ECtHR, 2 September 2021).
[27] ibid para 94.
[28] ibid para 100.
[29] ibid para 98.
[30] See Eleni Frantziou, *The Horizontal Effect of Fundamental Rights in the European Union: A Constitutional Analysis* (OUP 2019).
[31] ibid 1.
[32] Gavin Phillipson and Alexander Williams, 'Horizontal Effect and the Constitutional Constraint' (2011) 74 MLR 878.

relations between individuals or in conflicts between private parties and competing fundamental rights.[33] While not generally translated into distinct legal rights, positive obligations tend to be used as important arguments in Article 10 case law.[34]

As discussed below, the US approach is markedly different from the UK. A meaningful understanding of these differences is important to appreciating the benefits of a comparative examination in this area. These differences may be explained, in part, by the predominating principle of negative liberty as well as limited conceptions of horizontality in the American constitutional framework.

Strong protections for online speech in the US

The US is well-known as an outlier in the expansive protection it affords to freedom of expression, both online and offline. This is due, in large part, to the Supreme Court's (USSC) interpretation of the First Amendment, which is predicated on the marketplace of ideas, a metaphor first introduced in 1919 by Justice Oliver Wendell Holmes in his dissenting opinion in *Abrams v. U.S.* In this opinion, Justice Holmes opined that 'the best test of truth is the power of the thought to get itself accepted in the competition of the market … '.[35] Based on this reasoning, the Court holds that 'regardless of how pernicious an opinion or idea may seem, Americans must depend on its correction not from judges or juries but from the competition of other ideas'.[36] As a result, the Court generally prohibits the government from proscribing speech because of disapproval of the subject matter or viewpoint expressed and subjects such restrictions to strict scrutiny, its most rigorous standard of review.[37] Thus, unlike in the UK, content-based restrictions on expression, including those restricting speech based on viewpoint, are regarded as presumptively unconstitutional and rigorously scrutinised by courts.[38]

The USSC's inherent scepticism of content and viewpoint-based restrictions on expression is largely absent from the ECtHR's interpretation of Article 10, which permits regulation of categories of harmful expression.[39] Part of this scepticism may be explained by the USSC's emphasis on the

[33] Jean-François Akandji-Kombe, 'Positive Obligations under the European Convention on Human Rights: A Guide to the implementation of the European Convention on Human Rights (Human Rights Handbooks No 7)' (Council of Europe, 2007).
[34] Brittan Heller and others, 'Freedom of Expression: A Comparative Summary of United States and European Law' (*Transatlantic Working Group*, 3 May 2019) 10–11.
[35] *Abrams v US*, 250 US 616, 630 (1919).
[36] *Gertz v Welch, Inc*, 418 US 323, 339–40 (1974).
[37] Strict scrutiny requires that the government demonstrate that a particular regulation serves a compelling governmental interest and is narrowly tailored to achieve said interest, i.e. is the least restrictive means. See *Sable Commc'ns of Cal, Inc v Fed Commc'ns Comm'n*, 492 US 115, 126 (1989).
[38] See *RAV v City of St Paul*, 505 US 377, 395 (1992).
[39] See *Lilliendahl v. Iceland*, App no 29297/18 (ECtHR, 11 June 2020). This is not to suggest that the European approach does not place a high value on speech in public discourse. See, e.g., *Dichand and others v Austria*, App no 29271/95 (ECtHR, 26 May 2002).

development of an individual's character, opinion, and belief without government interference, which reflects the broader notion of negative liberty enshrined in the American Constitution, as well as the prevailing principle that the government has limited authority to regulate the ways in which rights are exercised.[40] In particular, the emphasis on individual identity in American free speech jurisprudence reflects the extent to which constitutional rights in the US are understood almost exclusively in terms of the relationship between the individual and the government.

This reflects the steadfast commitment to individualism and long-standing tradition of negative liberty that lie at the heart of American legal and cultural identity. For this reason, the rights guaranteed in the Bill of Rights – the first ten Amendments to the US Constitution – only protect rights holders from the acts of government, not those of private actors.[41] Additionally, the Constitution does not impose upon the government an obligation to protect the life, liberty, or property of citizens from invasion by private actors.[42] The predominant emphasis placed on individual identity in the American constitutional framework, may be explained, in part, by the country's historical origins. For example, in distinguishing the First Amendment from the English free speech tradition, the USSC remarked that it 'cannot reasonably be taken as approving English practices prevalent at the time of its adoption, but on the contrary the unqualified prohibitions laid down by the framers thereof were intended to give to liberty … the broadest scope that could be countenanced in an orderly society'.[43]

The state action doctrine is relevant in this context because it draws the line between the acts of government and the acts of private actors, which is fundamental to the USSC's approach to the First Amendment.[44] The Court opines that the doctrine protects 'a robust sphere of individual liberty' by distinguishing the government from individuals and private entities and by enforcing the boundary between the governmental and the private.[45] Accordingly, in each case in which the First Amendment is implicated, a court must determine whether a challenged act is that of the government, i.e. whether a particular restriction is sufficiently governmental in

[40] *Bowers v Devito*, 686 F2d 616, 618 (7th Cir 1982) (emphasising that the US Constitution is 'a charter of negative liberties; it tells the state to let people alone'). See also Michael Rosenfeld, 'Hate Speech in Constitutional Jurisprudence' in Michael Herz and Peter Molnar (eds), *The Content and Context of Hate Speech: Rethinking Regulation and Responses* (CUP 2012).

[41] *Hudgens v NLRB*, 424 US 507, 513 (1976). See also *Manhattan Cmty Access Corp v Halleck*, 139 SCt 1921, 1928 (2019) (holding that '[t]he Free Speech Clause prohibits only *governmental* abridgment of speech … The Free Speech Clause does not prohibit *private* abridgment of speech'.).

[42] *Castle Rock v Gonzales*, 545 US 748, 755 (2005).

[43] *Bridges v California*, 314 US 252, 264–65 (1941) (the Court further observes that 'no purpose in ratifying the Bill of Rights was clearer than that of securing' for the American people 'much greater' freedom of expression 'than the people of Great Britain had ever enjoyed').

[44] *Denver Area Ed Telecomm Consortium v FCC*, 518 US 727, 737 (1996).

[45] *Manhattan* (n 41) 1930.

character to constitute 'state action', thereby triggering the application of the First Amendment.[46]

While outside the scope of this article, it is worth noting that while, at first glance, the state action doctrine and horizontal effect appear to be contradictory principles – the former premised on the idea that fundamental rights may only be enforced against the acts of the government while the latter provides for the protection of fundamental rights in private relationships – many constitutional scholars argue that horizontality is present in the American constitutional framework. However, there is significant disagreement regarding to what extent and on what basis.[47] The point here is that the USSC's approach to the First Amendment is fundamentally rooted in negative liberty, which aids in understanding the significance of the distinction between state and private action in the context of speech regulation and the hostility to the types of speech regulations directed to ameliorating online (and offline) harm that are commonplace in the UK and other Council of Europe Member States.

Sweeping immunities for online platforms under the Communications Decency Act

Another key difference between American and British free speech frameworks is that online platforms in the US are granted sweeping statutory immunity. In 1996, the US Congress enacted Section 230 of the Communications Decency Act (CDA).[48] Section 230 provides that platforms may not be treated as the publisher or speaker of any content provided by users.[49] The immunity extends to the exercise of a publisher's traditional editorial functions, including deciding whether to publish, withdraw, or alter content created by a third party.[50] In enacting Section 230, Congress sought to encourage intermediaries to screen content without fear of liability, thus overriding the traditional treatment of publishers and distributors under statutory and common law.[51] In contrast to the EU's e-Commerce Directive, the CDA immunity applies regardless of whether an intermediary is aware of

[46] *Edmonson v Leesville Concrete Co*, 500 US 614, 619 (1991).

