Hate speech online: an (intractable) contemporary challenge?

Kate O'Regan*

Abstract: Internet users generate billions of pieces of online content weekly across a number of social media platforms, and that content includes hate speech. How to respond to hate speech online is a question that is troubling democracies all over the world and there is no easy solution in sight. The question of hate speech has long given rise to dispute in international human rights law, a dispute that arises because the protection of freedom of speech on the one hand and the prohibition of hate speech on the other are rooted in different normative principles that need to be accommodated. This enduring dispute makes it particularly difficult to design solutions to the problem of hate speech online. The article describes and assesses the current rules regulating hate speech online in four jurisdictions, the USA, the UK, Europe and Germany and suggests that it is not clear that any of the systems satisfactorily address the issue of hate speech online.

Keywords: hate speech, freedom of speech, Internet regulation, article 19 of the International Covenant on Civil and Political Rights, article 20(2) of the International Covenant on Civil and Political Rights.

1. *Introduction*

The global spread of the Internet is breath-taking. In 2018, the number of Internet users exceeded 4 billion, more than half the global population. 3.19 billion people used social media platforms,[1] with Facebook claiming it has more than 2 billion users, and estimating that every day more than 800 million people across the globe 'like'

* Professor of Human Rights Law, Director of the Bonavero Institute of Human Rights, University of Oxford. I am grateful to Dr Stefan Theil for his assistance with translating the German Networks Act, discussed below, and for his comments on an earlier draft of this paper. I am also grateful to Professor Jeff King, Mr Oliver Butler and to two anonymous peer reviewers of Current Legal Problems for their helpful comments on earlier drafts of this article.

[1] See the report *Global Digital Report 2018* on the wearesocial website which draws on data from a wide range of sources, including Google's Consumer Barometer, Statista, StatCounter, and GlobalWebIndex. Simon Kemp, 'Digital in 2018: World's internet users pass the 4 billion mark' (*wearesocial,* 30 January 2018) <https://wearesocial.com/blog/2018/01/global-digital-report-2018> accessed on 31 October 2018.

something on Facebook.[2] YouTube claims one billion users, and claims also that one billion hours of YouTube content is watched daily.[3]

Access to the Internet is global, but the developed world still has markedly higher internet penetration than the developing world, with nearly 90% Internet penetration in western Europe and North America, approximately 66% penetration in eastern Europe, Latin America and Oceania, around 50% in East Asia and the Middle East, 33% in south Asia and only 29% in Africa. Nevertheless the global picture is changing too. For example, the Asia Pacific region now accounts for more than half of all Internet users globally and accounted for 70% of total growth in Internet users across the globe in 2017.[4]

Since the early years of this century, the Internet has become a form of communication in which users generate content to share with other users, often referred to as Web 2.0.[5] Web 2.0 is an indiscriminate vehicle for all forms of speech: the virtuous and the violent, the true and the false, idle chatter and fighting words. And one of the questions that the Internet presents is how to manage hate speech online. There is no universally accepted definition of hate speech, but for the purposes of this article, I am going to use the term hate speech to refer to speech that incites hatred or is degrading of individuals or groups of individuals based on their race, ethnic origin, sexual orientation, gender identity or other similar attribute.

Whether and how to regulate hate speech online is a question that is being asked with increasing concern in democracies all over the world and has been the subject of recent legislation in Germany,[6] is the subject of an on-going enquiry in the

---

[2] See M Nowak and G Spiller, 'Two Billion People Coming Together on Facebook' (*Facebook newsroom*, June 27 2017) <https://newsroom.fb.com/news/2017/06/two-billion-people-coming-together-on-facebook/> accessed on 30 October 2018.

[3] See *YouTube for Press* (*Youtube*) <https://www.youtube.com/yt/about/press/> accessed on 31 October 2018.

[4] See *Global Digital Report 2018* (n 1).

[5] See discussion in M Yar, 'A Failure to Regulate? The Demands and Dilemmas of Tackling Illegal Content and Behaviour on Social Media' (2018) 1 International Journal of Cybersecurity Intelligence and Cybercrime 5, 6.

[6] See the full discussion of the *Netzwerkdurchsetzungsgesetz* (Network Enforcement Law) in Section 5.D. below. Its long title is Law for the Improvement of the Legal Regulation of Social Networks, and it came into force on 1 October 2017.

