Journal of Media Law

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjml20

The Draft Online Safety Bill and the regulation of hate speech: have we opened Pandora's box?

Peter Coe

To cite this article: Peter Coe (2022) The Draft Online Safety Bill and the regulation of hate speech: have we opened Pandora's box?, Journal of Media Law, 14:1, 50-75, DOI: 10.1080/17577632.2022.2083870

To link to this article: https://doi.org/10.1080/17577632.2022.2083870

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 01 Jun 2022.

Submit your article to this journal

Article views: 6150

View related articles

View Crossmark data

Citing articles: 1 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
2022, VOL. 14, NO. 1, 50–75
https://doi.org/10.1080/17577632.2022.2083870

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

The Draft Online Safety Bill and the regulation of hate speech: have we opened Pandora's box?

Peter Coe

School of Law, University of Reading, Reading, UK

ABSTRACT
In thinking about the developing online harms regime (in the UK and elsewhere[1]) it is forgivable to think only of how laws placing responsibility on social media platforms to prevent hate speech may benefit society. Yet these laws could have insidious implications for free speech. By drawing on Germany's Network Enforcement Act I investigate whether the increased prospect of liability, and the fines that may result from breaching the duty of care in the UK's Online Safety Act - once it is in force - could result in platforms censoring *more* speech, but not necessarily *hate* speech, and using the imposed 'responsibility' as an *excuse* to censor speech that does not conform to their objectives. Thus, in drafting a Bill to protect the public from hate speech we may unintentionally open Pandora's Box by giving platforms a statutory justification to take more 'control of the message'.

KEYWORDS Online Safety Bill; online harms; hate speech; free speech; online speech

Introduction

According to the UK government's response to the Online Harms White Paper consultation[2] 'online harms' encapsulates a broad variety of content published online that can be harmful to certain groups within society, and/or society and the public sphere at large. The 'harms' identified in the response include hate speech.[3] Although not easy to define, hate speech is classified by various commentators as abusive speech that targets members

CONTACT Peter Coe peter.coe@reading.ac.uk

[1]For example, the European Commission's ongoing development of the Digital Services Act, Ireland's Online Safety Media Regulation, Germany's 'network enforcement law' known as *Netzwerkdurchsetzungsgesetz* law, or 'NetzDG', the European Union's *Code of conduct on countering illegal hate speech online*.

[2]Department for Digital, Culture, Media and Sport, *Online Harms White Paper: Full government response to the consultation* (CP 354, 15 December 2020).

[3]According to the response a 'limited number of priority categories of harmful content, posing the greatest risk to users, will be set out in secondary legislation' which will include 'hate crime'. Furthermore, 'hate content' is one of the 'priority categories' that will be set out by the government in secondary legislation. See ibid, paras 2.3 and 2.29.

© 2022 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 License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

of certain groups, which are, typically, minority groups.[4] The increasing role played by online platforms, such as Facebook, Twitter and Instagram, in our daily lives, and the extent of their responsibility for preventing the publication of hate speech on their platforms, has been a source of controversy for some time; a debate intensified in the UK following the publication of the Draft Online Safety Bill (the Bill) in May 2021.

When we think about this developing online harms regime (both in the UK and elsewhere), we can be forgiven for thinking only in terms of how laws placing responsibility on social media platforms to prevent hate speech and other harmful speech may benefit society and the public sphere. Yet, despite this seemingly obvious and welcome benefit, the Bill, like legislation introduced in other jurisdictions, has met with resistance from a variety of actors due to its potential negative impact on freedom of expression.[5]

It is important to note at this juncture that at the time of writing the overall shape of the new UK regime remains to be seen. Because this is a Draft Bill, it is likely to change. Additionally, because the Bill itself is rather vague, much of the legalistic detail is uncertain and undefined. As a result, future secondary legislation will follow the Bill's enactment. Therefore, debates about the legislation and its effectiveness will continue to rumble on, and only time will tell what its ultimate impact on free speech will be.[6] Notwithstanding this uncertainty, the purpose of this article is to ask whether, based on the current content of the Bill, in passing a law to protect the public sphere from online hate speech (and other illegal and legal, yet 'harmful' speech), we may unintentionally open Pandora's Box by giving social media platforms the opportunity, or even a de facto justification, to take more 'control of the message'. By asking this question in the context of the Bill, my hope is that this article will provide guidance, or perhaps even forewarning, for UK legislators, as well as legislators in other jurisdictions who are looking to develop, or are in the process of developing, similar regimes.

[4]For example, according to Alon Harel, these groups include 'racial groups, ethnic groups, religious groups, groups defined on the basis of sexual orientation and so on [and] other groups such as individuals targets on the basis of class', see A. Harel, 'Hate Speech' in A Stone and F Schauer, *The Oxford Handbook of Freedom of Speech* (Oxford University Press 2021), 455-476, 455. This corresponds with the UK government's classification of 'hate crime', which occurs where hostility is demonstrated towards an individual on the grounds of their actual or perceived race, religion, sexual orientation, disability or transgender identity, see HM Government, *Online Harms White Paper* (CP 354, April 2019), para 7.16.

[5]For example, see House of Lords Communications and Digital Committee, 'Free for all? Freedom of expression in the digital age', 1st Report of Session 2021-22, HL Paper 54, 22 July 2021; M Earp, 'UK online safety bill raises censorship concerns and questions on future of encryption', *Committee to Protect Journalists*, 25 May 2021; L Kirkconnell-Kawana, 'Online Safety Bill: Five thoughts on its impact on journalism' *Media@LSE*, 3 June 2021; C Elsom, 'Safety without Censorship. A better way to tackle online harms' *Centre for Policy Studies*, September 2020; J Petley, 'Online Safety and the Press: A Thoroughly Unsafe Bill, Parts 1 to 3', *Inforrm's Blog*, 6, 7 and 8 July 2021, <https://inforrm.org/>. See also the body of work from Graham Smith on *Inforrm' Blog* relating to the Bill.

[6]P Coe, 'Misinformation, Disinformation, the Online Safety Bill, and its Insidious Implications for Free Speech' (2021) *Communications Law* 26(3) 127-129, 129.

To answer this question, I draw on Germany's Network Enforcement Act, known as *Netzwerkdurchsetzungsgesetz* law, or 'NetzDG', which foreshadowed the current regulatory zeitgeist sweeping Europe and beyond,[7] and has much in common with the UK Bill, both in respect of why it was introduced and its operation. In constructing the Bill, the UK government took inspiration from its 'international partners' and regimes that are already in place, or in development, in those jurisdictions, including Germany and NetzDG.[8] Furthermore, and importantly, prior to and soon after its enactment NetzDG was the subject of concern over its potential impact on free speech. However, unlike the abstract fears surrounding the Bill, NetzDG has been in force since 2017.[9] Thus, its impact on free speech can, to an extent, be more accurately gauged. For this reason, it provides a tangible comparator for exploring whether the Bill, and in particular the duties it places on social media platforms, could lead to the feared insidious implications for free speech that have been advanced.

This article begins with an explanation of the current regulatory zeitgeist sweeping Europe and beyond, the impetus for the Bill, and how it mirrors the introduction of NetzDG. Next, the article explains how the Bill may operate once it is enacted, its potential implications for free speech, and how this compares to NetzDG. Finally, it examines the German experience following the passage of NetzDG.

[7]See: European Commission, 'Recommendation on Measures to Effectively Tackle Illegal Content Online' COM 2018 1177 final, 1 March 2018; The European Commission's ongoing development of the Digital Services Act; Ireland's Online Safety Media Regulation Bill; In Canada, in June 2021 the government announced a new Bill to amend the Canadian Human Rights Act, the Criminal Code and the Youth Criminal Justice Act to better tackle hate speech and hate crimes. The Bill 'will be complemented by a regulatory framework to tackle harmful content online': Department of Justice Canada, 'Government of Canada takes action to protect Canadians against hate speech and hate crimes' (online, 23 June 2021) <https://www.canada.ca/en/department-justice/news/2021/06/government-of-canada-takes-action-to-protect-canadians-against-hate-speech-and-hate-crimes.html>. NetzDG was preceded by the European Commission's *EU Code of Conduct on countering illegal hate speech*. Between its introduction in May 2018 to June 2021 its signatories include Facebook, Microsoft, Twitter and YouTube, Instagram, Snapchat, Dailymotion, Jeuxvideo.com and LinkedIn.