[47] For example, Stephen Gardbaum argues that while private actors are not bound by constitutional rights in the US, they are indirectly subject to (and may be adversely affected by) them because such rights govern the laws that private actors invoke and rely on against one another. As a result, constitutional rights may either prevent such laws from protecting certain interests, choices, and actions of one private actor against another altogether, or place significant limits on their ability to do so. He argues that the extent of the reach of individual rights into the private sphere defies the standard understanding of the US as creating a rigid public-private distinction in constitutional law, thereby epitomizing the vertical approach to this issue. Stephen Gardbaum, 'The "Horizontal Effect" of Constitutional Rights' (2003) 102 Mich L Rev 387. See also Mark Tushnet, 'The Issue of State Action/Horizontal Effect in Comparative Constitutional Law' (2003) 1 ICON 79.

[48] 47 USC s 230.

[49] ibid.

[50] ibid.

[51] See *Zeran v Am Online, Inc*, 129 F3d 327, 330 (4th Cir 1997).

objectionable content and/or whether such content is removed or disabled.[52] Thus, while there is some degree of overlap with respect to the protections afforded to intermediaries under the CDA and the e-Commerce Directive, the former provides significantly broader protection for intermediaries and is predicated upon fundamentally different policy considerations.

Concerns regarding offline harms caused by online platforms, particularly social media, have led to several bills at the federal level to amend or repeal the CDA, as well as leaders of large social media platforms like Google and X (formerly Twitter) testifying before Congress regarding efforts to tackle the purported harms flowing from their platforms.[53] To date, however, all legislative efforts have failed and Section 230 remains a key protection for platforms from liability for user content.[54]

Open questions regarding the application of the First Amendment to online speech

The USSC has observed that in the digital age, 'one of the most important places to exchange views is cyberspace, particularly social media, which offers relatively unlimited, low-cost capacity for communication all kinds, to users engaged in a wide array of protected First Amendment activity on any number of diverse topics'.[55] With that said, it has yet to rule directly on the question of the extent to which First Amendment protections attach to online speech, with lower courts acknowledging where social media fits in traditional First Amendment jurisprudence is 'not settled'.[56]

The Court has, however, weighed in on some of the First Amendment questions raised by online expression. For example, it holds that its jurisprudence 'provide[s] no basis for qualifying the level of First Amendment scrutiny that should be applied' to the internet.[57] Additionally, in 2019, it held that 'merely hosting speech … does not alone transform private entities into state actors subject to First Amendment constraints'.[58] Some lower

[52]See 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. The EU's Digital Services Act incorporates the e-Commerce Directive's existing liability exemption rules.

[53]See, e.g., 'Preview: Senate Judiciary Committee to Press Big Tech CEOs in Failures to Protect Kids Online During Landmark Hearing Today' (*US Senate Committee on the Judiciary*, 31 January 2024) <https://www.judiciary.senate.gov/press/releases/preview-senate-judiciary-committee-to-press-big-tech-ceos-on-failures-to-protect-kids-online-during-landmark-hearing-todayreview: Senate Judiciary Committee to P … | United States Senate Committee on the Judiciary> accessed 28 February 2024.

[54]See Chris Riley and David Morar, 'Legislative efforts and policy frameworks within the Section 230 debate' (*Brookings Institution*, 21 September 2021) <https://www.brookings.edu/articles/legislative-efforts-and-policy-frameworks-within-the-section-230-debate/> accessed 15 February 2024. See also Jeff Koseff, 'A User's Guide to Section 230, and a Legislator's Guide to Amending It (or Not)' (2022) 37 Berkeley Tech L J 757.

[55]*Packingham v North Carolina*, 137 SCt 1730, 1732 (2017) (internal quotations and citations omitted).

[56]See *Netchoice v Moody*, 546 FSupp3d 1082, 1090 (ND Florida 2021) (overruled on other grounds).

[57]*Reno v ACLU*, 521 US 844, 870 (1997).

[58]*Manhattan* (n 41) 1930.

courts have relied on these and similar holdings in finding that social media platforms have a First Amendment right to moderate content disseminated on their platforms.[59] There is a dispute, however, among lower courts regarding this issue. To date, the USSC has not opined directly on this and related questions.[60]

The expansive protection of free speech in the American constitutional framework, underpinned by a firm commitment to negative liberty and limited conceptions of horizontality, coupled with the CDA immunity present significant challenges for any regulatory efforts directed to tackling online harms. Recent decisions of lower courts in cases involving challenges to state legislation in this area highlight both how difficult it is for regulatory measures to meet the strict requirements imposed by the USSC and the significant outstanding questions regarding the regulation of online platforms. These and other issues are discussed below.

The UK's (overly) ambitious Online Safety Act

At just over 280 pages, the OSA reflects the government's mission to make the UK 'the safest place in the world to be online' by introducing a sweeping regulatory framework targeting myriad online harms in a single piece of legislation.[61] These harms include cyber-bullying, harassment, hate speech, fraud, terrorism, disinformation, and a variety of content – both legal and illegal – deemed harmful to children. It does this by, among other things, placing duties of care on regulated services, including social media platforms, photo or video-sharing services and instant messaging apps, as well as search services, in relation to illegal and harmful content. These duties include a complex suite of risk safety measures aimed at reducing the risk of harm to users. Some services, including large social media platforms (designated as 'Category 1 services' under the Act), have additional duties.[62] These duties include providing user identity verification options and protections for news publisher and journalistic content.[63]

Throughout the legislative process, the government made repeated assurances that the new framework would adequately protect freedom of

[59]See, e.g., *Netchoice v Florida*, 34 F4th 1196, 1203 (11th Cir 2022).
[60]While the Court recently issued a ruling in these cases, the majority opinion did not reach the merits of the parties' claims. See *Moody v Netchoice*, 2024 WL 3237685 (2024). See also discussion of recent developments in the US.
[61]See 'Britain makes internet safer, as Online Safety Bill finished and ready to become law' (*Gov.UK*, 19 September 202) <https://www.gov.uk/government/news/britain-makes-internet-safer-as-online-safety-bill-finished-and-ready-to-become-law> accessed 3 December 2023.
[62]See Online Safety Act (OSA) 2023, ss 71(1), 72(3)(a).
[63]See ibid ss 14–19. See also 'Implementing the Online Safety Act: Additional duties for 'categorised' online services (*Ofcom*, 25 March 2024) <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/additional-duties-for-categorised-online-services> accessed 1 July 2024.

expression and the OSA contains provisions directed to this end. For example, Section 1 dictates that regulated services be designed and operated such that 'users' rights to freedom of expression … are protected'.[64] Peter Coe describes the free speech duties imposed on regulated services as 'softer-edged' because the statute's language only requires services to 'take account of' or 'have regard to' them.[65] From the beginning, free speech advocates expressed concerns regarding the potential deleterious impact on free speech posed by the new regulatory framework.[66] While some of the most troubling proposals – including regulating 'legal but harmful content' directed to adults – were abandoned during the legislative process, significant concerns remain.[67] And while the UK government clearly had regard for the right to freedom of expression throughout the legislative process, this is not the same as *grounding* a regulatory framework in human rights. As the below examination demonstrates, the OSA does not reflect a human rights based approach to regulation.

At the outset, it is important to highlight two points. First, given how much of the implementation of the legislation is left to the regulator, Ofcom, it remains to be seen whether and to what extent these concerns are subsequently addressed. Second, due to the sweeping scope and breadth of the OSA's regulatory measures, an exhaustive analysis of all free speech issues arising from the new framework is too ambitious for a single article. Accordingly, this article focuses on some of the most concerning elements of the OSA from a free speech perspective. These are the OSA's definition of 'harm', the new false communications offence, and the age verification requirement for legal content deemed harmful to children. The first two elements are addressed, in turn, below. The age verification requirement is examined in the penultimate section, alongside similar regulatory efforts in the US.

The OSA's nebulous definition of harm

The terms 'harm' and 'harmful' appear 315 times in the OSA. As the legislation is directed to preventing online harms, repeated use of these

[64]ibid (OSA) s1. See also ss 17–19, 22.

