Journal of Media Law

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

Platform regulation of hate speech – a transatlantic speech compromise?

Uta Kohl

To cite this article: Uta Kohl (2022): Platform regulation of hate speech – a transatlantic speech compromise?, Journal of Media Law, DOI: 10.1080/17577632.2022.2082520

To link to this article: https://doi.org/10.1080/17577632.2022.2082520

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

Published online: 02 Jun 2022.

Submit your article to this journal

Article views: 186

View related articles

View Crossmark data

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

JOURNAL OF MEDIA LAW
https://doi.org/10.1080/17577632.2022.2082520

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Platform regulation of hate speech – a transatlantic speech compromise?

Uta Kohl

Southampton Law School, University of Southampton, Southampton, UK

ABSTRACT

This paper argues that the binary opposition in the treatment of hate speech in the US and Europe hides non-binary preoccupations that reflect different *primary* fears which do not fall along the same 'scale'. European liberal democracies fear the consequences of hate speech being left uncensored in the public domain (a WHAT concern) whilst America fears the consequences of content interventions by government (a WHO concern). The paper then proposes that the German Network Enforcement Law of 2017 builds a bridge between American and European speech traditions. NetzDG requires major platforms to moderate content in response to user takedown notices based on legally imposed speech standards. The mechanism of *public standards being enforced through private processes* is arguably uniquely adept at simultaneously assuaging the *primary* European fear about the absence of effective speech controls in the public domain and the *primary* American fear about the presence of governmental censorship.

KEYWORDS Platform liability; NetzDG; First Amendment; Digital Services Act; hate speech

Introduction

Hate speech is not harmless.[1] Group libel – or the expressions of national, racial or religious contempt to incite hatred – is liable to result in discrimination, hostility and even violence,[2] and so has real-life consequences for

CONTACT Uta Kohl U.Kohl@soton.ac.uk

[1]Jeremy Waldron, *The Harm in Hate Speech* (Harvard University Press 2012); Ivan Hare and James Weinstein (eds), *Extreme Speech and Democracy* (OUP 2009); M Herz, P Molnar (eds), *The Content and Context of Hate Speech* (Cambridge University Press 2012); Eric Heinze, 'Viewpoint Absolutism and Hate Speech' (2006) 69(4) MLR 543; Eric Heinze, *Hate Speech and Democratic Citizenship* (OUP 2016); Eric Barendt, *Freedom of Expression* (OUP 2005); Law Commission, *Abusive and Offensive Online Communications : A Scoping Report* (Law Com No 381, 2018).

[2]Article 20 (2) of the International Covenant on Civil and Political Rights (1966); see also Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination (1965). On hate speech as group libel, see Waldron, ibid, 34ff.

© 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-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.

members of vulnerable minorities and for society at large.[3] On a more diffuse but perhaps even more important register, hate speech presents 'a calculated affront to the dignity of vulnerable members of society and a calculated assault on the public good of inclusiveness'.[4] The harm potentials of hate speech are – with variations in the detail – widely accepted, and not even principally disputed in the US where it is constitutionally protected speech.[5] In the words of Frederic Schauer: 'we must appreciate that freedom of expression protects the expression of information and ideas not because such expression is typically inconsequential or harmless, but *despite* the harm and consequences that expression may produce'.[6] First Amendment protection neither denies the harm of hate speech nor seeks to sanction or promote the ideas behind such speech; rather it seeks to guard the sanctity of the expressive act itself from governmental interference:[7] 'In the competition for influence, government remains on the sidelines, a neutral observer'.[8] The expressive act is sacred given its constitutive role for the individual – autonomy, self-disclosure and self-realisation – and for society, in particular for democratic self-government.[9] Other liberal democracies, most prominently European societies, agree that a commitment to freedom of expression is not about protecting niceties or – or as the European Court of Human Rights [ECtHR] put it – communications that are 'favourably received or regarded as inoffensive or as a matter of indifference' but rather about those 'that offend, shock or disturb the State or any sector of the population'.[10] Yet, for these liberal democracies, the self-realisation of the individual and democratic self-government would be undermined if hate speech in the public domain were to be left uncensored. So although America and other liberal democracies agree on the importance of freedom of expression and even broadly on the harmful consequences of

[3]At times captured through references to 'public order' or 'public peace' in the formulation of relevant offences, see e.g. s 130 of the German Criminal Code. In *King v Osborne* [1732] 94 Eng Rep 406, the group libel of Jews amounted to a breach of the 'public peace'; see also Chara Bakalis, 'Rethinking cyberhate laws' (2018) 27(1) Information & Communications Technology Law 86 (discussing different types of harm caused by online hate).

[4]Waldron (n 1) 3, 6 respectively.

[5]For example, Edward J Eberle, *Dignity and Liberty – Constitutional Visions in Germany and the United States* (Praeger 2002) 224; Robert Post, 'Hate Speech' in Hare and Weinstein (n 1) 123, 133ff, on the 'tendential' causal connection between the harms and the speech. The harms are often downplayed by U.S. authors e.g. Ronald Dworkin, *Foreword* in Hare and Weinstein (n 1): 'Many of the claims [of the malign consequences of bad speech] are inflated and some are absurd … we must protect it even if it does have bad consequences … '

[6]Frederick Schauer, 'The Exceptional First Amendment', Faculty Research Working Paper Series RWP05-21, Harvard University, John F. Kennedy School of Government (February 2005), 28 [emphasis added].

[7]Waldron (n 1) 38, discussing Catherine MacKinnon's concept of 'speech acts' to criticize the extension of the First Amendment to also protect intimidation, discrimination and subordination, see Catharine A MacKinnon, *Only Words* (Harvard University Press 1993).

[8]Eberle (n 5) 193.

[9]Barendt (n 1) 6ff; Frederick Schauer, *Free Speech: A Philosophical Enquiry* (CUP 1982).

[10]*Handyside v United Kingdom* App no 5493/72 (ECHR, 7 December 1976) [49]; *Observer and Guardian v United Kingdom* App no 13585/88 (ECHR, 26 November 1991) [59].

hate speech, they sharply disagree on whether it is beneficial for the State to intervene.[11] One might characterise this difference as one about *process* – what is the best avenue for neutralising hate speech – as opposed to one about *substance*, that is the nature of hate speech, its harmful consequences or the values protected by free speech more generally.

A reminder of the shared understandings, and of the nature of the disagreement, between American and European speech traditions provides a useful springboard for this paper which examines the clash of these speech cultures in the online world – and will do so through a 'process' lens. The Internet has changed speech concerns in two significant ways. *First,* it has led to an increase in the amount, visibility and permanence of unfiltered public speech, including hate speech, and thereby confronted European hate speech prohibitions with a problem of a different order of magnitude – one that demands a paradigmatic regulatory rethink. *Second,* the Internet has also 'globalised' speech and made the territorial silos within which divergent speech cultures could previously be enacted with relative coherence not impossible but much harder to realise.[12] This paper proposes that the pressure towards a new regulatory paradigm, on the one hand, and towards some reconciliation between the speech traditions, on the other hand, is unexpectedly bearing fruit. The German *Law to Improve Law Enforcement on Social Networks* (2017)[13] (hereafter 'NetzDG') is the first of a number of similar European initiatives,[14] that requires major platforms to moderate content in response to user takedown notices, and thereby delivers, it is argued, a compromise between American and European speech traditions. As all compromises, NetzDG has left both sides – US platforms and First Amendment advocates, on one side, and European advocates of speech restrictions to protect the rights of others, on the other side[15] – feeling compromised and not illegitimately so. Nevertheless, the argument here is that the mechanism which NetzDG adopts, namely a *public standard setting enforced via private processes*, is uniquely adept at simultaneously

[11]There are also those who argue that the difference is overblown: C Edwin Baker, 'Autonomy and Hate Speech' in Hare and Weinstein (n 1) 139, 140ff.

[12]See Martin Eifert, 'Rechenschaftspflichten für soziale Netzwerke und Suchmaschinen' (2017) Neue Juristische Wochenschrift 1450, 1454, commenting on the preference of the global tech players for adopting common global standards, thereby implicitly moving towards a convergence of national standards.

[13]*Gesetz zur Verbesserung der Rechtsdurchsetzung in Sozialen Netzwerken* (*Netzwerkdurchsetzungsgesetz* – 'NetzDG') (1 Sept 2017, BGBl I S 3352).