[8]In its *Online Harms White Paper*, the government acknowledged the international scope of the problem posed by online harms, and that, in developing its online harms regime, it was taking into account approaches from other countries, and was 'working closely with international partners as we develop our own approach that reflects our shared values and commitment to a free, open and secure internet', see *Online Harms White Paper* (n 4), paras 2.16-2.17, pp 38-39. The government reiterated this commitment to working with its 'international partners' in its full response to the White Paper consultation, see *Online Harms White Paper: Full government response to the consultation* (n 2), paras 6.4-6.14, 88-90. The White Paper also refers to approaches in Australia (its establishment of an eSafety Commissioner through its Enhancing Online Safety for Children Act in 2015), and the European Commission's Action Plan against disinformation. It also discusses a joint project between the United States and Canada to tackle online child abuse, see *Online Harms White Paper* (n 4).

[9]A first draft was published on the 14 March 2017, and the Bill was introduced on the 16 May 2017. It was passed on the 30 June 2017, and came into force on the 1 October 2017. Companies were given a grace period of three months, until the 1 January 2018, to comply with the legislation.

Reasons behind the regulatory zeitgeist

Governments around the world are facing acute pressure to sanitise the online environment.[10] This is largely based on the increasingly popular notion that despite the benefits to free speech and the public sphere wrought by social media its role in proliferating and intensifying harmful speech warrants rethinking its contribution to society and democracy, as well the motives and responsibilities of online platforms.[11] Until very recently, in the UK and elsewhere, the problems associated with social media have been managed with traditional legal tools that are supplemented with 'soft law' in the form of industry voluntary self-regulation.[12] In this section, I use the UK's two-tier liability system as an example of this framework. However, because the second tier of this system applies to online intermediaries and relates to the European Union's (EU) E-Commerce Directive[13] (which is applicable across the EU and has been 'retained' in UK law by the European Union (Withdrawal) Act 2018), and because a number of the 'soft law' self-regulatory initiatives are cross-jurisdictional, I draw on examples from both the UK and Germany to illustrate the system's failings.

Liability for online speech from the UK: a messy business

Currently in the UK liability for online speech is regulated by a complex and fragmented two-tier system.[14] The first tier consists of online publishers, i.e. any individual or organisation who publishes content online (this includes

[10]In its *Online Harms White Paper*, published in April 2019, the government acknowledged the international scope of the problem posed by online harms: *Online Harms White Paper* (n 4), paras 2.16-2.17, pp 38-39. In respect of hate speech specifically, the Alan Turing Institute has said that COVID-19 has intensified this pressure as, according to the United Nations High Commissioner for Human Rights, 'the pandemic may drive more discrimination, calling for nations to combat all forms of prejudice and drawing attention to the 'tsunami of hate' that has emerged'. Alan Turing Institute, 'Detecting East Asian prejudice on social media' <https://www.turing.ac.uk/research/research-projects/hate-speech-measures-and-counter-measures/detecting-east-asian-prejudice-social-media> accessed 21 May 2022.

[11]According to the European Commission: '…the constantly rising influence of online platforms in society, which flows from their roles as gatekeepers to content and information, increases their responsibilities towards their users and society at large'. European Commission, 'Communication: Tackling Illegal Content Online. Towards an enhanced responsibility of online platforms' COM 2017 555 final (28 September 2017), para 6.

[12]For example, in Germany, prior to the introduction of NetzDG, section 130 of the Criminal Code 1998 operated alongside a self-regulatory 'task force' which included Facebook, Twitter and Google (including YouTube). See: Task Force Umgang mit rechtswidrigen Hassbotschaften im Internet, 'Gemeinsam gegen Hassbotschaften', 15 December 2015. For a comparison of the regulation of hate speech in different jurisdictions, see M Rosenfeld, 'Hate Speech in Constitutional Jurisprudence: A Comparative Analysis' (2001) 24 *Cardozo Law Review* 1523.

[13]Directive 2000/31/EC. The Directive was implemented into UK law by the Electronic Commerce (EC Directive) Regulations 2002.

[14]Andrew Scott has described this system as an 'unwholesome layer cake', see A Scott, An unwholesome layer cake: intermediary liability in English defamation and data protection law' in D Mangan and L Gillies, *The Legal Challenges of Social Media* (Edward Elgar 2017), 222.

the mainstream media, individual social media users, bloggers, and anyone with a website). The second tier applies to online intermediaries that enable the sharing and dissemination of online content. These include user-to-user intermediary services (such as Facebook, Twitter, Instagram and other social media platforms) and search engines (including the likes of Google and Bing).[15] This second tier is supplemented by a variety of voluntary self-regulatory initiatives and schemes.

In respect of tier one, liability arises at the point of publication of illegal content, such as content that is defamatory, in breach of data protection law, infringes copyright or is criminal (which can include hate speech offences). The current criminal regime for dealing with online speech adds further complexity and fragmentation to the two-tier system. As a result, it is subject to extensive guidance from the Crown Prosecution Service[16] and the Sentencing Council.[17] In the case of social media, such content may involve the commission of a range of 'substantive offences', including offences against the person, public justice offences, sexual offences or public order offences.[18] Where social media is not used to commit a substantive offence prosecutors can consider communications offences contrary to section 1(1) of the Malicious Communications Act 1988[19] and / or section 127(1) of the Communications Act 2003.[20] In respect of hate crimes specifically, sections 29–32 of the Crime and Disorder Act 1998 create racially or

[15]Since the 1 of November 2020, UK-established video-sharing platforms (VSPs) have been subject to the Audiovisual Media Services Regulations 2020. These Regulations place requirements on UK-established VSPs to protect their users from certain types of harm, including incitement to hatred. According to the *Online Harms White Paper*, the government intends for the regulation of UK-established VSPs to eventually be part of the online harms regime (see *Online Harms White Paper* (n 4), para 3.5, p 54). Ofcom, as the regulator for UK-established VSPs, maintains a register of the platforms, which is available here: Ofcom, 'Notified Video Sharing Platforms', <https://www.ofcom.org.uk/tv-radio-and-on-demand/information-for-industry/vsp-regulation/notified-video-sharing-platforms>, accessed 21 May 2022.

[16]See: Crown Prosecution Service, Social Media - Guidelines on prosecuting cases involving communications sent via social media, 21 August 2018; Homophobic, Biphobic and Transphobic Hate Crime - Prosecution Guidance, 16 October 2020; Racist and Religious Hate Crime – Prosecution Guidance, 21 October 2020; Disability Hate Crime and other crimes against Disabled people - Prosecution Guidance, 21 October 2020. All available via: Crown Prosecution Service, 'Hate Crime', <https://www.cps.gov.uk/crime-info/hate-crime> accessed 21 May 2022.

[17]Sentencing Council, 'Hate Crime Sentencing', <https://www.sentencingcouncil.org.uk/explanatory-material/magistrates-court/item/hate-crime/> accessed 21 May 2022.

[18]For a non-exhaustive list of examples, see CPS, Social Media - Guidelines on prosecuting cases involving communications sent via social media (n 16).

[19]Pursuant to s 1(1), a person who sends to another person 'a letter, electronic communication or article of any description which conveys (a) (i) a message which is indecent or grossly offensive; (ii) a threat; or (iii) information which is false and known or believed to be false by the sender' is guilty of an offence if their purpose, or one of their purposes, in sending it is that it should 'cause distress or anxiety to the recipient or to any other person to whom they intend that it or its contents or nature should be communicated'. Under ss 1 2A(a) and (b), 'electronic communication' includes any oral or other communication by means of an electronic communications network and any communication (however sent) that is in electronic form. If convicted on indictment a defendant can receive a sentence of up to two years imprisonment, or a fine, or both. If tried and convicted summarily they can be sentenced to up to twelve-month imprisonment, or receive a fine, or both (ss 14(a) and (b)).

[20]Communications Act 2003, s 127(1) makes it an offence to send through a 'public electronic communications network' a message which is 'grossly offensive or of an indecent, obscene, or menacing

religiously aggravated forms of assault,[21] criminal damage,[22] public order offences[23] and 'harassment etc'.[24] However, if a substantive offence is caused by a social media communication, or if an offence has been committed under section 1 of the 1988 Act or section 127 of the 2003 Act, that is driven by 'hostility' toward a group or individual because of race, religion, sexual orientation or transgender identity, or disability, section 66(2) of the Sentencing Code[25] requires magistrates and judges to regard the 'hostility' toward the hate crime characteristics as an aggravating factor when determining sentence.[26]

The liability of online intermediaries, including social media platforms, falls under tier two. The scope of this liability is *limited* because the regime is subject to the 'safe-harbour' protections for intermediaries provided by Articles 12–15 of the E-Commerce Directive that were designed to protect free speech and the privacy rights of users. Under this regime, intermediaries are not subject to an *ab initio* duty to ensure that only lawful content is hosted or indexed, which means that liability for content that is, for example, defamatory, or in breach of data protection laws or criminal, *will only crystallise* if (i) the intermediary has been notified that they are hosting or indexing illegal content, including hate speech (thus, there is no obligation on platforms to pre-emptively block unlawful content) *and* (ii) they then fail to remove or de-index the unlawful content expeditiously.[27]

character'. Section 127(3) tells us that a person guilty of an offence under s 127 shall be liable, on summary conviction, to imprisonment for a term not exceeding six months or to a fine, or to both.