[65]E.g., '[w]hen deciding on, and implementing, safety measures and policies, a duty to have *particular regard* to the importance of protecting users' right to freedom of expression within the law'. ibid s 22(2) (emphasis added). This is opposed to the 'hard edged' duties of care in relation to protecting users from particular types of content. See Peter Coe, 'Tackling online false information in the United Kingdom: The Online Safety Act 2023 and its disconnection from free speech law and theory' (2024) 15 *Journal of Media Law* 213–42.

[66]See, e.g., 'A Legal Analysis of the Impact of the Online Safety Bill on Freedom of Expression' (*Index on Censorship*, 2022) <https://www.indexoncensorship.org/wp-content/uploads/2022/05/Legal-analysis-of-the-impact-of-the-Online-Safety-Bill.pdf> accessed 1 June 2024.

[67]See, e.g., 'UK: House of Lords must reject the Online Safety Bill (*ARTICLE 19*, 30 January 2023) <https://www.article19.org/resources/uk-house-of-lords-must-reject-the-online-safety-bill/#:~:text=ARTICLE%2019%20recognises%20that%20some,to%20freedom%20of%20expression%20remain.> accessed 1 June 2024; Coe (n 65).

terms is hardly surprising. However, given the unprecedented scope and complexity of the regulatory framework, it would be reasonable to expect the government to adopt a precise definition of 'harm' to guide both the regulator, Ofcom, and regulated services. Instead, the OSA offers this: '"Harm" means physical or psychological harm'.[68] In 2022, the government released Explanatory Notes for the version of the Online Safety Bill as introduced in the House of Commons on 17 March 2022.[69] In this document, the government attempted to offer some clarity in relation to the scope of harm covered by the legislation, advising that harm 'could include physical injuries, serious anxiety and fear; longer-term conditions such as depression and stress; and medically recognised mental illnesses, both short-term and permanent'.[70] However, rather than providing much needed clarity, the Explanatory Notes highlight the nebulous nature and expansive scope of the purported harm the legislation covers.

While platforms are already taking proactive steps to remove harmful content, including illegal content, in relation to their community standards, the OSA introduces a compulsory framework that endows Ofcom with the authority to impose 'grave consequences' for failures to fulfil duties and responsibilities relating to regulated content.[71] To date, Ofcom has issued 390 pages of guidance that 'should be used by services in all circumstances when they are required to make a judgement on whether a piece of content is illegal in order to fulfil their duties under the Act'.[72] Ofcom's enforcement authority includes the ability to issue fines of up to £18 million or up to 10 percent of a service's qualifying worldwide revenue (whichever is greater).[73] In a press release following the passage of the OSA, the Department for Science, Innovation and Technology highlighted the severity of the sanctions framework, stating that companies that fail to comply with the Act 'will face significant fines that could reach billions of pounds' and 'their bosses may even face prison'.[74]

The UK government asserts that OSA 'mak[es] sure what is illegal offline is illegal online'.[75] However, policing online speech is of a different nature and scale than the offline context, in which public authorities, not private

[68]OSA (n 62) s 234(2).
[69]'Online Safety Bill: Explanatory Notes' (*Parliament.UK*, 2022) <https://publications.parliament.uk/pa/bills/cbill/58-02/0285/210285en.pdf> accessed 1 September 2023.
[70]ibid para 726.
[71]See, e.g. 'Protecting people from illegal harms online: Annex 10: Online Safety Guidance on Judgement for Illegal Content' (*Ofcom*, 9 November 2023). <https://www.ofcom.org.uk/__data/assets/pdf_file/0025/271168/annex-10-illegal-harms-consultation.pdf> accessed 15 February 2024.
[72]See ibid.
[73]OSA (n 62) Schedule 13 (4)(1).
[74]Gov.UK (n 61).
[75]'Britain makes internet safer, as Online Safety Bill finished and ready to become law' (*Department of Science, Innovation and Technology*, 19 September 2023) <https://www.gov.uk/government/news/britain-makes-internet-safer-as-online-safety-bill-finished-and-ready-to-become-law?ref=everythinginmoderation.co> accessed 2 July 2024.

actors, bear the responsibility of enforcing the law. Of particular importance here is the OSA's sanctions regime, which provides an incentive for regulated services to err on the side of over-removing potentially harmful (and/or illegal) content or not hosting it in the first place. This relates to longstanding concerns over displacing decisions regarding the line between legal and illegal speech from the public authorities to private actors absent safeguards such as judicial oversight or due process.[76]

Concerns regarding the likelihood of over-removal in the context of compulsory platform regulation are also well documented in the literature. By way of example, Daphne Keller, Director of Intermediary Liability at the Center for Internet and Society at Stanford Law School, warns of the dangers to the free speech rights of users when states delegate interpretation and enforcement of speech regulations to private actors. Drawing on data from studies concerning intermediary liability in the sphere of copyright law, Keller observes that '[t]wenty years of experience with these laws in the United States and elsewhere tells us that when platforms face legal risk for user speech, they routinely err on the side of caution and take it down'.[77] Additionally, in a 2024 report assessing social media content removal in France, Germany, and Sweden, the Future of Free Speech found that 87.5 to 99.7 percent of deleted comments on Facebook and YouTube in these countries were legally permissible, suggesting that platforms may be over-removing content to avoid regulatory penalties.[78] The report concluded that 'the over-removal of legal content on social media platforms raises concerns about the chilling effect on free expression and the potential suppression of legitimate discourse online'.[79]

Free speech advocates raised similar concerns during the legislative process for the OSA.[80] Compounding this concern is the increasing reliance of platforms on automated tools to identify problematic content – which the UN expressly cautions against in the context of adopting human rights based regulatory approaches – particularly in relation to the types of voices that may be excised from public discourse as a consequence.[81] In a 2021 report on social media content moderation, the Brennan Center for Justice

[76]See, e.g., Joelle Fiss and Jacob Mchangama, 'The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship' (*Justitia*, 2019) <https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2019/11/Analyse_The-Digital-Berlin-Wall-How-Germany-Accidentally-Created-a-Prototype-for-Global-Online-Censorship.pdf> accessed 28 June 2021; Daphne Keller, 'Internet Platforms: Observations on Speech, Danger, and Money' (2018) Hoover Institution's Aegis Paper Series No 1807 <https://ssrn.com/abstract=3262936> accessed 28 June 2021.

[77]ibid (Keller) 2.

[78]See 'Preventing "Torrents of Hate" or Stifling Free Expression Online?' (*The Future of Free Speech*, May 2024). The collected comments were analysed by legal experts to determine whether they were illegal based on the relevant laws in effect in each country.

[79]ibid.

[80]See. e.g., Joe Mullin, 'The UK Online Safety Bill Attacks Free Speech and Encryption' (*Electronic Frontier Foundation*, 5 August 2022) <https://www.eff.org/deeplinks/2022/08/uks-online-safety-bill-attacks-free-speech-and-encryption> accessed 3 August 2023.

examined the continuous use of automated systems on Facebook, Twitter (now X), and YouTube to identify and remove content that contravened community standards, highlighting how these systems put speech from marginalised communities at risk of over-removal.[82] This is an important point because it sheds light on how reliance on automated tools for content moderation creates a disproportionate risk of the over-removal of speech that is critical for healthy public discourse.