Home Affairs Committee of the House of Commons here in the UK,[7] and is under scrutiny by the European Commission.[8]

This article seeks to explore some of the issues that arise when we seek to answer the question of what to do about hate speech online. First, I am going to talk about freedom of speech, and describe how it is widely proclaimed, but far less widely protected, and outline the key values that inform our commitment to freedom of speech. Secondly, I shall consider the values that inform the commitment to outlawing hate speech and argue that the balance to be struck between the protection of freedom of speech and hate speech is a shifting one, dependent in part on context and history, as well as the weight given in different constitutional settings to the different values at play. I suggest that it is the shifting balance between protecting freedom of speech and prohibiting hate speech gives rise to arguably the greatest challenge in addressing hate speech online.

Thirdly I shall consider how the publication of information on digital media differs from publication on traditional media, and what that might imply for the legal tools that we might use to address hate speech on line. Then I will describe how the regulation of Internet platforms in relation to hate speech is being managed in four different jurisdictions: the USA, the UK, the EU and Germany, and then I will assess these four approaches and conclude by noting that there is no easy answer to the question of how to regulate hate speech online, and that it will be important to monitor the initiatives that are being developed closely in the years ahead to ensure that appropriate balances are struck between the need to protect freedom of speech, on the one hand, and prohibit hate speech, on the other.

---

[7] See the report of the House of Commons Select Committee on Home Affairs, *Hate Crime, Abuse, hate and extremism online* (HC 2017) <https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/60902.htm> accessed on 30 September 2018. Following the 2017 general election, the Select Committee on Home Affairs resumed its inquiry into Hate Crime and its Consequences. The enquiry is ongoing.

[8] The European Commission has not issued a directive on countering illegal content on line, but issued a *Communication on Tackling Illegal Content Online – Towards Greater Responsibility of the Internet Platforms* (Communication) COM (2017) 555 final <https://ec.europa.eu/digital-single-market/en/news/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms> accessed on 30 September 2018 and a *Recommendation on Measures to Effectively Tackle Illegal Content Online* (Recommendation) C (2018) 1177 final <https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms> accessed on 30 September 2018. These are discussed further below.

2. *Freedom of Speech*

Considering what we should do about hate speech online requires us to start with what we mean by freedom of speech and why it matters. A commitment to freedom of speech is widely accepted as a core component of the international human rights framework and of the constitutional framework of most contemporary democracies.

It is entrenched in many international human rights documents, such as Article 19 of the Universal Declaration of Human Rights, Article 19 of the International Covenant on Civil and Political Rights,[9] Article 10 of the European Convention on Human Rights, and in most domestic Bills of Rights. 169 countries of 197 have, according to the UN Office of the High Commissioner for Human Rights, ratified the ICCPR, a further six have signed it (which includes China, who has signed, but not ratified), and only 22 have taken no steps with regard to it (a number which includes Saudi Arabia, South Sudan and Myanmar).[10]

Article 19 of the ICCPR declares that

> [e]veryone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media…[11]

On paper, freedom of expression is thus one of the most widely asserted rights in the world. Yet this apparent unanimity belies two worrying facts: the first is that the acknowledgement of freedom of speech in the ratification of conventions and entrenchment in constitutions does not match the practice in many countries; and the

---

[9] International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR).

[10] For a list of the countries that have signed and ratified the ICCPR, see Office of the High Commissioner for Human Rights, 'Status of Ratification' <http://indicators.ohchr.org> accessed on 30 October 2018.

[11] The full text of Article 19 ICCPR provides as follows:

> 1. Everyone shall have the right to hold opinions without interference.
> 2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.
> 3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:
> (a) For respect of the rights or reputations of others;
> (b) For the protection of national security or of public order (ordre public), or of public health or morals.

second is that what we mean (and protect) when we speak of freedom of speech varies across jurisdictions, and one of the issues on which it varies most sharply is the question of hate speech.

So the apparent universal recognition of freedom of speech belies a wide gap between the formal provisions entrenching freedom of speech in legal documents and the actual protection it receives in practice. Of course it is not unusual for there to be a discrepancy between law in the books and law in practice, but the cleft between freedom of speech and its protection is disturbingly wide, and it seems to be growing.