[21]Crime and Disorder Act 1998, s 29.

[22]ibid, s 30.

[23]ibid, s 31.

[24]ibid, s 32.

[25]This Code is found in s 66 of the Sentencing Act 2020. This provision consolidates ss 145 and 146 of the Criminal Justice Act 2003 which provided a sentencing uplift for hate crime offending, and which applied to communications offences where the defendant demonstrates hostility to the victim based on the victim's protected characteristic (race, religion, disability, sexual orientation or transgender identity). Sections 145 and 146 have been repealed by the Sentencing Act 2020, sch 28.

[26]See also Sentencing Council, 'Hate Crime Sentencing', <https://www.sentencingcouncil.org.uk/explanatory-material/magistrates-court/item/hate-crime/> accessed 21 May 2022.

[27]ibid, art 14. Notwithstanding the E-Commerce Directive, these principles have also been consistently applied in UK case law relating to internet service providers (ISPs), user-to-user platforms, and search engines. For example, see *Bunt v Tilly and Others* [2007] 1 WLR 1243 (ISP); *Metropolitan International Schools Limited v Designtechnica Corp. and Others* [2011] 1 WLR 1743 (search engine); *Tamiz v Google Inc* [2013] 1 WLR 2151 (user-to-user social media platform). However, in other jurisdictions these decisions have not always been followed. For example, see *Oriental Press Group v Fevaworks Solutions* [2013] HKFCA 47; *Yeung v Google Inc* [2014] HKCFI 1404; *A v Google New Zealand Limited* [2012] NZHC 2352; *Google v Trkulja* [2016] VSCA 333; *Defteros v Google LLC* [2020] VSC 219. In Germany, online intermediaries can only rely on the Directive's safe-harbour exceptions if they act promptly on complaints of third parties whose rights have been infringed by their users. The *Bundesgerichtshof*, the Federal Court of Justice, has ruled that if intermediaries do not comply to a take-down request 'without culpable hesitation' (s 121 of the German Civil Code), they are to be treated as if they posted the content (s 7 of the Telemedia Act), see BGH, *Internet-Versteigerung I* (11 March 2004) I ZR 304/01; BGH, *Jugendgefährdende Medien bei eBay* (12 July 2007) I ZR 18/04, [42]; BGH, *Alone in the Dark* (12 July 2012) I ZR 18/11, [28].

Moreover, pursuant to Article 15, courts cannot order intermediaries to undertake general monitoring of hosted or indexed content in order to detect something that may be unlawful.

What about self-regulation?

These legal tools have been supplemented by self-regulatory voluntary frameworks that, until relatively recently, were the preferred approach to addressing harmful online speech.[28] These frameworks are wide-ranging, both in terms of their substance and application, in that some responsibilities are generally applicable (and therefore 'attached' to platforms as opposed to jurisdictions), whereas others are country or region-specific. For instance, they include platforms adopting hate speech policies[29] and making regular public statements that they are acting appropriately to tackle harmful content.[30] In Germany, Facebook, Twitter, Google and YouTube are part of a self-regulatory, albeit non-binding, task force committed to removing harmful content quickly, introducing or improving internal reporting mechanisms, and employing more local experts and lawyers to undertake supervision.[31] In 2016, a year after the creation of the task force, a similar, non-binding, commitment to self-regulation was made by Facebook, Microsoft, Twitter and YouTube who, together with the European Commission, launched *The EU Code of conduct on countering illegal hate speech online*.[32] In line with Article 14 of the E-Commerce Directive the companies agreed to review the majority of notifications received by online users in less than twenty-four hours and to implement procedures to remove notified content when considered illegal. Additionally, the signatories committed to, inter alia, publishing community guidelines setting out the prohibition of incitement to violence and hateful conduct on their platforms, raising the awareness of their staff, working more closely with state authorities, and providing information regarding their rules on reporting and notification processes for illegal content.[33]

[28]See Joint Special Rapporteurs, *Declaration on Freedom of Expression and the Internet* (Organisation for Economic Co-Operation and Development, 1 June 2011), general principle 1(e); Council of Europe Committee of Ministers, *Declaration on freedom of communication on the Internet* (28 May 2003), principle 2, encouraging 'self-regulation or co-regulation regarding content disseminated on the Internet'. See also Quintel and Ullrich (n 4), 197.

[29]See Part III section 12 of Facebook's Community Standards; Twitter's Hateful conduct policy.

[30]For instance, at the time of writing, Instagram announced new features to its platform that will restrict hate speech and abusive content, see Adam Mosseri, 'Introducing New Ways to Protect our Community from Abuse' (Instagram 10 August 2021) <https://about.instagram.com/blog/announcements/introducing-new-ways-to-protect-our-community-from-abuse>.

[31]Task Force Umgang mit rechtswidrigen Hassbotschaften im Internet, 'Gemeinsam gegen Hassbotschaften', 15 December 2015.

[32]See (n 7).

[33]For a detailed examination of the Code, see Quintel and Ullrich (n 4), 197.

Does the current system work?

In recent years, the inadequacy of existing liability systems for harmful online speech have been exposed. While an examination of the manifold reasons for this inadequacy is beyond the scope of this article, there are arguably three primary causes that are illustrated by real-world examples, as well as a theoretical reason that fundamentally undermines the rationale upon which the E-Commerce Directive is based.

Primary causes and real-world examples

Firstly, liability systems that target publishers of harmful content, such as the tier-one regime in the UK, were not designed to deal with online speech and, perhaps understandably, are unable to cope with the practicalities of the online environment.[34] In the UK the Law Commission has recommended that section 1(1) of the 1988 Act and section 127(1) of the 2003 Act be repealed and replaced with a consolidated harm-based model.[35] This is because the methods and frequency of communication, the types of content that may be published, and the number of publishers disseminating harmful content, have 'fundamentally changed' because of the internet and social media.[36] Yet, notwithstanding the Law Commission's recommendation for reform, this 'suite' of criminal offences has changed very little to adapt to this communication revolution. Indeed, the section 127(1) offence is largely based on section 10(2)(a) of the Post Office (Amendment) Act 1935 and, as a result, 'it is perhaps no surprise that criminal laws predating widespread internet and mobile telephone use (to say nothing of social media) are now of inconsistent application'.[37]

The amount of individuals publishing harmful content at an exponentially increasing frequency, combined with the inadequacy of existing offences to address the online environment, has led to two paradoxical phenomena. On the one hand, these laws *under-criminalise*, in that despite causing substantial harm many culpable and damaging communications evade appropriate criminal sanction because the offences allow for some abusive, stalking, and bullying behaviours to 'simply fall through the cracks'.[38]

[34]For a critique of the UK's tier-one regime, see generally P Coe, 'The social media paradox: an intersection with freedom of expression and the criminal law' (2015) 24 *Information & Communications Technology Law* 1, 16-40.

[35]Law Commission, *Modernising Communications Offences. A final report* (Law Com No 399, 2021) 24, paras 2.38-2.39. At the time of writing, the UK Government has announced that the new offences recommended by the Commission will be added to the Draft Online Safety Bill. The government has also confirmed that they will not apply to 'regulated media such as print and online journalism, TV, radio and film', see Department for Digital, Culture, Media and Sport, 'Update on Law Commission's Review of Modernising Communications Offences' (4 February 2022) <https://questions-statements.parliament.uk/written-statements/detail/2022-02-04/hcws590>.

[36]ibid 1, para 1.3.

[37]ibid 2, para 1.4.

[38]ibid 2, para 1.5.