For example, the study examined how upload filters are typically applied to content related to nudity and suspected terrorism.[83] These filters, which largely rely on hashing systems that 'fingerprint' offending content in order remove duplicates quickly, may capture satirical and artistic expression and, in so doing, reflect a decision by the platforms to accept errors for the purpose of rapid enforcement.[84] The report also quotes from a Twitter employee working on machine learning issues, who shared that 'while the company viewed possibly restricting ordinary Arabic-speaking users and journalists as an acceptable trade-off in the fight against ISIS, deploying a similar tactic to fight white supremacy in a manner that would constrain American users and Republican politicians was not'.[85] Additionally, studies into the use of automated tools for identifying hate speech revealed that models for automatic hate speech detection were 1.5 times more likely to flag tweets written by self-identified Black people as offensive or hateful.[86]

The foregoing examples highlight the challenges and complexities of using automated tools to identify and remove particular types of content and how inherent biases in these tools may function to silence marginalised voices. Indeed, filters and other automated tools are not an appropriate substitute for human judgment, particularly in the context of legal determinations.[87] As Keller notes, '[t]o an algorithm, an album cover image used to promote illegal downloads is indistinguishable from the same image in a concert review. An ISIS video looks the same, whether used in recruiting or in news reporting'.[88] The primary concern with the OSA in this regard is that it combines a nebulous definition of harm with unprecedented

[81] See Robert Gorwa, Rueben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance (2020) 7 Big Data & Society 1.
[82] Ángel Díaz and Laura Hecht-Felella, 'Double Standards in Social Media Content Moderation' (*Brennan Center for Justice*, 4 August 2021) <https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation?ref=welcometohellworld.com> accessed 3 November 2023.
[83] ibid 11.
[84] ibid.
[85] ibid.
[86] See Maarten Sap et al., 'The Risk of Racial Bias in Hate Speech Detection' *Proceeding of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 28–August 2, 2019, 1671 <https://www.aclweb.org/anthology/P19-1163.pdf> accessed 10 June 2024.
[87] Keller (n 76) 7.
[88] ibid.

enforcement powers, raising legitimate concerns that the threat of non-compliance is likely to result in platforms *increasing* their reliance on such tools to err on the side of caution in removing – or failing to post in the first place – large swaths of legal political speech from marginalised communities. This risks further reducing the variety of voices on the platforms on which so much of contemporary public discourse takes place.

The flawed false communications offence

In addition to placing duties and responsibilities on regulated services, the OSA also introduces new communications offences, including the false communications offence. According to the government, this offence 'will bring internet trolls to justice' by outlawing the intentional sending of false information that could cause harm.[89] Additionally, the government claims the offence will 'bolster [its] strong commitment to clamping down on dangerous disinformation and election interference online'.[90] The Ofcom guidance on this offence suggests a less ambitious approach, advising that '[t]his offence is not intended to capture all "fake news"'.[91] This is more in line with the Law Commission's recommendations upon which the offence is based.[92]

A person commits this offence if they send a message that conveys knowingly false information with the intent of causing 'non-trivial psychological or physical harm' to the likely audience (without a reasonable excuse for sending).[93] A 'likely audience' includes any individual who, at the time the message is sent, is either reasonably foreseeable as someone who would encounter the message or a subsequent message that forwards or shares the content of the message.[94] In a case where several or many individuals are a likely audience, it is not necessary that the person intended to cause harm to any one of them in particular (or to all of them).[95] There is an exemption for recognised news publishers, which cannot commit this offence.[96] The CPS guidance on this offence notes that there is no requirement that 'such harm should in fact be caused, only that it be intended'.[97]

[89]'Cyberflashing, epilepsy-trolling and fake news to put online abusers behind bars from today' (*Department for Science, Innovation and Technology Press Release*, 31 January 2024) <https://www.gov.uk/government/news/cyberflashing-epilepsy-trolling-and-fake-news-to-put-online-abusers-behind-bars-from-today> accessed 14 February 2024.
[90]ibid.
[91]*Ofcom* (n 71) 139.
[92]See Law Commission, *Modernising Communications Offences A final report*, HC 547, Law Com 399 (2021).
[93]OSA (n 62) s179.
[94]ibid s179(2).
[95]ibid s179(3).
[96]ibid s180.
[97]'Communications Offences – Legal Guidance' (*CPS*) <https://www.cps.gov.uk/legal-guidance/communications-offences> accessed 1 February 2024.

It is also worth noting that, at present, we do not know what 'non-trivial psychological or physical harm' means. The Law Commission consultees noted that this type of harm is difficult to define, which raises further concerns regarding a lack of clarity in the framework that could, in practice, negatively impact free speech rights.[98]

The free speech concerns arising from this offence are apparent on its face and reflect a trend in the UK – and beyond – of incorporating increasingly broad speech related offences into its criminal law framework, the result of which is proscriptions on expression that bear no causal link to any risk of demonstrable harm.[99] Here, a person may commit the offence absent evidence that anyone was exposed to the impugned expression or that any harm was likely to result therefrom, let alone that any harm actually resulted. These types of offences raise serious concerns regarding coercive overreach in the realm of freedom of expression in the digital age.

Additionally, while Ofcom acknowledges that there are 'issues around freedom of expression and the difficulty for services in determining falsity' it has, to date, failed to go into detail concerning what these issues are and how regulated services should address them.[100] Instead, it advises that there will be instances when platforms may be able to infer knowledge of falsity and advises them to consider whether the message is actually false and whether the user intended to cause nontrivial (which is not defined in the legislation) physical or psychological harm.[101] In such cases, Ofcom advises that it is 'likely that the service will have reasonable grounds to infer that the content is illegal content' under the OSA.[102] As with the OSA's vague definition of 'harm', this offence suffers from a lack of clarity that risks regulated services increasingly relying on automated tools to err on the side of over-removal to avoid potential exposure to the sanctions regime.[103]

Finally, it is worth noting that this offence is grouped as a 'Foreign Interference' offence in Ofcom's guidance.[104] Conspicuously absent from the OSA and Ofcom's guidance to date is a focus on domestic disinformation, which research suggests is a significant problem in contemporary British politics.

[98]See Coe (n 65).
[99]See, e.g. Eliza Bechtold, 'Scotland's New Hate Crime Law Imperils Freedom of Expression' (2022) 26 Edin LR 250; Eliza Bechtold 'Terrorism, the Internet, and the Threat to Freedom of Expression: The Regulation of Digital Intermediaries in Europe and the United States' (2020) 12 Journal of Media Law 13–46; 'Human Rights Council Report on the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms high countering terrorism' (*UNGA*, 1 March 2019) UN Doc A/HRC/40/52.
[100]See *Ofcom* (n 71) 138.
[101]ibid.
[102]ibid 139.
[103]For a more detailed examination of the free speech issues surrounding the false communications offence, see Coe (n 65).
[104]See *Ofcom* (n 71) 144.

By way of example, a 2020 report compiled by the Oxford Internet Institute found that computational propaganda – defined as the use of automation, algorithms and big-data analytics to manipulate public life – is a widespread tactic amongst multiple actors in the British political system.[105] The study highlights the Coalition for Reform in Political Advertising's description of political advertising coming from the main parties during the 2019 General Election as 'illegal, indecent, dishonest and untruthful'.[106] It further notes that analysts reported on the 'apparent impunity with which the main parties … employed overt disinformation to secure votes'.[107] If the government is truly dedicated to making Britain safe from disinformation, perhaps it should turn its focus closer to home.

Failed efforts (so far) to regulate online harms in the US

At the state level in the US, legislatures are introducing a flurry of legislation directed to tackling online harms.[108] While the OSA attempts to regulate myriad online harms, US states are taking a more piecemeal approach. This section examines some of these legislative efforts and the free speech issues they raise, and analyses First Amendment challenges.