[14]Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM/2020/825 final; UK Draft Online Safety Bill (May 2021) CP405; Ireland's Online Safety and Media Regulation Bill (10 Jan 2020); France's *Loi* no 2021–1109 of 24 August 2021, Article 42; see also the decision of the French Constitutional Court invalidating a similar law known as the Avia Law, CC decision n° 2020–801 DC of 18 June 2020; European Commission, *EU Code of Conduct on countering illegal hate speech* (31 May 2016, together with four major IT companies: Facebook, Microsoft, Twitter and YouTube).

[15]This is a generalisation for the purposes of simplicity in this article; the author is fully aware that not all American commentators support the First Amendment (although most do), and not all European commentators are in favour of some speech prohibitions.

assuaging the *primary* European fear about the absence of effective speech controls in the public domain and the *primary* American fear about the presence of governmental censorship. Both fears are grounded in their respective conceptions of citizenry, which are *prima facie* as legitimate in their own right as they are different from each other.

Two preliminary points are in order. *First*, it may be queried why European regulatory initiatives aimed at removing hate speech from *national* online spaces, should at all be cognizant of American speech sensibilities. After all, customary international law on jurisdiction permits each State to apply its laws to activities that affect its territory even where they originate from the outside, and this has long been extended to online activities.[16] By implications, online hosts generally, and the large, predominately American platforms that target European users in particular, can legitimately be expected to comply with European obligations to moderate online content in respect of their localised services.[17] This is jurisdictionally uncontroversial. Still, a subject's hostility to the substance of a law cannot but translate into reluctant, even recalcitrant compliance. This is problematic for the effectiveness of any law, especially where the law is squarely grounded in private-public cooperation, as NetzDG and like laws are, and so requires more than the normal acceptance that render laws imposing obligations effective. Taking a social scientist's lens to the importance of legitimacy for obedience, David Beetham observed in *The Legitimation of Power*: [18]

> when legitimacy is eroded or absent … power does not necessarily collapse, or obedience cease, since it can continue to be kept in place by incentives and sanctions. However, coercion has to be much more extensive and omnipresent, and that is costly to maintain … Less dramatic, but equally important, is the effect a lack of legitimacy has on the *degree* of cooperation, and the *quality* of performance, that can be secured from them, and therefore on the ability of the powerful to achieve goals other than simply the maintenance of their position.'[19]

Thus, beyond any academic interest in the constitutional compatibilities of NetzDG's regime, legitimacy matters practically as it affects how American platforms *and* European users engage with the content moderation frameworks within which they are active participants. As a *second* and related point, the discussion does not suggest, nor depend upon the suggestion, that accommodating American speech sensibilities was a conscious design behind NetzDG, even if feasibility concerns may have brought issues

[16]Uta Kohl, *Jurisdiction and the Internet – Regulatory Competence over Online Activities* (CUP 2007); Julia Hörnle, *Internet Jurisdiction in Law and Practice* (OUP 2021).
[17]For a prominent example, see Article 3(2) of the GDPR.
[18]David Beetham, *The Legitimation of Power* (Humanities Press International Inc 1991).
[19]ibid, 28 [emphasis in the original]; see also Waldron (n 1) 20ff, commenting on early sedition laws as precisely designed to protect the legitimacy (or respectability) of the government from verbal assaults in order to maintain the cooperation with authority.

about its acceptability within the field of vision of the law makers.[20] Rather, the paper seeks to reposition the trend towards a *public standard setting enforced through private processes* as a restructuring of legal processes in response to a more complex, more heterogeneous and more global speechscape, which incidentally embeds the conflicting constitutional demands.

With this in mind, the paper locates the legitimacy of *each* speech perspective in their historically grounded construction of citizenry and the attendant primary fear to which each speech regime speaks. It then proposes that although NetzDG appears to feed both fears for different reasons, on closer inspection its central mechanism of *public (content and procedural fairness) standards embedded in private processes* creates a bridge of sorts between these historically grounded speech traditions and addresses their structural differences that go to the heart of their legitimacy concerns. Finally, it is argued that whilst this bridge is imperfect, it is *good enough* to meet the core concerns and objections from either side.

Divergent conceptions of citizenry and their primary fears

The American and European divergence on the legitimacy of hate speech regulation goes hand in hand with their particular conceptions of empowered citizenry,[21] with each being 'founded on deeply felt socio-political ideals, whose histories reach back to the revolutionary era of the later eighteenth century'.[22] In *The Two Western Cultures of Privacy: Dignity versus Liberty* (2004),[23] James Whitman historically traced *liberty* and *dignity* as the respective animating forces behind the divergent American and Continental European privacy sensibilities and conceptions. The same underlying values patterns also underlie, and help to explain, their different attitudes to hate speech.[24]

For Europeans, hate speech regulation is a matter of recognising the dignity of all members of society and thereby upholding their citizenry – not understood in the strict legal sense, but rather denoting an entitlement to being treated with basic respect based on one's humanity.[25] The uncensored presence of hate speech in the public domain is seen as a continuous attack on 'a shared sense of the basic elements of each person's status,

[20]For the role played by lobbying in communicating alternative values: Adam Satariano, Matina Stevis-Gridneff, 'Big Tech Turns Its Lobbyists Loose on Europe, Alarming Regulators' *The New Yok Times* (14 Dec 2020).

[21]The word is not used in a legal sense, as connoting voting and other rights reserved to 'citizens'.

[22]James Q Whitman, 'The Two Western Cultures of Privacy: Dignity versus Liberty' (2004) 113 Yale Law Journal 1151, 1219; James Q Whitman, Enforcing Civility: Three Societies (2000) 109 Yale Law Journal 1279.

[23]ibid.

[24]Whitman (2000) (n 22) on hate speech regulation.

[25]Waldron (n 1) 86f, where Waldron distinguishes between 'appraisal respect' (variable) and 'recognition respect' (invariable as it recognised the basic dignity of every person).

dignity, and reputation as a citizen or member of the society in good standing.'[26] Hate speech prohibitions must thus be understood not so much about protecting finer sensitivities, or cushioning the minority or the majority from offence but rather about stating and reinstating, on a continual basis, the community's commitment to inclusivity, diversity and equality as basic tenets of society. Historically, Whitman traced the Continental European preoccupation with dignity to the strictly hierarchical societies within which dignity and honour were the prerogative of the aristocracy and upper middle classes and embodied in numerous privileges.[27] By the same token, public insult and humiliation were (and remain) tools for disempowerment and exclusion. The pressure toward a '*levelling up* of dignity' occurred and found expression in the emerging European idea of citizenship in the late 18th and early nineteenth century.[28] The historian Ute Frevert observed, in *The Politics of Humiliation: A Modern History* (2020),[29] that 'lower-class people increasingly objected to disrespectful treatment … [and] used the language of honour and concepts of personal and social self-worth – previously monopolised by the nobility and upper-middle classes – to demand that they not be verbally and physically insulted by employers and overseers'.[30] This, in turn, went hand in hand with 'a new type of honour that followed the invention of 'citizens' (rather than subjects) in democratising societies. Citizens who carried political rights and duties were also seen as possessing civic honour … . [in contrast, traditional] social honour had been stratified according to status and rank.'[31] Unlike shaming effected within a social group, 'humiliation works by distinguishing radically between those who are in and those who are out: we are us, you are different and count for less'.[32]

Hate speech is designed to undermine the common humanity of a vulnerable minority, and thereby to silence and exclude its members. From a European perspective, it is the function of the State to protect minority groups and assure them of their standing as equal members of society.[33] This, in turn, requires removing affronts to their equal good standing, from the public domain, or, in Waldron's words, from the 'permanent visible fabric of society'.[34] This rationale points both to the core task hate speech

[26]Waldron (n 1) 47.
[27]Whitman (2004) (n 22) 1151f, where Whitman traces hate speech prohibitions back to dueling law: 'In the nineteenth century, continental courts protected the right to respect only of the dueling classes.'
[28]Whitman (2004) (n 22) 1151, 1166ff.
[29]Ute Frevert, *The Politics of Humiliation – A Modern History* (OUP 2020).
[30]Ute Frevert, 'The history of humiliation points to the future of human dignity' *Psyche Newletter,* 20 Jan 2021.
[31]ibid.
[32]Frevert (n 29) 13.
[33]Waldron (n 1), 46f, 98ff.
[34]Waldron (n 1), 3, 100, 45: '[w]hat attracts the attention of the criminal law] is the fact that something expressed becomes established as visible or tangible feature of the environment.'