On the other hand, by proscribing content on the basis of 'apparently universal standards' – such as 'indecent' or 'grossly offensive' content – the law as it stands criminalises without regard to the potential for harm in a given context,[39] thereby *over-criminalising*.[40] Although these issues existed prior to the internet and social media, the exponential increase in expression facilitated by the internet and social media platforms has exacerbated them. Thus, notwithstanding the exercise of prosecutorial discretion, this approach has the potential to interfere with freedom of expression by criminalising speech, and therefore possibly preventing it from being heard, without a proper contextual assessment of the harm it causes and whether it *actually* meets an objective standard of criminality. Furthermore, over-criminalisation could 'swamp the criminal justice system'.[41] Even if, for a moment, you set aside the number of publishers that could theoretically be prosecuted for publishing harmful content, which in itself would require enormous police and prosecution resources and would likely bring any national prosecuting agency and court service to a standstill, the transience of online publishers, the fact they operate across different jurisdictions, and the frequency with which they publish anonymously or pseudonymously, means that even locating and identifying them is challenging.[42] Secondly, in respect of intermediaries, as we have seen above, the E-Commerce Directive restricts liability for these actors. Finally, social media platforms have consistently failed to meet their commitments to self-regulate.

I could point to any number of real-world events that animate these three causes which have intensified calls for an overhaul of the current framework for intermediary liability from a multitude of actors. These calls have led to the introduction of online-specific legal tools, such as the Bill and NetzDG,[43] that are designed to tackle hate speech (and other forms of harmful speech) within the online environment.[44] However, for the purpose of this article, I will briefly sketch two particularly high-profile events from the UK and Germany respectively.

[39]For example: 'Two consenting adults exchanging sexual text messages are committing a criminal offence, as would be the person saving sexual photographs of themselves to a 'cloud' drive', ibid 2, para 1.6.

[40]ibid.

[41]ibid.

[42]See generally P Coe, *Media Freedom in the Age of Citizen Journalism* (Edward Elgar 2021), ch 7; P Coe, 'Anonymity and pseudonymity: Free speech's problem children' (2018) 22 *Media & Arts Law Review* 2, 173-200.

[43]See (n 7).

[44]In December 2020, the UK government confirmed that hate speech will fall within the remit of the Bill. According to the response, a 'limited number of priority categories of harmful content, posing the greatest risk to users, will be set out in secondary legislation' which will include 'hate crime'. Furthermore, 'hate content' is one of the 'priority categories' that will be set out by the government in secondary legislation, see: *Online Harms White Paper: Full government response to the consultation* (n 2), paras 2.3 and 2.29.

In the UK, the inadequacy of the current framework for intermediary liability is illustrated by the level of online hate speech targeting professional footballers during the 2020 UEFA European Championship.[45] This included racist abuse of Marcus Rashford, Jadon Sancho and Bukayo Saka after England's loss to Italy in the final. Much of this abuse took place on Twitter, and it has since transpired that although the platform permanently suspended the accounts of fifty-six persistently abusive users on the 12 July 2021 (the day after the final) thirty of those offenders continued to post, or 'respawn', on the network, often under slightly altered usernames.[46] Consequently, Dame Melanie Dawes, the Chief Executive of Ofcom, stated that these events brought '[t]he need for regulation … into even sharper focus'.[47]

Germany's motivation for introducing NetzDG was largely fuelled by changes to the political climate in the country in 2016 as a result of a large influx of refugees into the country, which had started in 2015.[48] Although, according to Karsten Müller and Carlo Schwarz, the impact of social media on this change to the climate is difficult to quantify, what is clear is that at a time when the traditional institutional media outlets supported the government's migration policy, social media gave critics of the policy an alternative public arena to organise themselves and express their views.[49] Consequently, Thomas Wischmeyer explains that '[f]or some, social media proved to be not only a tool of communicative self-empowerment, but also a mechanism to fuel resentment and to spread hatred and defamation' which turned aspects of social media into a 'toxic environment for minorities and, in particular, refugees'.[50]

One high-profile event seemed to be the catalyst for this political maelstrom[51] and, at the same time brought into sharp focus not only the inadequacy of a liability regime severely hamstrung by the E-Commerce Directive, but also the failings of social media platforms to meet their self-regulatory commitments; the combination of which ultimately failed to protect an individual's rights. In September 2015, Anas Modamani, a Syrian refugee, photographed a 'selfie' with the German Chancellor, Angela Merkel, during her visit to his shelter in Berlin. The picture subsequently became a symbol for

[45]Due to COVID-19 the tournament was played in June and July 2021.
[46]P MacInnes, 'Twitter users banned after racist abuse of England players still posting online' *The Guardian*, 13 August 2021.
[47]Dame Melanie Dawes, Ofcom, 'In news we trust: keeping faith in the future of media', (Oxford Media Convention, 19 July 2021) (Keynote speech).
[48]T Wischmeyer, ''What is illegal offline is also illegal online': the German Network Enforcement Act 2017' in B Petkova and T Ojanen, *Fundamental Rights Protection Online. The Future Regulation of Intermediaries* (Edward Elgar Publishing 2020), 28-56, 31; K Müller and C Schwarz, 'Fanning the Flames of Hate: Social Media and Hate Crime' (SSRN, 5 June 2020), <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082972>.
[49]ibid. See generally: Müller and Schwarz (n 48); Wischmeyer (48) 31.
[50]Wischmeyer (48).
[51]ibid.

Chancellor Merkel's 'open borders' policy. However, in 2016 false content was published on Facebook stating that Modamani was involved in the 2016 Brussels bombings, which in turn suggested a link between Merkel and terrorism. Following a request from Modamani, Facebook removed and geo-blocked specific *existing* posts, yet it declined to pre-emptively filter all new posts, which led to Modamani applying for a preliminary injunction against the platform. Unfortunately for the applicant, in 2017 the Würzburg District Court ruled that, inter alia, because of Articles 14 and 15 of the E-Commerce Directive, Facebook, as the host platform, could not be made to pre-emptively block any offensive content that may violate Modamani's rights.[52] Thus, the circumstances of the case, and the decision itself, served to highlight to the world-at-large that: (i) the E-Commerce Directive was depriving victims of hate speech and other harmful content of their rights; (ii) existing laws on the limits of free speech were not being, and could not be, effectively enforced within the online environment;[53] (iii) social media platforms generally were simply paying lip service to their self-regulatory commitments and, in Germany, the self-regulatory task force set up in 2015 was not coming close to meeting its promises.[54]

Theoretical arguments: the E-Commerce Directive and the active/passive distinction

Recital 42 of the E-Commerce Directive explains that the limitations it places on the liability of online intermediaries (which it refers to as 'information society service providers') exist because of their passivity in the curation and dissemination, and hosting or indexation of content on their platforms. Accordingly, the Recital says that they do no more than engage in 'the technical process of operating and giving access to a communication network over which information made available by third parties is transmitted or temporarily stored for the sole purpose of making the transmission more efficient'. The reasoning for this is that, according to the Directive, such activity is of a 'mere technical, automatic and passive nature, which implies that the information society service provider has neither knowledge of nor control over the information which is transmitted or stored'.

[52]Landgericht Würzburg, Case 11 O 2338/16 UVR [2017].

[53]Wischmeyer (n 48) 32.

[54]Indeed, research published in 2017 by the German public watchdog *jugendschutz.net* found that despite the requirements under German law and the Directive itself, that to be able to rely on its safe-harbour exception intermediaries must act expeditiously to remove illegal content upon being notified of its existence: (i) all major online platforms were very slow to act on take-down requests; (ii) the notification processes set up by platforms were overly complex; and (iii) the removal quota for illegal content ranged from a satisfactory 90 per cent (YouTube), to an unsatisfactory 39 per cent (Facebook), to an appalling 1 per cent (Twitter), see jugendschutz.net, 'Löschung rechtswidriger Hassbeiträge bei Facebook, YouTube und Twitter' (March 2017).

Although the Recital's rationale certainly fits with the public message that is often conveyed by social media platforms that they are merely passive technology companies as opposed to active media companies that perform editorial functions, their actions consistently suggest that they are, in fact, operating as both,[55] thereby undermining the Recital's theoretical basis. This is illustrated by how social media platforms use algorithms to curate news content and influence what users see.[56] These algorithms shape how content is aggregated, presented and distributed, and how users consume content by producing a personalised news feed for each and every user using settings that are dependent on, but not entirely under the control of, the respective user.[57] By presenting content in a particular way, or by removing material because it conflicts with the respective platform's business goals or ideology, or contravenes its own policies, Facebook, and Twitter et al are playing an editorial-like role.[58] Google has also found itself at the centre of this debate in Europe and in Australia. For example, Frank Pasquale found that depending on the issue and commercial interest at stake Google opportunistically characterises itself as a passive speech conduit and/or an active content provider;[59] a practice illustrated by the European Commission fining the company €2.42 billion for manipulating the search rankings of its search engine in favour of its own products.[60] Similarly, in Australia, in *Defteros v Google LLC*[61] Justice Richards found that Google is a publisher because its search engine is 'not a passive tool' as it is 'designed by humans who work for Google to operate in the way it does, and in such a way that identified objectionable content can be removed, by human intervention'.[62]

[55]For detailed discussion of this debate, see Coe (n 42) 60-65.