Legislation targeting online hate speech falters under strict scrutiny

In May of 2022, a white supremacist used a social media platform to live-stream himself perpetrating a hate fuelled mass shooting on Black shoppers at a grocery store in New York. Shortly thereafter, a recording of the shooting 'went viral' and a manifesto expressing the shooter's racist ideology was also shared on social media. These events spurred New York's legislature to enact the 'Hateful Conduct Law' in 2022.[109] The law applies to 'social media networks'[110] and prohibits the use of such networks to 'vilify, humiliate, or incite violence' on the basis of race, colour, religion and several other

[105]Samantha Bradshaw and others, 'Country Case Studies Industrialised Disinformation: 2020 Global Inventory of Organised Social Media Manipulation' (*Oxford Internet Institute*, 2020) <https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2021/03/Case-Studies_FINAL.pdf> accessed 3 March 2021, 415.
[106]ibid.
[107]ibid.
[108]In 2023, more than half of state legislatures targeted social media use by young people for regulation. See Jason Kelly and Aaron Mackey, 'States Attack Young People's Constitutional Right to Use Social Media: 2023 Year in Review' (*Electronic Frontier Foundation*, 30 December 2023) <https://www.eff.org/deeplinks/2023/12/states-attack-young-peoples-constitutional-right-use-social-media-2023-year-review> accessed 3 January 2024.
[109]NY Gen Bus Law § 394 – ccc(5).
[110]Defined as 'service providers, which, for profit-making purposes, operate internet platforms that are designed to enable users to share any content with other users or to make such content available to the public.' ibid. § 394 – ccc(1)(b).

characteristics.[111] It further requires 'social media networks' to create complaint mechanisms for users with respect to specific types of 'hateful conduct' and to have 'clear and concise' policies regarding how they will respond to reports of incidents of such conduct on their planforms.[112]

The law was immediately challenged on First Amendment grounds by operators of online platforms falling within its definition of a 'social media network'.[113] The Plaintiffs allege both as applied and facial First Amendment challenges.[114] First, they argue that the legislation violates the First Amendment as it applies to them because it is a content-based regulation on expression. Second, they argue that the legislation violates the First Amendment because it is facially overbroad or vague.[115] Plaintiffs sought a preliminary injunction to block the law going into effect during the pendency of the lawsuit. To prevail, the Plaintiffs bore the burden of demonstrating a likelihood of success on the merits.

With respect to the 'as applied' claim, the district court interpreted the law as compelling Plaintiffs to speak about 'hateful conduct', noting that this was a particularly onerous duty for Plaintiffs, whose websites have dedicated 'pro-free speech purpose[s]' that likely attract users who oppose censorship.[116] In the court's view, requiring Plaintiffs to endorse the state's definition of 'hateful conduct' forces them to weigh in on the debate about the contours of hate speech when they may otherwise choose not to speak. In so doing, the law deprives Plaintiffs of their First Amendment right to communicate freely on matters of public concern absent state coercion.[117] On this issue, the court further opined that private entities, including online platforms, have a First Amendment right to make decisions 'about whether, to what extent, and in what manner [to] disseminate speech … '.[118]

The court assessed the law as a content-based restriction on expression subject to strict scrutiny, affirming USSC precedent that 'government regulation of speech is content based if a law applies to particular speech because of the topic discussed or the idea or message expressed'[119] and 'when a state compels an individual to speak a particular message, the state alters the content of their speech, and engages in content-based regulation'.[120] While the court acknowledged that 'preventing and reducing the instances of

[111]ibid § 394 – ccc(1)(a).
[112]ibid § 394 – ccc(3).
[113]*Volokh v James*, 2023 WL 1991435 (SD NY 2023).
[114]In certain circumstances, US courts may, as an exception to ordinary standing requirements, entertain a claim that a law is unconstitutional even if potentially constitutional as applied to the claimant, i.e. that the law is facially unconstitutional. See *Hobbs v County of Westchester*, 397 F3d 133 (2nd Cir 2005).
[115]These are examples of facial challenges.
[116]*Volokh* (n 113) *6.
[117]ibid.
[118]ibid (internal citations and quotations omitted).
[119]ibid *4 (quoting *Reed v Town of Gilbert, Ariz*, 576 US 155, 163 (2015)).
[120]ibid (quoting *Nat'l Inst of Fam & Life Advocs v Becerra*, 138 SCt 2361, 2371 (2018)).

hate-fueled mass shootings' is a compelling government interest, it held that the law was not narrowly tailored toward that end. For example, it was 'unclear what, if any, effect a mechanism that allows users to report hateful conduct on social media networks would have on reducing mass shootings'.[121] For these reasons, among others, the court found that Plaintiffs had demonstrated a substantial likelihood of success on their as applied First Amendment challenges.

With respect to Plaintiff's facial challenges, the court found that the law 'clearly implicates the protected speech of social media users'.[122] It observed that while the law does not require social media networks to remove 'hateful conduct' from their websites and does not impose liability on users for engaging in such conduct, targeting this type of speech 'certainly could make social media users wary about the types of speech they feel free to engage in without facing consequences from the state'.[123] The court also asserted that this potential chilling effect is exacerbated by the indefiniteness of the law's key terms, querying whether a post using the hashtag 'BlackLivesMatter' or 'BlueLivesMatter' could be considered 'hateful conduct'. Due to this uncertainty, the legislation 'fails to put social media users on notice of what kinds of speech or content is now the target of government regulation'.[124] For all of these reasons, the court held that Plaintiffs had met their burden of demonstrating a likelihood of success on the merits of their facial challenges.[125] At the time of writing, the law is subject to a preliminary injunction and the case is pending in the Southern District of New York.

In the current landscape of online harms legislation, the Hateful Conduct Law is an illustrative example of the challenges in regulating categories of harmful speech in the US, which face the most demanding level of review by courts. Even in circumstances where legislation is clearly aimed at a compelling government objective, the means to effectuate that objective will be rigorously scrutinised and will likely fail under such scrutiny. This raises a reasonable question as to whether the USSC has set such a high burden for the state to overcome in addressing online harms that any regulation will ultimately fail as a result of concerns over a potentially chilling effect on protected speech. This approach offers a stark contrast with the UK framework, which offers numerous examples of broadly articulated speech offences that threaten to capture speech that falls well outside of the government's stated aims and, as a result, raises concerns regarding the adequate protection of free speech. The OSA offers two such examples, with its

[121]ibid *8.
[122]ibid *9.
[123]ibid *10.
[124]ibid.
[125]ibid (internal quotations and citations omitted).

nebulous definition of harm and the broadly articulated false communications offence.

Must carry laws lacking legitimate aims

While New York's Hateful Conduct Law and the OSA offer examples of legislation aimed at reducing the amount of harmful content on online platforms, recent legislation out of Texas and Florida is directed to the opposite result by way of must carry provisions that prohibit, to a large extent, online platforms from removing such content. These laws are premised on the idea that social media platforms censor conservative voices and viewpoints.[126] While this is a popular narrative promulgated by the Republican Party, it is a false one. A study released in 2021 by the NYU Stern Center for Business and Human Rights revealed that the claim that social media platforms suppress conservative voices 'is itself a form of disinformation: a falsehood with no reliable evidence to support it' and that no trustworthy large-scale studies have determined that conservative content is being removed for ideological reasons or that searches are being manipulated to favour liberal interests.[127]

The Texas law prohibits social media platforms from 'censor[ing] a user, a user's expression, or a user's ability to receive the expression of another person based on the viewpoint of the user or another person'.[128] Florida's law, which applies to large social media platforms, prohibits banning any candidate for office and prohibits, among other things, post-prioritization or shadow banning algorithms for content posted by or about a user who is a candidate for office.[129] Removal of any material posted by a 'journalistic enterprise' based on content is also prohibited.[130]

Unsurprisingly, the Texas and Florida laws were immediately challenged by two trade associations that represent internet and social-media companies. The lawsuits (collectively referred to as 'the Netchoice cases' after the leading Plaintiff) allege, among other things, that the laws violate the companies' First Amendment rights and are pre-empted by Section 230 of the CDA.[131] The Netchoice cases raise an important set of questions about the role of the First Amendment in the context of regulating online speech, including whose free speech rights are implicated by such measures (the

[126]For example, a Texas State Senator tweeted that the objective of the legislation was to 'allow Texans to participate on the virtual public square free from Silicon Valley censorship'. @SenBryanHughes, 5 March 2021.
[127]Paul M Barrett and J Grant Sims, 'False Accusation: The Unfounded Claim that Social Media Companies Censor Conservatives' (*NYU Stern Center for Business and Human Rights*, February 2021) 1.
[128]Tex Civ Prac & Rem Code § 143A.002; § 143A.002(a)(1)–(3).
[129]See Fla Stat § 106.011(3)(e); Fla Stat § 501.2041(2)(h).
[130]ibid. § 501.2041(2)(j). 'Journalistic enterprise' is defined broadly to include any entity doing business in Florida that publishes in excess of 100,000 words online and has at least 50,000 paid subscribers or 100,000 monthly users.
[131]See *Netchoice, LLC v Paxton*, 49 F4th 439 (5th Cir 2022) and *Netchoice* (n 59), respectively.