prohibitions set themselves and the areas beyond their concern. Invariably, they focus either on *public* expressions[35] or on expression and other acts designed to disturb the *public* peace;[36] they are not concerned with what individuals think, or even say privately. What matters is what is 'broadcast' and would become etched into the public speech-scape. The visibility and permanence of online speech was stressed by the ECtHR in *Delfi AS v Estonia* (2015)[37] concerning an online intermediary's secondary liability for defamation, by observing that 'clearly unlawful speech, including hate speech and speech inciting violence, can be disseminated like never before, worldwide, in a matter of seconds, and sometimes remain *persistently* available online'.[38] Meanwhile, the ECtHR stressed the protective responsibility of the State in *Nix v Germany* (2018),[39] where it ruled on the compatibility of a criminal conviction for a blog containing hate symbols with Article 10. Reflecting on the danger of normalising hate speech in the absence of censorship, the ECtHR observed that 'States which have experienced the Nazi horrors may be regarded as having a special moral responsibility to distance themselves from the mass atrocities perpetrated by the Nazis' and banning hate symbols 'from all means of communication' is 'meant to pre-empt anyone becoming used [them] … '[40] Whitman and other have shown that the European preoccupation with equalised dignity which lies at the heart of its construction of citizenry goes back much further than World War II, but Nazi atrocities certainly gave those ideas a renewed intensity and urgency.[41]

Speaking in the context of privacy, Whitman observes that 'the primary enemy … according to this continental conception, is the media, which always threatens to broadcast unsavoury information about us in ways that endanger our public dignity. But … [a]ny other agent that gathers and disseminates information can also pose such dangers'.[42] Indeed, one can go further and say that from a European perspective, the identity of the 'primary enemy' is relatively unimportant; what matters is the presence or absence of the dignity-undermining speech in public. For sure, traditionally,

[35]Articles 184, 185, 186 of the German Criminal Code (humiliation and defamation).
[36]Section 130(1)2 of the German Criminal Code specifically frames hate speech as a disturbance of the public peace through violating 'the human dignity of others by insulting, maliciously maligning or defaming one of the aforementioned groups … on account of their belonging to … [such] groups … ', in contrast to civil actions
[37]*Delfi AS v Estonia* App no 64569/09 (ECHR, 16 June 2015) [140–162]; see also *MTE-Index.hu v Hungary* App no 22947/13 (ECHR, 2 February *2016*).
[38]*Delfi AS* (n 37) [110] [emphasis added], see also [147].
[39]*Nix v Germany* App no 35285/16 (ECHR, 05 April 2018)
[40]ibid [47] and [54] respectively.
[41]Whitman (2004) (n 22) 1166, controversially argued 'much of this levelling up took place *during* the fascist period' [emphasis in the original] but clearly this was a very limited levelling up amongst 'Aryan' Germans and essentially retained the idea that some had a right to dignity, whilst others did not.
[42]Whitman (2004) (n 22) 1161.

media companies could throw the widest broadcasting net and thereby create the greatest harm potential, but that was a factual rather than a normative reason for targeting them with dignity-driven laws. Whitman's search for a 'primary enemy' betrays a peculiarly American constitutional outlook, where the involvement of government irrevocably taints an act that would otherwise be considered uncontroversial or indeed worthy. If a social media company removes hate speech of its own volition, as many of the major ones have done for some times, e.g. under the 'EU Code of Conduct on countering illegal hate speech online' (2016),[43] this is entirely within its rights and, in terms of public policy, often desirable. Twitter was applauded when it permanently suspended Donald Trump's account in early 2021, following various other platforms that limited his activities.[44] If, however, that removal had been demanded by government, it would have been considered an intolerable threat to the freedom of citizens and would have been struck down as unconstitutional.

The American preoccupation with the State as the primary threat to free speech is legally anchored in the First Amendment which shields speech from governmental interventions in absolutist terms, that is without exceptions[45] - underwritten by the value of liberty, or freedom *from* the State. Although the First Amendment (1791) only emerged in its present absolutist interpretation in the second half of the twentieth century with cases like *Brandenburg v Ohio* (1969)[46] and *New York Times v Sullivan* (1964),[47] it is American settler history and the national mythology surrounding it that these cases reflect, celebrate and reinforce. First Amendment jurisprudence is deeply inscribed with settler mentality, or the ideal of hardiness, self-sufficiency and self-reliance and, by implication, of self-rule and distrust of government, as encountered by the settlers in the old world left behind, and then belatedly in the new world through the colonial administrations.

[43] European Commission (n 14), the Code was agreed between the European Commission and Facebook, Microsoft, Twitter and YouTube, followed subsequently by Instagram, Snapchat, Dailymotion, Jeuxvideo.com, TikTok and LinkedIn.

[44] Kate Conger, Mike Isaac, 'Twitter Permanently Bans Trump, Capping Online Revolt' *New York Times*, 8 Jan 2021.

[45] Schauer (n 6) 20 where he refers to the First Amendment as an 'imbalanced text'; see also Hugo L Black, 'The Bill of Rights' (1960) 35 New York University Law Review 865, 874. Note also, Michael Kang and Jacob Eisler, 'Trump's Challenge to Constitutional State Speech' (2022) University of Illinois Law Review (forthcoming) commenting on JS Mill's view on liberty of speech as not just entailing non-interference by the State but also as ensuring 'a substantive opportunity to … practice intelligent self-mastery … [free from the] despotism of opinion when society is itself the tyrant … ' [internal marks omitted].

[46] *Brandenburg v Ohio* 395 U.S. 444 (1969); see also *Cantwell v Connecticut* 310 U.S. 296 (1940). For a context of First Amendment developments being directed against attempts of the Southern states to deploy McCarthy-era indirect speech restrictions against civil rights organisations, see Seth Kreimer, 'Censorship by Proxy: the First Amendment, Internet Intermediaries, and the Problem of the Weakest Link' (2006) 155 University of Pennsylvania Law Review 11; Vincent Blasi, 'The Pathological Perspective and the First Amendment' (1985) 85 Columbia Law Review 449.

[47] *New York Times Co v Sullivan* 376 U.S. 254 (1964).

Justice Brandeis in *Whitney v California* (1927)[48] – although upholding a conviction for communist speech advocating the overthrow of the government – famously laid the foundation for the absolutist interpretation of the First Amendment, by grounding it in the self-reliant settler who does not need or want government to protect him from bad speech; quite the reverse, protection from government is what is needed:

> 'Those who won our independence by revolution were not cowards. They did not fear political change. They did not exalt order at the cost of liberty. To courageous, self-reliant men, with confidence in the power of free and fearless reasoning applied through the processes of popular government, no danger flowing from speech can be deemed clear and present unless the incidence of the evil apprehended is so imminent that it may befall before there is opportunity for full discussion. If there be time to expose through discussion the falsehood and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence. Only an emergency can justify repression. Such must be the rule if authority is to be reconciled with freedom.'[49]

Within this settler construction of speech entitlements, the European preoccupation with *levelling up* of dignity is not particularly meaningful. For settlers, the universal immigration status and common hardship in a hostile environment acted as a powerful equaliser that would have made any demand for formal equal respect at best irrelevant and at worse misplaced. Given America's view of itself as egalitarian in its very roots and inception, it has in fact nurtured a culture of disrespect which Whitman describes as America's *levelling down*: 'The American refusal to show respect to anybody is not just, as it were, a social refusal; it belongs to our sense of the political constitution of our form of egalitarian society … '[50] He continues: 'our free speech, to adopt a term from ancient cynic philosophy, tends to express itself as "parrhesia" – as speech that is not just about the sober expression of opinions, but also about the free and aggressive display of disrespect'.[51]

Yet, much as the First Amendment creates no room for prohibiting 'aggressive displays of disrespect' such as hate speech, it does not protect such disrespectful speech from being removed by *private* actors. The right of private gatekeepers to edit as they please is protected by the First Amendment, and editorial oversight is often implicitly encouraged by the law.[52]

[48] *Whitney v California* 274 U.S. 357 (1927).
[49] *Whitney v California* 274 U.S. 357 (1927), 377.
[50] Whitman (2000) (n 22) 1397, more generally see 1384ff. Clearly, this is more myth than reality in light of slavery and the treatment of Native Americans.
[51] ibid.
[52] S 230(c)(2) of the Communication Decency Act: 'No provider or user of an interactive computer service shall be held liable on account of … any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected … '

In other words, the American speech regime is not concerned with the presence or absence of particular kinds of speech in the public domain; it simply takes the State out of the equation in dealing with speech based on a 'skepticism about the ability of any governmental institution reliably to distinguish the good from the bad, the true form the false, and the sound from the unsound … '[53] So its primary concern lies in guarding against government interventions, particularly where it would involve government taking sides on what is good or bad speech:[54] 'The leitmotif of contemporary American free speech doctrine is its intensive hostility to the content-based regulation of public discourse.'[55] The least tolerated content regulation, in turn, is one that engages in viewpoint discrimination, that is a speech prohibition not just based on the content of the message but also on its ideological perspective: 'the American understanding is that principles of freedom of speech do not permit government to distinguish protected from unprotected speech on the basis of the point of view espoused'.[56] The government must, by and large,[57] stay on the sidelines of public discourse, which entails forbearance even where it considers the speech bad. In the words of Justice Scalia in *RAV v City of St Paul* (1992):[58] 'Let there be no mistake about our belief that burning a cross in someone's front yard is reprehensible. But St. Paul has sufficient means at its disposal to prevent such behaviour without adding the First Amendment to the fire'.[59] The State had other avenues for dealing with the behaviour, e.g. arson or trespass, without the need for passing judgment on its communicative content. Justice Harlan in *Cohen v California* (1971)[60] rested the positive case for the First Amendment absolutism, on the one hand, on it being a 'powerful medicine in a society as diverse and populous as ours'[61] and, on the other hand, on the ideals of capable citizenry and individual dignity:

> '[Free speech] is designed and intended to remove governmental restraints from the arena of public discussion, putting the decision as to what views shall be voiced largely into the hands of each of us, in the hope that use of such freedom will ultimately produce a more capable citizenry and more perfect polity and in the belief that no other approach would comport with

[53] Schauer (n 6) 24.
[54] Subject to exceptions: (1) narrow speech categories where governmental intervention is permitted e.g., obscenity, child pornography, fighting words and incitement to imminent lawless action, and (2) content-based regulation is constitutional where it survives the strict scrutiny test.
[55] James Weinstein, 'An Overview of American Free Speech Doctrine and its Application to Extreme Speech' in Hare and Weinstein (n 1) 81.
[56] Schauer (n 6) 9f.
[57] Unless the prohibition survives strict scrutiny, see e.g. *Burson v Freeman* 504 U.S. 191 (1992).
[58] *RAV v City of St Paul* 505 U.S. 377 (1992).
[59] *RAV v City of St Paul* 505 U.S. 377 (1992), 396.
[60] *Cohen v California* 403 U.S. 15 (1971).
[61] *Cohen v California* 403 U.S. 15 (1971), 24.

> the premise of individual dignity and choice upon which our political system rest'.[62]

So, European and American speech jurisprudence equally appeal to diversity and multiculturalism, and citizenry and human dignity as driving forces behind their divergent approaches to speech regulation. The difference in their approaches, however, results in the endorsement or rejection of hate speech regulation and thereby suggests a binary opposition. Yet, this binary hides non-binary preoccupations that reflect qualitatively different *primary* fears – fears that do not fall along the same scale. European liberal democracies fear the consequences of hate speech being left in the public domain (a WHAT concern), whilst America fears the consequences of content interventions by the government (a WHO concern). For sure, these fears lead to a direct confrontation when *government* is treated as the only entity equipped to deal with hate speech. Yet, this confrontation cannot detract from the fact that their essential preoccupations are not of the same kind and may be accommodated under one roof. This, then, brings the discussion to the legal framework requiring platform to remove hate speech in response to user takedown notices, as illustrated by NetzDG, the first law that formalises the *process* for platform takedown obligations.

A compromise solution? Public standards and private processes

Germany's NetzDG

Germany's NetzDG came into force in 2017[63] and requires social media platforms with more than two million German users[64] to remove hate speech and other illegal content in response to user complaints. The time span for doing so is 24 hours for 'manifestly illegal' material, and seven days for all other illegal material, with the possibility of further extensions.[65] The illegality or legality is reviewed by the platforms in light of *existing* content

[62] ibid 24.

[63] *Gesetz zur Verbesserung der Rechtsdurchsetzung in Sozialen Netzwerken* (n13) ; amended NetzDG available at https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_node.html; discussed in Eifert (n 12); William Echikson and Olivia Knodt, 'Germany's NextDG: A key test for combatting online hate' *CEPS Policy Insight* No 2018/09 (22 November 2018); Thomas Wischmeyer, 'What is illegal offline is also illegal online': The German Network Enforcement Act 2017' in Bilyana Petkova, Tuomas Ojanen (eds), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries* (Edward Elgar 2020) 28; Heidi Tworek and Paddy Leerssen, 'An Analysis of Germany's NetzDG Law' Working Paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, 15 April 2019, available at www.ivir.nl/twg/; Amelie Heldt, 'Reading between the lines and the numbers : an analysis of the first NetzDG reports' (2019) 8(2) Internet Policy Review; Andrej Lang, 'Netzwerkdurchsetzungsgesetz und Meinungsfreiheit' (2018) Archiv des Öffentlichen Rechts 220.

[64] S 1(2).

[65] S 3(2)2 and 3(2)3.

prohibitions in the German Criminal Code, including section 130 on hate speech.[66] NetzDG shifts censorship (here not used in a technical legal meaning, but broadly as the imposition of restrictions on speech by private or public actors),[67] traditionally adjudicated by the State, to the large social platforms. The platforms already occupy active gatekeeper roles and thereby shape the public sphere, albeit according to their own 'community standards' based on their profit-driven interests.[68] In this sense, NetzDG may be read as adding publicly or legally agreed concerns to the private or corporate list of unacceptable activities, conducts and expressions subject to takedown. In effect, it supplements US-centric global platform community standards imposed by the technology corporations, with democratically legitimated community standards of the German polity.[69]

Platforms can be fined of up to €50m[70] for non-compliance with their obligations under sections 2–5 of the NetzDG, as strengthened by amendments in 2021.[71] These obligations are: to publish transparency reports; to put in place adequate processes for allowing and dealing with user notices; for the appeal of content decisions and for arbitrating disputes between users making complaints and those whose content is to be taken down; obligations to keep these processes under review and remedy organisational shortcomings; an obligation to train and supervisor content reviewers; and finally to designate a contact person in Germany who must respond to information requests.[72] What is noteworthy about these obligations is, *first*, that they are reactive to user engagement, i.e. user takedown notices or complaints about unjustified takedown notices, and do not expect the platform itself to activate any actual case or monitor its site, as would be contrary to Article 15 of the E-Commerce Directive.[73] *Second*, the obligations are all process-focused: the fines do not attach to individual 'wrong' content

[66]NetzDG lists 22 statutory offences in s 1(1): ss 86, 86a, 89a, 91, 100a, 111, 126, 129 bis 129b, 130, 131, 140, 166, 184b, 185 bis 187, 201a, 241 and 269. There have been criticisms that the catalogue of offences is too long.

[67]For a perspective on online censorship and its meaning, see *András Koltay*,'The Private Censorship of Internet Gatekeepers' (2021) 59 University of Louisville Law Review 255, 264ff.

[68]For example, 'Facebook Community Standards', YouTube's 'Community Guidelines' and 'The Twitter Rules'.

[69]NetzDG enjoyed popular support in Germany with a 87% approval rating and only 5% disapproval rate in 2018, see Tworek (n 63), 2.

[70]S 4(2) of NetzDG in conjunction with ss 30, 130 of the Act on Infringements ('*Ordnungswidrigkeitengesetz*').

[71]*Gesetz zur Änderung des Netzwerkdurchsetzungsgesetzes* (3 June 2021) BGBl. I S441; for the consolidated version in German, see http://www.gesetze-im-internet.de/netzdg/BJNR335210017.html). The amendments seek to improve, inter alia, the user-friendliness of the complaint procedure (s 3(1)), provides a possibility of appeals procedure and dispute resolution (s 3b,3c), enhances the information on the transparency reports (s 2(1, (2)) and increases the powers of the supervise authority (s 4(a)).

[72]Section 4(1) 1–8 respectively.