[56]Although Twitter gives more control to its users over the curation of their news feeds, it still makes editorial decisions by, for example, removing content that infringes legislation or its own policies.

[57]A Koltay, New Media and Freedom of Expression Rethinking the Constitutional Foundations of the Public Sphere (Hart Publishing 2019) 158; L Andrews, Facebook, The Media and Democracy: Big Tech, Small State? (Routledge 2020) 60.

[58]Because of this, Natalie Helberger has argued that Facebook is a 'new breed of social editor'. N Helberger, 'Facebook is a new breed of editor: a social editor' (LSE Media Policy Project, 16 June 2017). In respect of Facebook, Tarleton Gillespie has argued that its switch from a chronological news feed to an algorithmically curated one, meant that it began to produce 'a media commodity', see T Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018), 43. Similarly, Timothy Berners-Lee has argued that Facebook is making billions of editorial decisions every day, see TB Lee, 'Mark Zuckerberg is in denial about how Facebook is harming our politics' (Vox, 10 November 2016).

[59]F Pasquale, 'Platform Neutrality: Enhancing Freedom of Expression in Spheres of Private Power' (2016) 17 *Theoretical Inquiries in Law* 487, 512.

[60]'European Commission, 'Press Release – Antitrust: Commission Fines Google €2.42 Billion for Abusing Dominance as Search Engine by Giving Illegal Advantage to Own Comparison Shopping Service' (27 June 2017), <https://ec.europa.eu/commission/presscorner/detail/en/IP_17_1784>.

[61][2020] VSC 219.

[62]ibid [40]. In 2019 the Australian Competition and Consumer Commission found that: 'Digital platforms like Google and Facebook are more than mere distributors or pure intermediaries in the supply of journalism … They increasingly perform similar functions as news media businesses such as selecting, publishing and ranking content, including significant amounts of news media content', see ACCC,

Contrary to the Recital's rationale, and despite corporate messages that they are simply passive technology companies, there is an abundance of evidence (such as what I have set out above) which points convincingly to the fact that online platforms are increasingly playing an active role in the curation and dissemination and hosting or indexation of content. This has resulted in a blurring of the active and passive activities and functions that they perform, which has in turn rendered the Directive's theoretical foundation - its active/passive distinction in this context - obsolete.[63] Moreover, from a practical perspective, as illustrated by the examples sketched above, the Directive's exemptions do not reflect the modern online environment, and social media as an industry; it does not take into account why content is managed as it is (for instance, to serve the respective platform's ideological or commercial agendas), internet business models (for example, the use of clickbait and the manipulation of content coverage to attract users/readers and therefore more advertising revenue), and how platforms have diversified (from simple hosting platforms to multinational and multimedia conglomerates).[64]

A comparison of the main principles of the Draft Online Safety Bill and NetzDG and what these may mean for free speech: have we opened Pandora's box?

The Bill and NetzDG represent the UK and German governments' solution to the online harms problem and a way of remedying the defects of the current system of liability. As previously stated, NetzDG came into force in 2017,[65] and therefore foreshadowed much of the legislative developments that have since swept Europe and beyond. In developing the UK's online harms regime, and in drafting the Bill, the government drew on NetzDG and the German experience.[66] Consequently, as we shall see, the two pieces of legislation share similarities, but there are also some important distinctions. In this section, for context, I begin by setting out the scope and oversight of both regimes. This leads into a critical discussion about core

'Examining the impact of digital platforms on competition in media and advertising markets' (27 February 2019).

[63]Quintel and Ullrich (n 4) 221.

[64]For a detailed discussion of these issues, see Coe (n 42) ch 3. Facebook has rebranded itself as 'Meta' (although the change does not apply to its individual platforms, such as Facebook, Instagram and Whatsapp, only the parent company that owns them). According to the company, the new name will better "encompass" what it does, as it broadens its reach beyond social media into areas like virtual reality. In announcing the new name, Mark Zuckerberg says he plans to build a "metaverse" - an online world where people can game, work and communicate in a virtual environment, often using VR headsets. See: D Thomas, 'Facebook changes its name to Meta in major rebrand' *BBC News,* 28 October 2021.

[65]See (n 9).

[66]See (n 7) and (n 8).

aspects of each piece of legislation and some of the key free speech concerns they have generated, which in summary is that the legislation leads to privatisation of censorship, which incentivises platforms to over-censor contested but legal speech, thereby reducing, or even silencing, legitimate debate. [67]

Scope and oversight

Who is in scope?

Services that are within the scope of the Bill are 'user-to-user services' (in other words, an internet service that enables user-generated content, such as Facebook or Twitter) and 'search services' (such as Google)[68] that have links with the UK (in that the service is capable of being used in the UK, or there are 'reasonable grounds to believe there is a material risk of significant harm to individuals' in the UK from the content or the search results).[69]

Clause 39 and Schedule 1 specify the services and content that are *excluded* from the regime, albeit these are limited by various caveats. These include emails,[70] SMS messages and MMS messages.[71] However, the exclusion applies only if the services or content represent 'the only user-generated content enabled by the service' meaning that, Facebook Messenger for example, is not exempt, and will therefore be regulated. The Schedule 1 exemption also applies to internal business services,[72] comments and reviews on provider content,[73] paid-for advertisements[74] and news publisher content (though the site needs to be a 'recognised news publisher' pursuant to clause 40),[75] certain public bodies services,[76] and 'one-to-one live aural communications'[77] (these are communications made in real time between users, although the exclusion applies only if the communications consist solely of voice or other sounds, and do not include any written message, video or other visual images, meaning that Zoom, for instance, does not qualify for the exemption, and is within the Bill's scope). Finally, the Bill

[67]It would be impossible within the scope of this article to provide analysis of all of the free speech concerns. For further analysis of NetzDG in this special issue, see Uta Kohl, 'Platform Regulation of Hate Speech – A Transatlantic Speech Compromise?', section 3 and Mathias Hong, 'Regulating hate speech and disinformation online while protecting freedom of speech as an equal and positive right – comparing the fundamental rights frameworks in Germany, Europe and the United States', section 3. For discussion on China's approach to dealing with online hate speech, and how it compares with NetzDG, see Ge Chen, 'How equalitarian regulation of online hate speech turns authoritarian: a Chinese perspective'.
[68]cls 1 to 3.
[69]cl 3(3) to (6).
[70]cl 39(2) and sch 1, cl 1.
[71]cl 39(2) and sch 1, cl 2.
[72]sch 1, cl 4.
[73]cl 39(5) and sch 1, cl 5.
[74]cl 39(7).
[75]cl 39(8)-(11).
[76]sch 1, cl 6.
[77]cl 39(6) and sch 1, cl 3.

gives significant power to the Secretary of State for Digital, Culture, Media, and Sport to amend Schedule 1, and either add new services to the list of exemptions or remove some of those already exempt, based on an assessment of the risk of harm to individuals.[78]

Although different terminology is used, prima facie, NetzDG regulates similar services to the Bill as, pursuant to section 1, it applies to online service providers which, 'for profit-making purposes, operate internet platforms which are designed to enable users to share any content with other users or to make such content available to the public'. However, its scope is more limited than the Bill in that under section 1(2) only platforms with more than two million registered users in Germany are obliged to apply the most relevant provisions. Like the Bill however, section 1 of the legislation excludes platforms 'offering journalistic or editorial content' and sites hosting only 'specific content' (such as online review sites, shops or games). Professional networks, such as LinkedIn, are also exempt.

What content is in scope?

The Bill is vague on the type of content that it covers. Essentially, it covers 'illegal content', which for user-to-user services is 'regulated content' (user-generated content) that 'amounts to a relevant offence',[79] and for search services is content that amounts to a relevant offence.[80] Thus, hate speech content is covered. Additionally, and controversially for reasons I discuss below, it imposes 'safety duties' on regulated services in relation to content that is legal but 'harmful' to adults and children. To the contrary, the obligations that NetzDG imposes on platforms pertain to specific types of illegal speech which are explicitly set out under section 1(3), all of which are existing offences under the German Criminal Code (GCC). These include, inter alia, incitement to hatred[81] and the defamation of religions, religious and ideological associations.[82] In limiting its scope to these existing offences NetzDG did not create new 'hate speech-specific' laws (for instance), but rather relied on criminal laws that were within the realm of hate speech.