private entities that own the platforms, the users, or both), and the legitimacy of a state's purported interest in regulating online expression. For example, Texas and Florida argue that social media platforms are 'common carriers' that, like phone companies, merely serve as a conduit for the speech of users. This is a key issue because the USSC holds that government regulation of common carriers does not raise significant First Amendment concerns.[132] Such designations are particularly important because they inform the level of scrutiny applied by courts. Less scrutiny is applied to legislation regulating common carriers than to individuals and entities engaging in First Amendment activity.[133]

In both cases, Plaintiffs sought preliminary injunctions, which resulted in a split among the federal circuit courts of appeal regarding shared First Amendment questions, including whether online platforms have First Amendment rights with respect to content moderation. The USSC granted certiorari in both cases and issued a single decision in June of 2024.[134] While the majority opinion offers some interesting dicta, there is no useful precedent concerning the important First Amendment questions at issue because the Court vacated remanded the cases back to the lower courts on procedural grounds.[135] As a result, these questions remain unanswered. While the parties in these cases offer diametrically opposed interpretations of the appropriate application of First Amendment precedent to online platforms, the Knight First Amendment Institute at Columbia University and other legal advocacy organisations in the US argue that the more reasonable approach likely lies somewhere in the middle.[136] Where to draw the line remains an open question and the subject of intense debate in the US.

The *Netchoice* cases also highlight why it is important to interrogate the stated objectives of the government in regulations impacting online speech. Here, the aims of the government are spurious, and the practical effect of the regulations is that platforms will be forced to host speech that, while legal in the US, most reasonable people would agree should not be available online, e.g., hate speech and disinformation.[137] Instead of reducing

[132]See *Denver Area Educational Telecommunications Consortium, Inc v FCC*, 518 US 727 (1996).

[133]For more detailed analysis of the Texas and Florida laws, see David Cole, 'Who Should Regulate Online Speech?' (*NY Times Review of Books*, 23 February 2024) <https://documentcloud.adobe.com/spodintegration/index.html?locale=en-us> accessed 28 February 2024.

[134]*Netchoice* (n 60).

[135]The majority opinion, authored by Justice Kagan, held that the lower courts did not properly consider Netchoice's facial invalidity claims.

[136]See, e.g. 'Knight Institute Urgers Supreme Court to Reject "Extreme" Arguments Made by States and Platforms in Cases Involving State Social Media Laws (*Knight First Amendment at Columbia University Press Statement*, 23 February 2024 <https://knightcolumbia.org/content/knight-institute-urges-supreme-court-to-reject-extreme-arguments-made-by-states-and-platforms-in-cases-involving-state-social-media-laws> accessed 1 March 2024.

[137]As discussed above, in the US, the question of whether such decisions are undertaken by private actors or the state is of primary importance as only the state is constrained by the First Amendment. See *Edmonson* (n 46) 619.

online harms, these laws would likely exacerbate them.[138] A federal court in the Florida litigation highlighted this issue by expressing concern with provisions that would prohibit child-friendly websites like YouTube Kids from removing soft core pornography posted by PornHub, which falls under the legislation's definition of a 'journalist enterprise'.[139]

Efforts to reduce online harms to children by way of age-verification requirements

Ofcom will be the age verification regulator in the UK under the new regulatory framework. The OSA requires a covered service to use age verification[140] or age estimation[141] (or both) to prevent children of any age from encountering 'primary priority content that is harmful to children' that it identifies on the service.[142] This category of content includes pornographic content and content that encourages suicide or an eating disorder.[143] In a 2023 Policy Paper, the UK Government argued that the OSA's age verification requirement in relation to pornographic content is related to the legitimate aim of 'protect[ing] children from suffering harm resulting from exposure to pornography during childhood'.[144] It contends that the age verification requirement 'interferes with the Article 10 rights of users no more than is necessary' because it only prevents children from receiving regulated content and 'does not prescribe how the age verification step must be done'.[145] This leaves regulated services with 'flexibility to introduce age verification in a way which is least burdensome for their adult users (as long as it still restricts children's access)'.[146]

In its statutorily mandated guidance, Ofcom acknowledges that technology around age assurance is still developing and, as a result, does not recommend any specific tool or technology.[147] However, it proposes criteria that a service provider should use when considering age assurance

[138]See Alex Chemerinsky and Erwin Chemerinsky, 'Misguided Federalism: State Regulation of the Internet and Social Media (2023) 102 NC L Rev 1.
[139]See *Netchoice* (n 59) 1229.
[140]Defined as 'any measure designed to verify the exact age of users of a regulated service'. OSA (n 62) s 230(2).
[141]Defined as 'any measure designed to estimate the age or age-range of users of a regulated service'. ibid s 230(3).
[142]ibid s12(4).
[143]ibid 61(1)–(5).
[144]'Online Safety Bill: European Convention on Human Rights Memorandum' (*Department for Digital, Culture, Media & Sport*, January 2023) <https://www.gov.uk/government/publications/online-safety-bill-supporting-documents/online-safety-bill-european-convention-on-human-rights-memorandum> accessed 1 June 2024.
[145]ibid.
[146]ibid.
[147]'Implementing the Online Safety Act: Protecting children online pornography' (*Ofcom*, 5 December 2023) <https://www.ofcom.org.uk/news-centre/2023/implementing-the-online-safety-act-protecting-children> accessed 15 January 2024.

methods and processes, including accuracy and reliability.[148] One problem with this guidance is that it fails to acknowledge or address the inherent free speech problems with age verification and assessment requirements. The US context offers useful lessons for the UK with respect to the free speech challenges in this area.[149] These laws, like the hate speech and must carry laws examined above, are facing legal challenges. Most recently, courts in Arkansas, California, and Texas have issued preliminary injunctions barring such laws from going into effect.

Of particular relevance here is the Texas law – House Bill 1181 – which is most analogous to the OSA. This law, successfully challenged in court, restricts access to websites deemed to be at least 'one-third' comprised of 'sexual material harmful to minors' by requiring digital age verification methods.[150] The legislation defines '[s]exual material harmful to minors' as any material the average person applying contemporary community standards would find is designed to appeal or pander to the prurient interest to minors, is patently offensive to minors, and lacks serious literary, artistic, political, or scientific value for minors.[151] It requires sites to verify a visitor's age via 'a commercial age verification system … using: (A) government-issued identification; or (B) a commercially reasonable method that relies on public or private transactional data to verify the age of an individual'.[152]

In its First Amendment analysis of the law's age verification requirement, the district court applied strict scrutiny because it 'restricts access to speech based on the material's content'.[153] With respect to whether the law is directed to a legitimate aim, while the court agreed that Texas has a legitimate goal in protecting children from sexually explicit material online, it emphasised the burden on Texas to demonstrate that the law survives strict scrutiny.[154] Identifying several problems with the law, the court found that the law was neither narrowly tailored nor the least restrictive means to achieving the state's aim. Among other things, the court held that the law is underinclusive because it does not regulate search engines, ignoring Defendant's own expert's suggestion that exposure to online pornography often begins with 'misspelled searches'.[155] Thus, while the law regulates adult video companies that post sexual material to their sites, it 'will do little else to prevent children from accessing pornography'.[156]

[148]ibid.
[149]See 'Social Media and Children 2023 Legislation' (*National Conference of State Legislatures*, 26 January 2024) <https://www.ncsl.org/technology-and-communication/social-media-and-children-2023-legislation> accessed 28 February 2024.
[150]HB 1181 (n 1).
[151]ibid § 129B.002.
[152]ibid § 129B.003.
[153]*Free Speech Coalition v Colmenero*, 2023 WL 5655712, *8 (WD Texas 2023).
[154]ibid *29.
[155]ibid *10.
[156]ibid.