[73]For the same reason, an obligation to take 'take effective measures against new uploads of illegal content' was taken out of the original draft, see Wischmeyer n 63, 9; see also 13: social networks already use technology to pre-emptively scan and interpret all content for reasons of fraud detection and risk management, but also to identify relevant content and, not least, to generate ad revenue.

decisions made by platforms but rather to *systemic* deficiencies in the processes.[74] This makes sense considering that more close-up oversight of individual decisions would largely defeat the purpose of the law of delegating the responsibility of making the content decisions. The law creates a public framework for private censorship, regulated self-regulation,[75] in order to address systemic risks. This constitutes a shift from earlier regimes such as the (conditional) intermediary immunity framework under Article 14 of the Electronic Commerce Directive[76] premised on the potential platform's secondary liability for illegal publications and thus inevitably entails the threat of a penalty for wrong review decisions (i.e. the lack of immunity). In effect, NetzDG metamorphoses platforms from *regulatees* subject to conditional liability for illegal content to *regulators* through their obligations to put processes in place to assess and suppress illegal content with (all) the due process trappings of a regulator. Thus hate speech in the public domain is taken down through private processes.

Yet, NetzDG has attracted strong criticisms from both sides of the transatlantic speech divide, that is from proponents and opponents of hate speech restrictions.[77] From a European perspective, the law creates undue privatised censorship, whilst, from an American perspective, it presents undue public or governmental censorship. Both perspectives are legitimate against each peculiar speech tradition, but in their contradiction also lies their weakness. Precisely in so far as the European concern is that NetzDG privatises censorship that ought to be in the hands of the judiciary, it weakens the American claim of undue governmental censorship; equally in so far as through an American lens NetzDG imposes undue public censorship, the European claim of privatisation of censorship does not fully hold water. In fact, NetzDG hovers in between private and public censorship: it puts in place a private process ('who') to uphold publicly defined speech standards ('what'), and thereby goes towards appeasing the primary fears of European and American speech traditions, as now discussed in more detail.

Undue Private Censorship: 'privatised processes'

The main objections against NetzDG from a European perspective are, *first,* that is privatises law enforcement, notably the judicial function, and so the

[74]For the latest version: Bundesamt für Justiz, '*Leitlinien zur Festsetzung von Geldbußen im Bereich des Netzwerkdurchsetzungsgesetzes* (NetzDG)' (11 June 2018), available at https://www.bundesjustizamt.de/DE/SharedDocs/Publikationen/NetzDG/Leitlinien_Geldbussen_de.pdf

[75]On the different modes of self-regulation, see Christopher T. Marsden, *Internet Co-Regulation: European Law, Regulatory Governance, and Legitimacy in Cyberspace* (CUP 2011) 54.

[76]Note, however, that NetzDG does not affect the application of Article 12–15 of the Electronic Commerce Directive 2000/31/EC.

[77]'European' and 'American' will be used as a convenient shorthand for the proponents and opponents of hate speech prohibitions

State abdicates its responsibility, and, *second,* that the private actors burdened with the law enforcement are liable to overblock to avoid liability and thereby unduly interfere with freedom of expression.[78] The latter objection tends to be treated as a practical instantiation of the former: unaccountable private actors cannot be trusted to apply the law neutrally and objectively in disregard of their own interest. The overblocking objection is a familiar one from earlier notice-and-takedown regimes: it is based on the idea that a law with review obligations on private actors backed with penalties creates incentive structures for private actors to err on the side of caution in taking down content in order to avoid liability or fines, and especially so when under time pressure.[79] Such law is bound to lead to the suppression of legally permissible communications.[80] Thus, the effort to protect the dignity of vulnerable members of society has the effect of violating the speech rights of others. The concern is in principle legitimate but appears not to have materialised in respect of other notice-and-takedown regimes, as intermediaries have strong countervailing interests in allowing user content.[81] It is even less likely to materialise under NetzDG given that fines do not attach to individual decisions but to failures in the processes.[82] Thus individual wrong decisions are not caught by NetzDG, and only *systemic* underblocking would raise questions about the adequacy of the platform's processes.[83] Systemic overblocking is in any event not caught by NetzDG as it does not prevent the platform's own content moderation policies to limit speech over and above the legal requirements (see below, Section 3.3).

Still, the first objection that private platforms ought not to be the final arbiters of compliance with legal standards remains intact quite regardless of whether or not NetzDG incentivises overblocking, given that this objection is only partly anchored in the untrustworthiness of private actors as legal arbiters as an empirical matter. Mainly, the objection against privatised censorship strikes a constitutional chord: the rule of law, or equality before the law, is constituted by laws being adjudicated by the judiciary; it is the function of independent and transparent courts to be the final and authoritative adjudicator on whether a law has been broken and whether sanctions should be applied.[84] Through NetzDG, the State appears to abdicate that responsibility. Yet, *formally* this is not the case given that the judiciary

[78]Echikson (n 63) 7; see also Molly K Land, 'Against Privatized Censorship: Proposals for Responsible Delegation' (2020) 60 Virginia Journal of International Law 363.
[79]See also the decision of the French Constitutional Court: CC decision n° 2020–801 DC of 18 June 2020, stating that the notice-and-takedown measures violated freedom of expression, particularly considering the very short timeframes within which platforms had to respond to notices.
[80]A valid point has been made that private companies already block much illegal content.
[81]Wischmeyer (n 63).
[82]Wischmeyer (n 63) 20.
[83]Such as the obligation to use independent and competent content reviewers.
[84]Wischmeyer (n 63) 15.

retains the right to make *final* assessments on content decisions by platforms. Through the continuing application of existing substantive legal regimes outside of NetzDG, including the possibility of accessory liability of platforms, takedown decisions by platforms can be challenged before courts as has occurred on occasions.[85] NetzDG itself also provides in section 4(5) that if the oversight administrative authority wants to challenge a platform decision not to remove content, as part and parcel of making a case of systemic deficiencies, it needs a preliminary ruling by a court to confirm the unlawfulness of the content. So, *formally*, the State through its judiciary remains the final arbiter of content review decisions, and thus provides an assurance of ultimate oversight.[86]

Still, despite the final formal authority of the State, the *raison d'être* of NetzDG *is* indeed to heighten the legitimacy of the private review processes which platforms have had in place for some time, and thereby their relative self-sufficiency, and thus to minimise the expected *substantive* role of the State to deal with problematic online content. Given the sheer quantity of online content, discharging that substantive role would otherwise require the State's heavy investment in 'its own infrastructure … to improve the enforcement of its laws – by strengthening law enforcement authority, prosecutors and court … '[87] Content assessment and takedown is thus 'transferred' to platforms as active in situ gatekeepers to act as a kind of first instance tribunals – albeit on the assumption that their processes will be conclusive in most cases, given the cost of litigation vis-à-vis the relatively low-value of most complaints.[88] The term 'transfer' misrepresents the situation in so far as the State had not previously discharged a comparable takedown role, other than in the isolated cases brought before courts. In any event, such 'transfer' of responsibility is neither in itself novel[89] nor constitutionally problematic under German or EU law.[90] Still, by imposing minimum transparency and procedural fairness standards on the review process, NetzDG fundamentally breaks with existing intermediary takedown regimes,[91] which had simply

[85]Eifert (n 12) 1451. See also s 3b (4) which states that the judicial avenue to challenge a decision by a platform is unaffected by the Act.

[86]For a conception of contract law as a (convenient) regulatory device: Aditi Bagchi, 'Interpreting Contract in a Regulatory State' (2020) 54 University of San Francisco Law Review 35.

[87]Wischemeyer (n 63) 17.

[88]Eifert (n 12) 1451.

[89]All secondary liability based on notice of the wrongdoing incentivises the 'publisher' to make a judgment about the presence of the wrongdoing by the primary author and to act upon it, as is common in defamation and IP law. The 'right to be forgotten' in C–131/12 *Google Spain SL and Google Inc v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* [2014] EU:C:2014:317, imposed adjudicative and takedown responsibility on data controllers such as Google (see now Article 17 of the GDPR 2016/679).

[90]Most recently, see BVerfG, *Stadiumban* (11 April 2018), BVerfGE 148, 267, on the limitations of the horizontal effect of basic rights. See also Eifert (n 12) 1451.