Oversight

Once enacted, under clause 29 of the Bill, the legislation will require Ofcom to issue codes of practice which will outline the systems and processes that companies need to adopt to fulfil their duty of care. It will have

[78]cl 3(8).
[79]'Relevant offence' is defined in cl 41(4).
[80]cl 41(2).
[81]German Criminal Code, s 130.
[82]ibid, s 166.

the power to fine companies up to £18 million, or 10 per cent of annual global turnover, whichever is higher, if they are failing in their duty of care.[83] Ofcom will also be given the power by the legislation to block non-compliant services from being accessed in the UK.[84] The government's response to the White Paper also suggests that Ofcom will be empowered, via secondary legislation, to impose criminal sanctions against individual executives or senior managers of regulated services if they do not respond fully, accurately and in a timely manner to information requests by the regulator.[85]

In Germany there is not a NetzDG regulator per se, rather section 4(1) makes it an administrative offence, punishable with a fine of up to €50 million,[86] for platforms to fail to produce a report or to implement sufficient procedures, or otherwise not comply with the requirements of the legislation. Pursuant to sections 4(4) and (5) the Federal Office of Justice is responsible for making determinations on the issuing of fines.

Duties of care/responsibilities and free speech

Overview

The existing liability regime in the UK, like the regime in Germany prior to the enactment of NetzDG, is only interested in the *output* – in that what matters is that illegal content is removed expeditiously once notice has been given. How platforms manage this is entirely up to them, and is therefore a rather opaque process, at least to the outside world. The Bill, once enacted, will change this, as the extensive and multi-layered duties of care imposed on regulated services operate at the systems and processes level *and* the content level.[87] Similarly, pursuant to sections 2 and 3, NetzDG regulates the design and performance of the internal systems used by platforms to deal with the large number of justified and unjustified requests to remove content.[88] However, unlike the Bill, which, as detailed below, requires regulated services to protect users from illegal content *and* content that is 'harmful' but not illegal, the purpose of NetzDG was not to regulate or criminalise previously legal speech, or in other ways extend the zone of what is 'unspeakable'. Rather, its novelty lies solely in the new procedural and organisational obligations placed on regulated services.

[83] cl 85(4).
[84] cl 91.
[85] *Online Harms White Paper: Full government response to the consultation* (n 2), 12, para 38.
[86] See s 4(2) NetzDG in conjunction with ss 30 and 130 of the German Act on regulatory offences (*Ordnungswidrigkeitengesetz*).
[87] Online Safety Bill, Explanatory Notes, Bill CP 405-EN, [7], 5. See the general duties imposed on user-to-user and search engine services in Part 2 Chapters 2 and 3 of the Bill.
[88] Section 2 prescribes a reporting obligation on platforms, and s 3 imposes obligations on regulated services relating to the handling of complaints about unlawful content.

The Bill's duties of care

The Bill, by contrast, sets out layers of duties that include: (i) general duties of care applying to user-to-user services[89] and search services;[90] (ii) additional duties for user-to-user services relating to children;[91] (iii) additional duties for 'Category 1 Services' (which are currently undefined user-to-user services to be included in a register maintained by Ofcom, pursuant to clause 59(6)). Essentially, these duties consist of, inter alia, 'harder' and manifold safety duties obliging services to protect users from 'illegal content',[92] which will include hate speech (although, as discussed below, this is undefined), and protecting children[93] and adults from legal yet (again, as discussed below, undefined) harmful content (in respect of adults this duty only applies to Category 1 Services).[94] The 'hard-edge' of these safety duties is, perhaps, best exemplified by clause 9(3), which has been described as being at 'the heart of draft Bill',[95] and clause 10(3), which, as I discuss below, are significant, in not only how they differ from NetzDG and the E-Commerce Directive, but also because of what they may mean for free speech when one takes into account the 'softer-edged' free speech duties.

Clause 9(3) imposes:

> 'A duty to operate a service using proportionate systems and processes designed to (a) minimise the presence of priority illegal content;[96] (b) minimise the length of time for which priority illegal content is present; (c) minimise the dissemination of priority illegal content; (d) where the provider is alerted by a person to the presence of any illegal content, or becomes aware of it in any other way, swiftly take down such content.'

Clause 10(3)(a) imposes:

> 'A duty to operate a service using proportionate systems and processes designed to (a) prevent children of any age from encountering, by means of the service, primary priority content that is harmful to children.'[97]

Duties (a) to (c) of clause 9(3) and clause 10(3)(a) in theory fall foul of Article 15 of the Directive which, as discussed, prohibits states from imposing general monitoring obligations on hosting providers. Furthermore, although

[89]cl 5(2).
[90]cl 17(2).
[91]cl 10 and 22.
[92]cls 9 and 21.
[93]cls 10 and 22.
[94]cls 11.
[95]G Smith, 'Harm Version 3.0: the draft Online Safety Bill', *Inforrm's Blog* (1 June 2021).
[96]Clause 41(5) tells us that 'Illegal content is 'priority illegal content' if the relevant offence is an offence that is specified in, or is of a description specified in, [currently undrafted] regulations made, pursuant to clause 44, by the Secretary of State'.
[97]Clause 22(3)(a) mirrors clause 10(3)(a) albeit it applies to 'search results' and 'prevent' is replaced with 'minimise'.

duty (d) of clause 9(3) appears to mirror the hosting liability shield provided by Article 14 this duty is put in an entirely different light by being cast in terms of a *positive* regulatory obligation to operate take down processes, rather than *potential exposure* to liability for a user's content should the shield be disapplied on gaining knowledge of its illegality. Thus, in imposing these duties, these clauses signal a clear policy departure from the rationale that underpins the Directive's safe-harbour protections, and from twenty years of EU and UK policy aimed at protecting the freedom of expression and privacy of online users.[98] This policy 'departure' is made more acute when the 'softer-edge' of the free speech duties is considered. In the following section, I turn to this concern, and three other related concerns with the overall vagueness of the Bill that could contribute to a significant interference with free speech.

Free speech concerns raised by the bill

Article 10(1) of the European Convention on Human Rights protects freedom of expression by providing: 'Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers'. Article 10(2) qualifies this right, in that a state can restrict the Article 10(1) right in the interests of, inter alia, 'the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others'. In respect of the offline world, the European Court of Human Right's jurisprudence gives the protection afforded by Article 10(1) considerable scope, in that it consistently holds that it is applicable not only to information or ideas ' … that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the State or any sector of the population. Such are the demands of that pluralism, tolerance and broadmindedness without which there is no "democratic society"'.[99] More recently however, the Strasbourg Court's case law has indicated that it is prepared to limit this wide scope to take account of the amplification of the threat posed to countervailing fundamental rights by the internet and online speech, so long as this limitation falls legitimately within the parameters imposed by Article 10(2).[100]

[98]ibid.

[99]*Handyside v United Kingdom* App no 5493/72 (ECHR, 7 December *1976*) [49]; see also *Sunday Times v United Kingdom (No. 1)* App no 6538/74 (ECHR, 26 April *1979*) [65]; *Lingens v Austria* App no 9815/82 (ECHR, 8 July 1986) [41]; *Axel Springer AG v Germany (No. 1)* App no 39954/08 (ECHR, 7 February 2012) [78]; *Thorgeir Thorgeirson v Iceland* App no 13778/88 (ECHR, 25 June *1992*) [63].