Another fatal flaw in the law, for the court, is that it 'deters adults' access to legal sexually explicit material, far beyond the interest in protecting minors'.[157] Here, the court noted the USSC's disapproval of 'content-based restrictions that require recipients to identify themselves affirmatively before being granted access to disfavored legal speech' on the basis that this chills adults from accessing such speech.[158] Another problem with the law is that Texas did not show that it undertook any analysis regarding the difference between age verification and content filtering, despite USSC precedent favouring the latter.[159] The court opined that because it was 'clear that age verification is considerably more intrusive while less effective than other alternatives', the law could not survive strict scrutiny.[160] For example, content-filtering – such as adult controls on children's devices – 'is the modern version of the "blocking and filtering"' software that the USSC has proposed as a preferred alternative to age verification methods.[161] For all of these reasons, and many more, the court held that Plaintiffs were likely to prevail on their facial claims that the law violates the First Amendment rights of internet users.

Unlike the must carry laws in Texas and Florida, laws directed to limiting minors access to pornographic and other sexually explicit material are directed to a legitimate government aim. However, how states are attempting to effectuate this aim are meeting strong First Amendment resistance in the courts. As the Texas case demonstrates, one of the main free speech problems with age verification requirements is that they are not the least restrictive means of protecting children from online harm because they ignore more potentially effective alternatives, like those identified by the USSC. Additionally, they are not narrowly tailored, which risks chilling adults from accessing constitutionally protected speech.

These shortcomings are also present in the OSA, which requires age verification for specific types of content that is legal for adults to access, ignoring potentially more effective – and less intrusive – alternatives that do not risk chilling adults from accessing legal content. While the USSC's interpretation of the First Amendment applies more expansive protections to expression as well as a more rigorous standard of review of regulatory efforts directed to restricting access to harmful content than the ECtHR with respect to Article 10, in both the jurisdictions limitations on fundamental rights must be *proportionate* to legitimate government interests. As discussed above, the ECtHR applies a proportionality test in determining whether an act of a public authority violates Article 10. Part of the ECtHR's analysis

[157]ibid *15
[158]ibid (citing *Ashcroft v ACLU*, 535 US 564 (2002)).
[159]ibid *19.
[160]ibid *20.
[161]ibid (citing *Ashcroft* (n 158)).

involves a balancing test that determines whether an interference is proportionate to the legitimate aim pursued by the regulation. If the answer to that question is yes, then the interference is deemed necessary in a democratic society.

To date, the ECtHR has not ruled on the precise question of whether age verification requirements violate Article 10, and any such decision would be based on a fact specific inquiry of the measures at issue in a given case. However, in its 2018 'Guidelines to respect, protect, and fulfil the rights of the children in the digital environment', the Council of Europe advised that Member States should 'require use of effective systems of age-verification to ensure children are protected from products, services and content in the digital environment which are legally restricted with reference to specific ages, using methods that are consistent with the principles of data minimisation'.[162] In his Concurring Opinion in the 2019 case of *Pryanishnikov v Russia*, Judge Pinto De Albuquerque referenced the Guidelines in arguing there is a 'positive obligation' on Member States to prevent 'under-18s from accessing pornographic material and content through mandatory age verification.[163] While Judge Pinto De Albuquerque acknowledged that 'there are tools that would allow for circumventing the age-verification restrictions' this 'is no excuse and [t]he fact that the law can be broken by some does not justify its not being imposed on the many, otherwise no prohibitive law would ever be adopted'.[164]

To date, neither the Council of Europe nor the ECtHR have addressed the significant deficiencies in age verification tools and technology or the human rights concerns that arise in this context. Observations regarding proportionality in this context and lessons available from legal challenges to such regulations in the US are discussed below.

Some preliminary observations

The discourse concerning the free speech implications of regulating online harms is constantly evolving due to changes in regulatory frameworks, doctrinal developments, as well as new developments in technology and the policies and practices of online platforms. Accordingly, this section offers some preliminary observations regarding the potential implications for freedom of expression from an examination of recent regulatory efforts in the UK and the US with the aim of offering a contribution to the broader, evolving discourse in this area.

[162]'Recommendation CM/rec(2018)7 of the Committee of Ministers to member States on a Guidelines to respect, protect and fulfil the rights of the child in the digital environment' CM/Rec(2018)7 (4 July 2018)), 10.
[163]App No 25047/05 (10 September 2019), 26.
[164]ibid 28.

The OSA and similar regulatory efforts in the US contravene UN guidance directed to ensuring adequate protection for human rights

An examination of the OSA and recent regulatory efforts by US states reveal surprisingly similar approaches to tackling the challenges posed by harmful online speech given the much stronger protections for free speech in the American framework. Another notable similarity between the OSA and state laws is the extent to which they contravene UN guidance on platform regulation in important ways.[165] For example, the OSA and New York's Hateful Conduct Law offer examples of poor definitions of key terms. In the OSA, examples include the nebulous definition of harm and the absence of definitions for key terms in the false communication offence, including 'nontrivial'. New York's Hateful Conduct Law lacks precision with respect to key terms like 'vilify' and 'humiliate'.

While such poor definitions create significant constitutional problems in the US, where such restrictions are subject to strict scrutiny by courts, they do not present as clear a challenge in the UK due to the ECtHR's tolerance of broadly articulated content and viewpoint-based speech offences.[166] However, while the OSA's definition of harm may not violate Article 10, it fails to comply with UN guidance on adopting human based regulatory approaches, which calls for more specificity with respect to key terms.[167] Relatedly, both the OSA and state laws in the US incentivise platforms to rely more heavily on automated tools to ensure compliance with overly broad and/or vague regulations. The OSA, in particular, due to its expansive scope, arguably poses a risk of over-removal of the types of speech that are necessary for healthy public discourse, including voices from marginalised communities and unpopular political speech, by incentivising platforms to increase reliance on automated decision marking. In so doing, it also increases the likelihood that legal speech that is available offline will not be available online. A human rights based approach to platform regulation in line with UN Guidance would provide more precision in terms of defining key terms as well as carefully considering the implications of how complying with such regulatory requirements may increase reliance on automated tools – and how such reliance may impact public discourse and put at risk the free speech rights of internet users.

While protecting children from online harm is a legitimate aim, regulatory efforts targeting this type of content may fail to adequately protect the free speech rights of internet users

As discussed above, human rights defenders in the US are challenging state laws requiring age verification or assurance technology to access certain

[165]The exception to this is must carry laws that, as discussed above, create a different set of concerns.
[166]See Natalie Alkiviadou and Jacob Mchangama, 'Hate Speech and the European Court of Human Rights: Whatever Happened to the Right to Offend, Shock, or Disturb?' (2021) 21 HRL Rev 4.
[167]See UN (n 4).

types of content on the basis that they imperil the free speech and privacy rights of internet users. The arguments against these types of requirements offer lessons for the UK as well as other jurisdictions enacting similar regulations.