[91]Most prominently Article 14 of the Electronic Commerce Directive 2000/31/EC but see also s 1 (1) of the Defamation Act 1996 (UK) or Article 8(3) of the Copyright Directive 2001/29/EC.

outsourced decisions on the legality of communications without paying attention how those decisions would be made. NetzDG shifts the focus *from* the validity of the decisions *to* the propriety of the decision-making process: 'the existing liability regime was only interested in the output – what mattered was that illegal content was removed, how providers managed to do so, was their business … [NetzDG] introduces new organizational obligations for intermediaries … '[92] It requires platforms to 'supply users with an easily recognisable, directly accessible and permanently available procedure for submitting complaints about unlawful content' and 'maintain an effective and transparent procedure for handling [these] complaints.'[93] For actual content decisions, platforms may give 'the user an opportunity to respond [on factual matters] to the complaint before the decision is rendered' and can delegate difficult decisions to 'a recognised self-regulatory institution.'[94] Once a decision is made, the provider must 'immediately notify the person submitting the complaint and the user about any decision, while also providing them with reasons for its decision.'[95] In 2021, NetzDG was amended to strengthened the 'access to justice' rights of complainants and, even more so, of the users threatened with removal of their content. Now both parties can initiate a review of the platform decision to remove or not to remove the material,[96] or settle their dispute through arbitration.[97] These procedural rights are designed to guarantee a fair process for adjudicating between the private interests in the relevant content *and* between the public interest in freedom of expression and competing rights. They do so by requiring an adversarial design of the decision-making process and through the (relative) transparency of decisions and their bases.[98] Throughout this private adjudication and takedown process, public authority remains a background presence by having to approve the self-regulatory institution to which a platform *may* defer difficult cases,[99] the arbitration body to which disputes may be submitted,[100] and by supplying the supervisory authority to which the local platform representative is answerable.[101] In short, the State delivers a basic framework to ensure, and be seen to ensure, the integrity of

[92]Wischmeyer (n 63) 7.
[93]S 3(1).
[94]S 3(2) 3.b and 3(6)-(11)
[95]S 3(2)5.
[96]S 3b: right to ask for a review of the decision within two weeks of the decision having been made.
[97]S 3c.
[98]For the general transparency requirements through biannual reports, see s 2; the information requirements are either for information of a general nature (s 2(2)2: 'the criteria applied in deciding whether to delete or block unlawful content') or for aggregate information and thus not instructive about individual decisions; see discussion accompanying n 139. Affected parties, however, must be kept in the loop: s 3(5).
[99]S 3(2)3.b and s 3(6)-(7).
[100]S 3c.
[101]S 4a.

the private review processes.[102] This expectation of due process rights also extends, at least partially, to content moderation under a platform's *own* Community Standards, as confirmed by the German Federal Court of Justice in 2021 when it held that Facebook acted illegally when it took down racist posts without informing the user retrospectively, and when it failed to inform the user of its intention to block his or her account without giving reasons and without giving the user an opportunity to respond to those reasons.[103] Of course, in as much as the State becomes the guardian of the process, it also removes itself from the actual substantive adjudication; it moves to the sidelines.

In light of the above, it may be argued that far from privatising censorship, NetzDG injects significant public dimensions into what would otherwise be entirely unaccountable private content moderation by platforms: first, it spells out public content standards to be observed by platforms (more on this below, see Section 3.3); and second, it imposes minimum due process and transparency requirements for the adjudication of complaints based on these standards. Public authority does not thereby abdicate its final formal authority, nor does it substantively surrender a function it previously discharged (the amount of low-value speech in the public domain now is unprecedented).[104] Equally, whilst the German Criminal Code supplies a legal bar by which content complaints to platforms must now be assessed, the role performed by platforms under NetzDG neither displaces criminal prosecutions nor civil actions;[105] it is simply designed to effect the removal of hateful and other illegal material from the 'permanent visible fabric of society'.[106] From a European hate speech perspective, such removal lies at the very heart of protecting the inherent dignity of individuals, particularly from minorities, and their equal standing in the community.

[102]Waldron (n 1) 80f, draws on Emily Durkheim when arguing that 'penal laws have an important expressive as well as a coercive function; and one would expect that expressive function to be at the fore … particularly in connection with the public and visible assurance of just treatment … to all of its members.'

[103]BGH, *Hate Speech Posts Deletion and Account Closure* (29 July 2021), III ZR 179/20, III ZR 192/20. The Court reached its decision by balancing the fundamental rights of the platform to 'occupational freedom' (Article 12 German Constitution) with that of the user to freedom of expression, and it required the platform, as a proportional response to its content moderation, to include minimum due process rights into its Terms and Conditions

[104]Jacob Rowbottom, 'To Rant, Vent and Converse: Protecting Low Level Digital Speech' (2012) 71 Cambridge Law Journal 355; *Stocker v Stocker* [2019] UKSC 17; and Koltay (n 67) 301, discussing this 'unknown model of co-regulation that has never existed before.'

[105]Under s 3a(4) the network has a duty to report illegal speech to the Federal Crime Bureau.

[106]See above (n 34).

Undue Public Censorship: 'official speech standards'

NetzDG would have been struck down as unconstitutional under the First Amendment if passed in the US.[107] Through an American constitutional lens, it is almost indistinguishable from the law under scrutiny in *RAV v City of St Paul* (1992),[108] which was St Paul's Bias-Motivated Crime Ordinance that prohibited the display of symbols or objects 'which one knows or has reasonable grounds to know arouses anger, alarm or resentment in others on the basis of race, color, creed, religion or gender … ' Even assuming, as the Supreme Court did, that the Ordinance should be construed to only reach 'fighting words' (i.e. those that tend to incite *immediate* breaches of the peace[109]) which fall outside First Amendment protection, the Ordinance still fell foul of the First Amendment. On its face, it discriminated on the basis of the content of the fighting words by only penalising fighting words based on race, colour etc. and not others. This furthermore amounted to viewpoint-based discrimination given that the prohibition sought to supress particular selected ideas. So the distaste American constitutional law displays towards content-based and, even more so, viewpoint-based discrimination by government is so strong that it reaches statutes prohibiting speech which is not even considered worthy of protection. To pass the constitutional hurdle, there must be 'no realistic possibility that *official* suppression of ideas is afoot'.[110] In this respect, First Amendment jurisprudence extends the strict scrutiny analysis applicable to political speech even to low-valued, unprotected categories of speech, such as fighting words.[111] Against this background, section 130(1) of the German Criminal Code on hate speech clearly and purposefully seeks to supress certain disfavoured ideas, much like St Paul's Bias-Motivated Crime Ordinance:

> Whoever, in a manner which is suitable for causing a disturbance of the public peace, 1. incites hatred against a national, racial, religious group or a group defined by their ethnic origin, against sections of the population or individuals on account of their belonging to one of the aforementioned groups or sections of the population, or calls for violent or arbitrary measures against them … incurs a penalty of imprisonment for a term of between three months and five years.

Accepting NetzDG's incompatibility with the First Amendment,[112] the argument here is that its enforcement regime, whereby *official* speech standards

[107]But see Patrick Zurth, 'The German NetzDG as Role Model or Cautionary Tales? Implications for the Debate on Social Media Liability' (2021) 31 Fordham Intellectual Property, Media & Entertainment Law Journal 1084.

[108]*RAV v City of St Paul* 505 U.S. 377 (1992). A difference is that the standards in NetzDG are not in fact enforced by the State.

[109]*Chaplinsky v New Hampshire* 315 U.S. 568 (1942), 572.

[110]*RAV v City of St Paul* 505 U.S. 377 (1992), 390 [emphasis added].

[111]Edward J Eberle, 'The Architecture of First Amend Free Speech' [2011] Michigan State Law Review 1191, 1193.

are enforced by private platforms through their largely autonomous private processes, has a significant privatising 'laundering' effect. Through this private processing, the public speech standards are absorbed, transformed, and remodelled in line with the view of the platform and thereby substantially distanced from government. This privatisation effect operates on two levels.