[100]For example, see: *Delfi AS v Estonia* App no 64569/09 (ECHR, 16 June 2015) [110]: defamatory and other types of clearly unlawful speech, including hate speech and speech inciting violence, can be disseminated like never before, worldwide, in matter of seconds, and sometimes remain persistently available online'; *Editions Plon v France* App no 58148/2000 (ECHR, 18 August 2004); *Ovchinnikov v*

When referring to 'illegal content', by saying that for user-to-user services it is regulated content that amounts to a relevant offence, and for search services it is content that amounts to a relevant offence, the Bill delegates the definition of those offences to other legislation. Unfortunately, the definition of hate speech is murky and can lead to confusion amongst the public, platforms, Ofcom and even prosecutors, which in turn can have serious implications for the operation of free speech. Without a clear definition of hate speech, it is potentially very easy for the ECtHR's established free speech principles to be illegitimately, but perhaps accidentally, restricted; an issue summed-up in evidence presented to House of Lords Communications and Digital Committee by Ayishat Akanbi, who suggested that the distinction between hate speech and 'speech we hate' can be hard to see.[101] This is not helped by regular changes to definitions of hate speech,[102] and the different legal parameters of hate crime that exist across a raft of criminals laws, including: the Public Order Act 1986, the Crime and Disorder Act 1998, the Criminal Justice Act 2003, the Malicious Communications Act 1988, the Racial and Religious Hatred Act 2006, the Communications Act 2003 and even the Football (Offences) Act 1991. Consequently, at the time of writing, these laws are subject to an ongoing Law Commission consultation that is investigating how they should function in practice and possibilities for reform.[103]

Secondly, the duties outlined in clauses 10 and 22 and clause 11 relating to content that is legal yet 'harmful' to children and adults, respectively, although unlikely to apply to hate speech, are controversial[104] and therefore worthy of consideration. These duties require platforms to identify the potential risks from 'harmful' content, and to specify in their terms of service how they will protect children and adults from such content. The meaning of 'content that is harmful to' children and adults is prescribed by clauses 45 and 46 respectively, pursuant to which content is harmful if 'there is a material risk of the content having, or indirectly having, a

Russia App no 24061/04 (ECHR, 16 March 2011); *Perrin v United Kingdom* App no 5446/03 (18 October 2005); *Editorial Board of Pravoye Delo and Shtekel v Ukraina* App no 33014/05 (ECHR, 5 August 2011) [63]. For a comprehensive overview of the Strasbourg Court's Article 10 jurisprudence in the digital age, see D 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); O Pollicino, *Judicial Protection of Fundamental Rights on the Internet* (Hart Publishing 2021), 68-87.

[101]'Free for all? Freedom of expression in the digital age' (n 5), 17, para 7.

[102]For example, making misogyny a hate crime is included in Lord McNally's Private Members Bill: Online Harms Reduction Regulator (Report) Bill. However, at the time of writing, the current Home Secretary, Priti Patel, has rejected attempts to classify misogyny as a hate crime, arguing that it would deliver only 'tokenistic' change and that adding it to the scope of hate crime laws would make it harder to prosecute sexual offences and domestic abuse. See: M Dathan, 'Patel says misogyny will not be made a hate crime in new law', *The Times,* 21 February 2022, 2.

[103]Law Commission, *Hate crime laws. A consultation paper* (Law Com CP 250, 23 September 2020).

[104]ibid 36-42, paras 151-183.

significant adverse physical or psychological impact on a child [or adult] of ordinary sensibilities'.[105] Clauses 45(7) and 46(6) stipulate that where the platform has knowledge about a particular child or adult at whom relevant content is directed, or who is the subject of it, then the child's or adult's 'characteristics' must be taken into account. Unfortunately, this is the limit of the Bill's explanation of what amounts to legal yet 'harmful' content. It does not account for the fact that how we determine what is harmful will depend on the individual concerned, nor does it define a child or adult of 'ordinary sensibilities' or prescribe the 'characteristics' that would make them more susceptible to harm. As the Bill currently stands, evaluating user content will be entrusted to the subjective judgment of the platform. The implications for free speech are discussed below.

Thirdly, clauses 12 and 23 set out a general duty applicable to user-to-user and search services respectively to 'have regard to the importance of': (i) 'protecting users' right to freedom of expression' *and* (ii) 'protecting users from unwarranted infringements of privacy'. In addition, clause 13 provides 'duties to protect content of democratic importance' and clause 14 prescribes 'duties to protect journalistic content'. However, unlike the clauses 12 and 23 duty, the clause 13 and 14 duties only apply to 'Category 1 services'. The fact that the core free speech duties pursuant to clauses 12, 13 and 14 of the Bill only require platforms to 'have regard to' or, in the case of clauses 13 and 14, 'take into account', free speech rights or the protection of democratic or journalistic content, means that platforms may simply pay lip service to these 'softer' duties when a conflict arises with the legislation's numerous and 'harder-edged' 'safety duties'. This distinction between the harder and softer duties gives intermediaries a statutory footing to produce boiler plate policies that say they have 'had regard' to free speech or privacy, or 'taken into account' the protection of democratic or journalistic content. So long as they can point to a small number of decisions where moderators have had regard to, or taken these duties into account, they will be able to demonstrate their compliance with the duties to Ofcom. It will be extremely difficult, or perhaps even impossible to interrogate the process. Furthermore, as explained above, the Strasbourg Court is clear that although it is prepared to accept greater limitation of the scope of Article 10(1) in the context of online speech, this limitation must still fall within the parameters of Article 10(2). Arguably the requirement that clause 12 imposes on platforms to merely '*have regard to the importance*' of 'protecting users' right to freedom of expression within the law' does not go far enough to ensure the Bill complies with this jurisprudence.

Thus, by making online intermediaries responsible for the content on their platforms, the Bill requires them to act as our online social conscience, thereby

[105]cls 45(3) and 46(3).

making them de facto gatekeepers to the online world. Although 'privatised censorship' has taken place on platforms such as Facebook and Twitter since their creation, the Bill gives platforms a statutory basis for subjectively evaluating and censoring content. This, along with the potential conflict between the harder and softer duties, could lead to platforms adopting an over-cautious approach to monitoring content by removing anything that *may* be illegal (including content that they think *could* be hate speech) or *may* be harmful, and that would therefore bring them within the scope of the duty and regulatory sanctions. This risk is amplified by the lack of clear definitions of what is hate speech, and what is legal yet 'harmful' content, and who is a child or adult of 'ordinary sensibilities' and what 'characteristics' this includes. Such an approach could lead to legitimate content being removed because it is incorrectly thought to be illegal or harmful. And, cynically, it may provide platforms with an opportunity, or an excuse, to remove content that does not conform with their ideological values on the basis that it *could* be illegal or harmful.[106]

There is a further challenge to free speech to add to this Pandora's Box of confusion caused by the vagueness of the Bill. As stated above, clause 14(2) imposes a duty on Category 1 Services to take into account the importance of the free expression of journalistic content when making decisions about 'how to treat such content and whether to take action against a user generating, uploading or sharing such content'. Journalistic content is defined by the Bill as 'generated for the purposes of journalism' which is 'UK-linked'.[107] Thus, it does not need to have been generated by a recognised media organisation. In a media environment where citizen journalists are growing in numbers and are increasingly contributing to public discourse,[108] the fact that the Bill does not define citizen journalists is problematic. Without a clear definition it is unlikely that platforms, Ofcom, and the public will be able to consistently distinguish citizen journalism from other forms of expression by individuals. In the context of hate speech, the potential implications of this were identified by Twitter in evidence given to the House of Lords Communications and Digital Committee:

> '… there are accounts we have suspended for Hateful Conduct and other violations of our rules who have described themselves as 'journalists'. If the Government wishes for us to treat this content differently to other people and posts on Twitter, then we would ask the Government to define it, through the accountability of the Parliamentary process. Without doing so, it risks confusion not just for news publishers and for services like ours, but for the people using them.'[109]

[106]Coe (n 42) 85.
[107]cls 14(8) and (9).
[108]See generally: Coe (n 42).
[109]'Free for all? Freedom of expression in the digital age' (n 5) 44, para 190.

NetzDG 'obligations', the E-Commerce Directive and free speech

As explained above, unlike the Bill, the obligations imposed by NetzDG relate solely to procedural and organisational processes that must be adhered to by in-scope platforms. The purpose of these new obligations was to create more transparent reporting and complaint-handling processes.[110] Thus, section 2 requires platforms to 'produce and publish half-yearly German-language reports on the handling of complaints about unlawful content on their platforms'. Of arguably greater importance for platforms, section 3(1) requires them to 'maintain an effective and transparent procedure for handling complaints about unlawful content'. Section 3(2)(ii) says that content that is *manifestly unlawful* must be removed or blocked within twenty-four hours of receiving the complaint; however, this does 'not apply if the social network has reached agreement with the competent law enforcement authority on a longer period for deleting or blocking … [the] content.' Content that is (merely) *unlawful* must be removed or blocked 'immediately', which means within seven days of receiving the complaint, although this deadline can be extended if the platform needs to verify the facts or refer the decision to a 'self-regulation institution'. On the 30 March 2021 an amendment, which came into effect on 1 February 2022, was made to section 3 in the form of a new section 3a(2). The amendment requires platforms to report, in addition to removing or blocking, certain criminal expressions, including incitement to hatred, to the Federal Criminal Police Authority.[111]