In an Amici Curiae brief in the Texas age-verification case, the American Civil Liberties Union (ACLU) highlights the primary free speech objections to age verification requirements.[168] These include the burdens imposed on all internet users associated with verification methods when there are less restrictive alternatives that would likely be at least as effective in achieving the legitimate purpose of protecting children from harmful content.[169] Specifically, the ACLU points to policies enabling users (or their parents) to control their own access to information through user-installed devices and filters or affirmative requests to companies, which the USSC recognises as more effective than age verification tools.[170] The ACLU also argues that age verification regulations rob internet users of the anonymity otherwise available on the internet with respect to protected speech activities, encroaching on the personal lives of those who use the internet precisely due to the anonymity it provides.[171] As recognised by the Texas court, this may result in a chilling effect on adults accessing legal content online.[172]

Additionally, it is reasonable to question whether age verification is a particularly useful method for ensuring that children are blocked from accessing particular content, as they 'can easily switch between most physical devices or cards, and automated analysis of facial features or online presence relies on a range of potentially biased or inaccurate assumptions'.[173] It is also worth noting here that internet users are legitimately concerned with the privacy implications of current age identification techniques, including credit cards, government-issued identification, analysis of online search history, and facial recognition. These methods require companies to process additional information from all users, thereby increasing the risks of cybersecurity attacks.[174]

Given existing USSC precedent holding that age verification requirements are not narrowly tailored or the least restrictive means of protecting children from harmful sexual content and impermissibility interfere with

[168]Brief of *Amici Curiae* American Civil Liberties Union and others, Case No 23-50627, Doc 85, 8 (26 September 2023). The ACLU was joined by other human rights advocates, including the Centre for Democracy and Technology.
[169]ibid 8.
[170]ibid 9 (quoting *Ashcroft* (n 158) 657.
[171]ibid 10.
[172]See *Colmenero* (n 153).
[173]Caitlin Chin-Rothman and Taylor Rajic 'A New Chapter in Content Regulation: Unpacking the UK Online Safety Bill (*Center for Strategic and International Studies*, 18 October 2023) <https://www.csis.org/analysis/new-chapter-content-moderation-unpacking-uk-online-safety-bill> accessed 12 February 2024.
[174]See ibid.

constitutionally protected speech, it is unlikely that state (or federal) laws requiring such methods will survive legal scrutiny by US courts. While the OSA's age verification requirement is tied to a legitimate government interest – protecting children from online harm – it suffers from similar deficiencies as recent state laws.

While the Council of Europe and at least one ECtHR judge have advised that age verification requirements do not pose Article 10 problems, the adjudication of age verification laws by US courts suggests that, at present, these regulations are not proportionate to the legitimate aim of protecting children from online harm. As a result, it is reasonable to inquire as to whether such measures are 'necessary in a democratic society'. To engage in such an inquiry is not to suggest that there is no place for government regulation in this area, that the UK should adopt the American approach, or that the public should throw up its hands in collective resignation of the inevitability of children freely accessing pornographic and other harmful content online. Rather, it highlights the importance of the government's burden in interfering with human rights to establish that the interference is proportionate to the legitimate aim pursued which, in this context, is mitigating harm to children from particular types of online content. While the scrutiny applied to such measures is stricter in the American framework, ultimately, the question in both jurisdictions is whether the means used to effectuate a government objective are proportionate. Answering this question requires an analysis of the efficacy of such means in relation to the aim and consideration of alternatives that do not pose as substantial a threat to human rights. In this respect, the US offers instructive lessons and guidance for the UK.

A hyper-focus on regulations restricting access to content distracts from meaningful consideration of regulatory options and approaches that do not imperil free speech

The OSA reflects a regulatory approach that focusses on placing duties of care on online platforms to protect users from myriad types of online harms and imposing potential criminal liability if they fail to do so. Such an approach reflects a belief that it is primarily the responsibility of platforms to prevent online harms to victims.[175] What is lacking is in this area is a sustained focus on other avenues for reducing online harms that do not directly target content moderation. These include improved education for children and young people and training for teachers, police and prosecutors in the

[175]See Andy Phippen and Emma Bond, 'Why do legislators keep falling victims in online harms' (2024) 38 International Review of Law, Computers & Technology 195–214, 206.

context of mitigating risks to these populations from harmful online content.[176] Such alternative approaches to ameliorating the harms to children flowing from exposure to pornographic content were absent from government debates over the Online Safety Bill.[177] Relatedly, in opposing the introduction of the false communications offence, the Electronic Frontier Foundation argued that legitimate concerns about widespread misinformation and conspiracy theories would be better countered by promoting high-quality information and digital media literacy as well as diverse, independent media sources to ensure that online users hear a plurality of political or scientific views.[178]

Also missing from regulatory efforts in this area is a focus on the business model of social media companies. While the business model of traditional media can lead to significant polarisation, limited bandwidth and editorial oversight generally incentivise attempting to reach broader markets, which disincentives publishing extreme content.[179] This is not the case for social media, which relies on leveraging individual users' data to push highly personalised content in order to maximise engagement, which translates into profit, thus incentivising more customised and potentially more extremist content.[180] Instead of focusing on these issues, UK and US state regulatory efforts largely prioritise restricting access to content, which, as discussed throughout this article, raises significant free speech concerns.[181] The point here is that a preoccupation with regulating content moderation – which results in either taking down content or restricting access altogether – stymies the discourse around how to address the problem of online harms while drawing attention away from other tools and approaches that do not so clearly and directly implicate free speech.

Conclusion

While recent efforts to regulate online harms at the state level in the US are facing intense scrutiny from courts, and largely failing as a result, the OSA

[176]ibid.
[177]ibid 207.
[178]See Corynne McSherry and others, 'Privacy First: A Better Way to Address Online Harms (*Electronic Frontier Foundation*, 14 November 2023) <https://www.eff.org/wp/privacy-first-better-way-address-online-harms> accessed 1 December 2023.
[179]Dipayan Ghosh, 'Are we Entering a New Era of Social Media Regulation?' (2021) Harvard Business Review <https://documentcloud.adobe.com/spodintegration/index.html?locale=en-us> accessed 12 September 2023.
[180]ibid. ARTICLE 19 highlighted during the legislative process that the OSA focusses on censorship instead of addressing the problematic business models of large online platforms. 'UK: Online Safety Bill is a serious threat to human rights online' (*ARTICLE 19*, 25 April 2022) <https://www.article19.org/resources/uk-online-safety-bill-serious-threat-to-human-rights-online/> accessed 3 September 2023.
[181]While the OSA places some transparency obligations on regulated services, what those obligations will be and the efficacy thereof remain to be seen.

will likely not encounter such scrutiny due to the ECtHR's more permissive approach to regulations on expression, both online and offline. However, neither jurisdiction is prioritising a human rights based approach to the regulation of online harms in line with UN guidance. Instead, both the OSA and US state laws suffer from imprecise definitions and a disproportionate focus on placing onerous obligations on platforms to remove or not host particular types of speech. In so doing, these laws incentivise platforms to rely more heavily on artificial intelligence and automated systems, potentially at the expense of the free speech rights of users.

These criticisms of the current regulatory landscape of online harms in the UK and the US do not suggest that the status quo is sufficient. Indeed, self-regulation brings its own host of problems and challenges. Elon Musk's purchase and subsequent control of X is a particularly troubling example of how concentrated private control over platforms creates problems for public discourse and exacerbates the harms arising from disinformation and other types of harmful speech. With that said, government efforts to tackle online harms, particularly those that place onerous requirements on platforms to police user content and adopt sanctions regimes, raise significant human rights issues that warrant our collective attention.

Regulatory approaches directed to addressing these challenges should be grounded in human rights, guided by the UN, so as to protect the important rights implicated by the regulation in this area, including free speech and privacy. At present, neither the UK nor US states are striking the right balance. While American courts are reigning in these laws based on the expansive protection of free speech under the First Amendment, the constitutional restraints in the UK are weaker, with the ECtHR tolerating content-based restrictions on expression and interpreting Article 10 as placing duties and responsibilities on internet users and platforms. This raises concerns regarding the adequate protection of the free speech rights of internet users in the UK. While the US approach brings its own set of challenges in relation to the predominance of free speech over other interests and principles,[182] in the context of the regulation of content moderation of online platforms, it is more in line with UN guidance.

Acknowledgements

The author would like to thank the anonymous reviewers and Dr Hedvig Schmidt for their invaluable feedback on previous drafts of this article.

[182]See, e.g. *Murthy v Missouri,* 2024 WL 3165801 (2024). While the USSC vacated and remanded this case back to the lower court on procedural grounds, it provides a useful summary of the significant First Amendment hurdles imposed on government efforts to coordinate with social media platforms (in the absence of statutory requirements) to address harms flowing from disinformation and other forms of harmful speech.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Eliza Bechtold is a Lecturer, School of Law, University of Aberdeen. Former attorney at the American Civil Liberties Union of New Mexico and DLA Piper LLP (US).