First, on an operational level, platforms subject to NetzDG have structured their review processes such that all complaints are assessed, first, under their terms of service and private community standards and only content *not* offending those standards is then assessed against the requirements of NetzDG. Analysing the first two biannual reports from various platforms, Tworek et al comment:

> Google, Facebook, and Twitter all prioritize compliance checks with their community guidelines; with each complaint, they first consider whether it violates their community standards. Any content that fails this check is removed … Accordingly, as Google's transparency report shows, a majority of removal decisions are based on the platform's private standards, and not on German speech laws. Facebook and Twitter do not specify this data in their reports, but they do review complaints in the same order, prioritizing community guidelines.[113]

In other words, the platforms largely absorb the public content standards within their generally more restrictive private house rules and so 'privatise' significant chunks of the German speech laws, which likely also already occurred at earlier stages when the house rules were formulated.[114]

This review hierarchy is not accidental but facilitated and sanctioned by the status of platforms as *private* actors who are at liberty to decide on their own content standards. Under NetzDG, illegal content must be taken down, but content legal under German law need not be kept up. This issue of whether platforms should be permitted to take down legal content was highly contested in the lead-up to NetzDG, with some commentators advocating that dominant platforms should not just be under a positive duty to take down illegal content but also – much like state actors – under a negative one *not* to censor *legal* content. This, in turn, would have removed the possibility of private community standards.[115] Yet, making platforms *directly* bound by fundamental rights comparable to state actors, based on their dominant role in the communication sphere, was rejected

[112]Jack Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 University of California Davis Law Review 1149, arguing, on the basis of U.S. constitutional law, against the participation of intermediaries in online speech regulation.

[113]Tworek (n 63) 5; see also Heldt (n 63) 8f. Facebook was in fact fined for misrepresenting the amount of unlawful content in its first transparency report of 2018, by steering users towards its standard feedback and reporting channels German Justice Ministry, 'Federal Office of Justice issues Fine against Facebook', 2 July 2019, https://www.bundesjustizamt.de/DE/Presse/Archiv/2019/20190702_EN.html.

[114]The community standards of the big platforms would have already been influenced by the *EU Code of Conduct on countering illegal hate speech* (2016), see (n 14).

[115]For a critical account of private censorship and its democratic deficit, see Koltay (n 67) 301ff.

as 'it had no firm ground in German constitutional law'.[116] However, that private parties can be *indirect* human rights bearers – through, for example, interpreting a platform's Terms and Conditions in line with fundamental rights – is well established in German law.[117] Furthermore, this indirect horizontal effect of fundamental rights may at times not be very different from the direct responsibility of state actors, particularly when – as in the words of the German Constitutional Court in *Beer-Cans-Flashmob for Freedom* (2015)[118] – 'private enterprise provides the infrastructure for public communications and so discharges a function that previously belonged *solely* to the State'.[119] Thus, in that case, the private owner of a town square, otherwise freely used by the public, had to allow a public assembly on the square as freedom of speech and assembly trumped his proprietary rights. By implication, there is a legal avenue for the imposition of negative 'keep-up' duties of legal content on private platforms under German law. However, NetzDG has not gone down that route of creating keep-up duties, and the decision by the German Federal Court of Justice in *Hate Speech Posts Deletion and Account Closure* (2021) requiring respect for users' due process rights was premised on the *prima facie* right of platforms to impose their own communication standards on users and sanction breaches through the takedown of posts and the closure of accounts, *even if* the posts were legal.[120]

A similar debate has ensued in the US – one that has relied on the Supreme Court's decision in *Marsh v Alabama* (1946),[121] according to which a private entity that owned and operated a town did perform municipal functions and thus must abide by the First Amendment. This holding was later restricted to private companies which 'perform the full spectrum of municipal powers'[122] and conceptualised as part of the 'state action' doctrine. In *Prager University v Google LLC* (2020),[123] the Court of Appeals for the Ninth Circuit rejected the argument that the state action doctrine should extend to YouTube as, according to the Court, despite its ubiquity and its

[116]Wischmeyer (n 63) 4; Malte Engeler, 'Meinungsfreiheit: Warum Facebook (zu Recht) nicht an Grundrechte gebunden ist' (18 Sept 2018) *Netzpolitik.org*; see also *Fraport* 1 BvR 699/06 [2011] BVerfG ECLI:DE:BVerfG:2011:rs20110222.1bvr069906, where the State owned and had a controlling influence (70% of the shares) of the airport in question and thus was held to be *directly* bound by fundamental rights

[117]For a discussion of German judgments on that point, see Koltay (n 67) 283ff.

[118]*Beer-Cans-Flashmob for Freedom* 1 BvQ 25/15 [2015] BVerfG, ECLI:DE:BVerfG:2011:rs20110222.1bvr069906; see also *Stadiumban* 1 BvR 3080/09 [2018] BVerfG ECLI:DE:BVerfG:2018:rs20180411.1bvr308009 which affirms that private actors, such as event organisers, who can control the access of others to participate in social life, must not use their discretionary powers to exclude specific persons from such events without factual reason.

[119]ibid, para 6 [translation by author, emphasis added].

[120](n 109). For a number of inconsistent judgments by lower courts on the negative duties of platforms, see Echikson (n 63) 11 and (n 117).

[121]*Marsh v Alabama* 326 U.S. 501 (1946)

[122]*Lloyd Corp v Tanner* 407 U.S. 551 (1972), 569; *Hudgens v NLRB* 424 U.S. 507 (1976), 518ff.

[123]*Prager University v Google LLC* No. 18–15712 (9th Cir. 2020).

public-facing nature, it remained a private forum: 'merely hosting speech by others is not a traditional, exclusive public function and does not alone transform private entities into state actors subject to First Amendment constraints'.[124] As a result, YouTube's content moderation in line with its Terms of Service and Community Guidelines would not be subject to constitutional scrutiny.

So although under both German and US law certain privately owned *physical* spaces freely accessible to the public have attracted constitutional protection of speech, religion and assembly,[125] these privileges (and attendant burden on the proprietors) have not been extended to comparable online spaces – although, as noted above,[126] under German law there is room for negative 'keep-up obligations' in particular cases. Doctrinally, this has been justified (in the US), on the basis that platforms do not perform a function that was 'traditionally the *exclusive* prerogative of the State'.[127] Regardless of the merits of that reasoning,[128] the effect is that the online domain is constructed as a quintessentially private one in which proprietary and contractual rights create the overarching legal framework for activities. Practically, it means that platforms enjoy editorial discretion and can moderate content (largely in the German context) as they please, which they could not do if they were 'state actors' under US law or indirectly *fully* bound by free speech obligations under German law. Taking it one step further, through the construction of platforms as solely private actors, the State *de facto* delegates content moderation to platforms – as something that it could not do itself (US) or would not be willing to do itself (Germany). This 'delegation' operates on a wholesale level, as opposed to the moderation based on specified offences, discussed above (Section 3.1). Whilst NetzDG – in conjunction with German criminal law – makes some public inroads into the privately owned and governed online space, it does not shake its essentially private governance framework and the default supremacy of 'Terms and Conditions' and 'Community Guidelines'.[129] The private super-structure explains why user complaints, no matter their basis, are first and foremost tested against house rules, and only residually subjected to the 'official' demands made under NetzDG. This has been criticised for

[124]ibid 5, citing *Manhattan Community Access Corp v Halleck* 139 SCt 1921 (2019), 1930.
[125]Note *Marsh v Alabama* 326 U.S. 501 (1946) has not been extended to shopping malls, see *Lloyd Corp v Tanner* 407 U.S. 551 (1972).
[126]See above (n 120).
[127]*Prager University v Google LLC* No. 18–15712 (9th Cir 2020), citing *Rendell-Baker v Kohn* 457 U.S. 830 (1982), 842 [emphasis in the original].
[128]For critical engagements with this position, see Matthew P Hooker, 'Censorship, Free Speech & Facebook: Applying the First Amendment via the Public Function Exception' (2019) 15(3) Washington Journal of Law, Technology & Arts 35.
[129]The same principle also underlies the frameworks to be put in place by the EU Digital Services Act and the UK Online Safety Bill (n 14).

sidestepping NetzDG[130] and reducing its effect to 'swifter and more consistent removal within Germany under the companies' community guidelines'[131] – a criticism that has validity especially if platform thereby also bypass the due process obligations.[132] Still, the review hierarchy is consistent with the legally sanctioned private governance framework of cyberspace.