There are some clear similarities between clauses 9(3) and 10(3)(a) of the Bill and section 3 NetzDG that have resulted in comparable arguments being made regarding NetzDG's compatibility with the E-Commerce Directive, despite adaptations to the original draft of the legislation to attempt to comply with the Directive's safe-harbour protections. These changes included the: (i) substitution of a strict one-week deadline for the removal of unlawful content with flexible deadlines, so as to comply with Article 14; (ii) removal of a requirement for platforms to 'take effective measure against new uploads of illegal content' due to its likely incompatibility with Article 15.[112] Although these amendments were made to try to ensure compatibility with the Directive, arguably, like clauses 9(3) and 10 (3)(a), they push the limits of the Directive too far.[113] Taking them in turn, firstly, the 'pre-structured' twenty-four-hour and seven-day deadlines prescribed by section 3(2) are clearly incompatible with the Article 14

[110]Wischmeyer (n 48) 39.
[111]*Bundeskriminalamt*. For analysis of this amendment see J Bayer, 'Germany: New law against right-wing extremism and hate crime' *Inforrm,* 24 April 2021.
[112]Wischmeyer (n 48) 40.
[113]ibid 43-46. T Hoeren, 'Netzwerkdurchsetzungsgesetz europarechtswidrig' *beck-community*, 30 March 2017.

flexible 'acts expeditiously' timeframe. Secondly, although the obligation placed on platforms to 'take effective measures against new uploads of illegal content' was not included in the final version of the Act, the complaint management system required by the legislation is only viable if platforms constantly and actively monitor all new content, which effectively violates Article 15.[114]

Thus, these requirements imposed by the legislation gave rise to two interrelated free speech fears that mirror the concerns regarding the Bill set out above. Namely, that the legislation would (i) lead to a privatisation of censorship, which in turn would (ii) incentivise platforms to over-censor contested but legal speech, thereby reducing, or even silencing, legitimate debate. In the final section, I will consider whether these fears have been realised and what this may portend for the Bill and free speech in the UK upon its enactment.

Lessons from Germany

In this article I have examined some of the implications for free speech that arise from the Bill. Chief among these is that it will create a regime that allows for the privatisation of censorship, in which the platforms become arbiters of free speech in the place of parliament or the courts, and which encourages the over-censorship, or 'over-blocking', of online speech.[115] At the time of its enactment NetzDG was the subject of similar concerns. Thus, in this section I briefly consider what clues the German experience may give us as to the longer-term impact of the Bill on free speech.

NetzDG was frequently described by a variety of actors as an 'invitation' to privatised censorship,[116] with a committee report of the *Bundesrat* articulating the fears that the legislation transfers the review of the legality of content from the state and courts to platforms; in doing so the government is avoiding its human rights obligations by imposing duties on private organisations to restrict expression: 'The review if the legality of content must not be delegated fully to the providers. In the view of the Bundesrat, section 3 of … [NetzDG] effectively transfers the review procedure to the private sector, which is contrary to the principles of the rule of law. The supervisory authorities or, as the case may be, the prosecution and, finally, the courts are responsible to authoritatively assess whether the law has been broken'. [117] Despite this fear, *at least technically* (although perhaps not *practically*), the Act does not give effect to such a

[114]ibid. Wischmeyer (n 48) 44.
[115]For example, see generally: 'Free for all? Freedom of expression in the digital age' (n 5) 36-43, para 151-183.
[116]For example, see: 'Germany: Flawed Social Media Law', *Human Rights Watch*, 14 February 2018; M Scott and J Delcker, 'Free speech vs. censorship in Germany' *Politico*, 4 January 2018.
[117]Bundesrat, *Empfehlungen der Ausschüsse* (Drs. 315/1/17, 23rd May 2017), 4 (tr. T.W.).

transfer. This is because, in all cases, upon notification, intermediaries must decide whether or not to remove the potentially illegal content which, ultimately, is a decision that can be subject to challenge in court. Thus, from a positivist perspective, 'the final say on the legality of the posting always remains with the courts'.[118] However, unfortunately, the position in the UK will be less clear than in Germany. Although in theory Ofcom has the final say as to whether the removal of particular content by a platform is a breach of a core free speech duty pursuant to clauses 12, 13 and 14, this duty only requires platforms to 'have regard to' or 'take into account' free speech rights or the protection of democratic or journalistic content. This gives platforms a statutory footing to produce boiler plate policies to that effect. As emphasised above in relation to the Bill, so long as the platform can point to decisions where moderators have had regard to or taken these duties into account, they will be able to demonstrate compliance. Consequently, Ofcom's role as a free speech-backstop is to a large extent a hollow one, as it will be extremely difficult, or perhaps even impossible, to meaningfully interrogate the process.

In relation to over-blocking of content, the arguments made in respect of NetzDG have been made about the Bill and other forms of online harms legislation. For instance, on the one hand, it has been recognised by commentators that NetzDG incentivises platforms to establish monitoring systems and processes that minimise their exposure to liability by 'deleting or blocking content in all cases in which the determination, whether or not the content is illegal, is more costly than potential losses the network might suffer from the exit of some users who take offence at over-blocking and who feel limited in their exercise of free speech'.[119] Yet, on the other hand, it is not inconceivable that platforms, in fact, shield their users from the impact of the legislation (and will do the same upon the Bill's enacted), to make them more attractive. In any event, although over-blocking is not a symptom of NetzDG (or the Bill, or any other form of online harms legislation), in that intermediaries have consistently used their terms of service to block legal content (and to refuse to delete illegal content) without any form of due process,[120] what we have with NetzDG and the Bill is the state, through the legislation, enabling platforms to dictate what to remove. Arguably, this is something altogether different and more concerning for free speech when one considers that, generally, Article 10 ECHR prohibits the state from interfering with freedom of expression.

[118]Wischmeyer (n 48) 49.
[119]ibid 51; A Lang, 'Netzwerkdurchsetzungsgesetz und Meinungsfreiheit' (2018) *Archiv des öffentlichen Rechts* 220, 227-228, 234-236. For this argument in respect of the Bill, see Coe (n 42) 84-85.
[120]Coe (n 42) ch 3.

Fortunately, we are not left to hypothesise about the practical effect of the Act, as section 2(1) NetzDG provides for a reporting mechanism which prescribes that '[p]roviders of social networks which receive more than 100 complaints per calendar year about unlawful content shall … produce half-yearly … reports on the handling of complaints about unlawful content on their platforms … and shall … publish these reports in the Federal Gazette and on their own website'. Although these reports neither confirm nor refute these inter-related concerns, they do reveal that NetzDG does not seem to have morphed platforms into unaccountable arbiters of the limits of free speech. Rather, it seems platforms block or delete far more content because it 'violates' their community standards.[121] Of course, whether the same pattern will apply in the UK once the Bill is enacted remains to be seen.

Conclusion

Although the German experience is not a 'crystal ball' it does provide some clues as to how the Bill, once enacted, may impact on free speech in the UK. However, this needs to be caveated with the fact that, as demonstrated throughout this article, the Bill goes further than NetzDG; its duties of care place more onerous obligations on platforms, the potential sanctions for breaching those duties are considerably more draconian and its in-built free speech protections are weaker. There are also a lot of unanswered questions about how the Bill will operate once in force because so much of the legal detail is currently un-drafted and will be subject to secondary legislation. Notwithstanding this uncertainty, for the reasons advanced in this article, there is reason for concern regarding the potential impact of the proposed framework on freedom of expression in the UK.

Acknowledgments

This article was presented as a paper at the British Association of Comparative Law's Annual Seminar 'The Regulation of Hate Speech Online and its Enforcement in a Comparative Perspective', Durham University, 31 August 2021. I am indebted to Dr Eliza Bechtold (University of Aberdeen), Dr Sophie Turenne (University of Cambridge) and Dr Oliver Butler (University of Nottingham) for their invaluable feedback on previous drafts of this article.

Disclosure statement

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

[121]Wischmeyer (n 48) 55.

Notes on contributor

Peter Coe is a Lecturer in Law at the School of Law, University of Reading, and an Associate Research Fellow at the Institute of Advanced Legal Studies and Information Law and Policy Centre, University of London. He is a member of the IMPRESS Code Committee, an independent member of the Council of Europe's Expert Committee on Strategic Lawsuits against Public Participation and he is currently serving as the UK's National Rapporteur on 'Freedom of Speech and the Regulation of Fake News' on behalf of the International Academy of Comparative Law and British Association of Comparative Law. He is also the Editor-in- Chief of Communications Law, and the Convenor of the Society of Legal Scholars' Media and Communications Law Subject Section.