Second, even for the minority of user complaints not captured by a platform's community guidelines and which have thus to be evaluated against the standards of German criminal law, the privatising effect of those official standards occurs through the combination of broad legal concepts and a lack of transparency of individual decisions. In fact, the broadness of the concepts that platforms have to apply has been a target for the detractors of NetzDG. David Kaye, UN Special Rapporteur on Freedom of Expression, criticised NetzDG's reliance on 'prohibitions … based on vague and ambiguous criteria … [and] highly dependent on context, context which platforms are in no position to assess',[133] a sentiment echoed by the OSCE and Reporters Without Borders,[134] and invariably premised on the concern that legal content may be taken down. As argued above (in this Section), this is in any event a liberty which platforms enjoy as a necessary incident of their legal construction as private actors. Platforms also already apply many of these 'vague' and 'ambiguous' criteria as part of their own community standards whenever these mirror the public ones. For example, Facebook's own prohibition of hate speech is very similar to s.130 of the German Criminal Code, albeit more restrictive:

> It [hate speech] creates an environment of intimidation and exclusion, and in some cases may promote offline violence. We define hate speech as a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.[135]

Whilst the broad criminal law concepts are invariably supplemented by sophisticated legal edifices that draw on subtle distinctions as, for example, between strongly expressed political views and denigrating personal

[130]Heldt (n 63) 13.
[131]Tworek (n 63) 6.
[132]But see judgment in *Hate Speech Posts Deletion and Account Closure* (n 103), where the German Court required basic due process rights in respect of a takedown based on Facebook's Community Standards.
[133]David Kaye, Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 'Open Letter to German Government' (I June 2017) *Office of High Commissioner of Human Rights* https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf.
[134]Bernd Holznagel, 'Legal Review of the Draft Law on Better Law Enforcement in Social Networks' (May 2017), commissioned by the Organisation for Security and Co-operation in Europe, Office of the Representative on Freedom of the Media, 14f; Reporter ohne Grenzen, 'NetzDG führt offenbar zu Overblocking', 27 July 2018; see also Article 19, 'Germany: The Act to improve Enforcement of the Law in Social Networks' (August 2017), available on *Article19.org*.
[135]Meta, 'Facebook Community Standards – Hate Speech' *Transparency Center* (accessed 29 Jan 2022).

attacks on a politician,[136] the broadness of the concepts creates a wide spectrum for legitimate variations in the final assessments. Indeed, in relation to the State, the ECtHR cautioned that 'where the State's discretion to prosecute for such offences becomes too broad … [it is] potentially subject to abuse through selective enforcement'[137] – a concern fully aligned with American fear of public abuse of power. Under NetzDG, however, it is private platforms that move into the discretionary space and import their perspectives and ideologies into the interpretation of 'hate' and like concepts. This importation of private values is partially counteracted by NetzDG's requirement of training the content assessor regularly.[138] Still, the evaluations themselves are distanced from government. Furthermore, given that individual decisions are not made publicly available, as rightly criticised by various commentators,[139] there is also no wider exposure to public scrutiny of the platforms' application of the official standards. NetzDG's general transparency obligations are all of an aggregate nature, and they do not give insight into the reasons for individual decisions,[140] to which only the parties have access and, it appears, the supervisory authority.[141] Last but not least, the 'natural' preference of platforms for autonomous inhouse decision-making also explains why the possibility of delegating difficult decisions to 'a recognised self-regulatory institution' has had a fairly low take-up.[142]

The fact that NetzDG imposes *official* standards based on view-point discrimination of speech and would thus be struck down as unconstitutional in America, only tells half the story. The other half unfolds when these official standards are privatised in their implementation by private platforms. NetzDG supports and legitimises a legal landscape whereby, on the one hand, public law is absorbed into private contractual normativity, as expressed in Terms of Service and Community Standards, and, on the other hand, residual complaints are resolved through largely autonomously inhouse processes applying broad legal concepts with much room for discretion and without routine scrutiny of individual decisions. This is not to say that NetzDG would escape the American constitutional wrath but to show

[136] *Renate Künast v Facebook* 27 AR 17/19 (2019) LG Berlin ECLI:DE:LGBE:2019:0909.27AR17.19.00, concerned the hate speech directed at the Green Party politician Renate Künast on Facebook which were both personally insulting and politically charged.

[137] *Savva Terentyev v Russia* App no10692/09 (ECHR, 28 August 2018) [85]; *Altuğ Taner Akçam v Turkey* App no 27520/07 (25 October 2011) [93–94].

[138] Section 3(4).

[139] Eifert (n 12) 1452ff; Tworek (n 63) 8 (mentioning the suggestion of a 'Clearing House' for cases where users disagreed with the handling of the decision, or alternatively a 'research repository' with data from all platforms).

[140] See s 3(2) on the aggregate information to be included in the biannual reports.

[141] For the affected parties, see s 3(5), and for the more ambiguous position of the supervising authority, see s 4(1)2 and s 4a(3) and s 3(5) which allows the supervisory authority to follow the complaint procedure.

[142] S 3(2)3.b and 3(6)-(11); see also Tworek (n 63) 6 (on low frequency of platforms seeking outside counsel); Echikson (n 63) 14f.

why the implementation of public content standards through relatively autonomous private processes should go towards assuaging the American primary fear of government's interference with speech.

Conclusion

The premise of this paper was that the internet has changed the public sphere profoundly. It has, by orders of magnitude, increased the amount, visibility and permanence of unfiltered public speech, and intensified the globalisation of communication that makes the traditional territorial speech silos more problematic, even if not impossible. NetzDG responds to both pressures by formalising the gatekeeping role of key platforms in Germany; it does so through injecting local public content standards and basic due process requirements into the platforms' own content moderation practices. It does not displace these practices but informs and complements them. Overall, NetzDG sanctions the private regulation of cyberspace and strengthens its legitimacy through the background presence of the State, that is, through *regulated* self-regulation.[143] It is neither at odds with the global nature of the platform activity nor out of tune with the platforms' loyalty to the American free speech tradition.

Indeed, this paper's main argument is that NetzDG strikes a workable compromise between the seemingly irreconcilable European and American speech traditions based on their divergent conceptions of empowered citizenry. Its key mechanism of enforcing public speech standards through private processes is uniquely adept at answering the German, or European, need to have hate and like speech removed from the public domain, as well as the American need for government to be kept on the sidelines. Still, the case in support of NetzDG's success in bridging the transatlantic speech divide is not flawless, and neither can it be. By showing how NetzDG injects significant public dimensions into otherwise entirely unaccountable private processes, and how the criminal speech standards upon which it is based are largely privatised in the course of the implementation process, the discussion also feeds the concerns of the other side. This may perhaps best be understood as an inevitable side effect of a convergence of the speech traditions, not in terms of their constitutional understanding, but rather as a convergence through process. The more general insight, however, is that the key difference between American and European free speech conceptions is not a binary one but reflects primary preoccupations that are different in kind: the First Amendment addresses the primary fear of government interfering with free speech (Who), whilst the primary

[143]Contrast to pure self-regulation: Kate Klonick, 'The New Governors: The People, Rules and Process Governing Online Speech' (2018) 131 Harvard Law Review 158.

concern behind the European conception lies in protecting the equality and inherent dignity of citizens in the public sphere (What). This divergence in their preoccupations creates room for negotiations.

Finally, America's adherence to First Amendment absolutism means that, as a government and regulator, it is largely immune to the savageries of online communications and to the pressures experienced by governments of other liberal democracies to respond to them. Yet, it also means that other States have taken the lead in the online governance thematic, often with effects well beyond their national borders. The 'Brussels effect' refers to the regulatory ripples caused by EU regulation, such as the GDPR, well beyond European borders.[144] This may be the result of combination of the extraterritorial design of the regulation, or the preference of global actors to operate under unified global Community Standards,[145] or other countries copying (worthwhile) regulatory initiatives.[146] In design, NetzDG is restricted to the German online space, yet it may well serve as a blue-print for responsible private regulation more widely.

Acknowledgement

University of Southampton. Many thanks to Mathias Hong, Lorna Woods, Chara Bakalis and Jacob Eisler for helpful comments and suggestions on an earlier draft, and to Sophie Turenne and Oliver Butler for organising the workshops in the lead up to this special edition.

Disclosure statement

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

Notes on contributors

Uta Kohl is Professor of Law at Southampton Law School; she has an interest in internet governance, particularly its transnational dimensions. She is the author of the monograph Jurisdiction and the Internet (CUP 2007), the editor of The Net and the Nation State (CUP 2017) and the co-editor of Data-Driven Personalisation in Markets, Politics and Law (CUP 2021, with Jacob Eisler).

[144]See Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (OUP 2020).
[145]Eifert (n 12).
[146]There has also been the concern that authoritarian regimes may be inspired by the mechanism introduced by the law to use it for their own problematic purposes, which in fact occurred when Russia copied parts of NetzDG for its anti-terrorism law, see, e.g. Tworek (n 63) 4.