The civil rights NGO, *Article 19*, has partnered with the social science database V-Dem Institute to produce an analysis of the state of freedom of expression in the world. The analysis is based on a metric that considers five indices of freedom of expression, based on 32 separate indicators. The five indices are civic space, the extent of media freedom and media pluralism, transparency (the extent to which individuals can gain access to information), regulation of the digital world in a manner compatible with freedom of speech and effective networks to protect those campaigning to defend freedom of speech.[12] The 2017 inaugural report draws on data that V Dem has been collecting for over a decade, and the picture it paints is a worrying one. In particular, the report notes the sharp decline in the extent of media freedom and media pluralism across the globe, as well as an alarming rise in the number of attacks on journalists and free speech defenders.[13] So the apparent global consensus on free speech masks widespread practices that limit freedom of speech in an increasing number of countries across the globe.

In thinking about the manner in which hate speech is or should be regulated, therefore we must take into account the existing global challenges for effective protection and promotion of freedom of expression. Although many countries proclaim a commitment to freedom of expression, in practice they do not protect freedom of expression. We should be alert in proposing new forms of regulation to address hate speech to the risk that they may be used for the perverse purpose of undermining freedom of speech.

---

[12] 'The Expression Agenda Report 2016/2017: The state of freedom of expression and information around the world' (*Article 19*, 2017) <https://www.article19.org/xpa-17/> accessed on 30 October 2018.
[13] ibid.

The second challenge for the global protection of freedom of speech is that there is deep disagreement as to what freedom of speech requires. And one of the core sources of disagreement is what we should do about speech that is hateful. The disagreement can be found at the heart of international human rights law. Article 20(2) of the ICCPR, the clause that follows article 19, provides that '[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.'

This is a clause that requires certain forms of speech to be prohibited, not protected. The speech it targets is speech that advocates national, religious or racial hatred in a manner that incites discrimination, hostility or violence. Many countries, including Australia,[14] Belgium,[15] New Zealand,[16] the United States of America and the United Kingdom[17] entered reservations when ratifying the ICCPR relating to article 20(2). The text of some of these reservations suggests that the state parties are anxious to ensure that the obligation imposed by article 20(2) does not unduly impair freedom of expression. The United States of America framed its reservation in the following terms:

> That article 20 does not authorize or require legislation or other action by the United States that would restrict the right of free

[14] The Australian reservation is in similar terms to the UK reservation, which is set out in n 17 below. UNTS, 'International Covenant on Civil and Political Rights' (Chapter IV, Human Rights) < https://treaties.un.org/Pages/ViewDetails.aspx?chapter=4&clang=_en&mtdsg_no=IV-4&src=IND> accessed 7 November 2018.

[15] The Belgian reservation reads as follows: 'The Belgian Government declares that it does not consider itself obligated to enact legislation in the field covered by article 20, paragraph 1, and that article 20 as a whole shall be applied taking into account the rights to freedom of thought and religion, freedom of opinion and freedom of assembly and association proclaimed in articles 18, 19 and 20 of the Universal Declaration of Human Rights and reaffirmed in articles 18, 19, 21 and 22 of the Covenant.' UNTS (n 14).

[16] The New Zealand reservation reads as follows: 'The Government of New Zealand having legislated in the areas of the advocacy of national and racial hatred and the exciting of hostility or ill will against any group of persons, and having regard to the right of freedom of speech, reserves the right not to introduce further legislation with regard to article 20.' UNTS (n 14).

[17] The United Kingdom reservation states that the Government of the United Kingdom interprets '… article 20 consistently with the rights conferred by articles 19 and 21 of the Covenant and having legislated in matters of practical concern in the interests of public order *(ordre public)* reserve[s] the right not to introduce any further legislation.' UNTS (n 14)

> speech and association protected by the Constitution and laws of the United States.[18]

Article 4 of the Convention on the Elimination of All Forms of Racial Discrimination (CERD) contains a similar provision to Article 20(2).[19] The United States entered a similar reservation in relation to Article 4,[20] stating that the Constitution and laws of the United States provide extensive protection for freedom of speech and association and the United States does not accept an obligation to restrict those rights.

In assessing this disagreement at the heart of freedom of expression law, it is useful to return to what are often described as the three key purposes of the protection of freedom of expression – the reasons that are commonly given to explain why we should protect freedom of speech.

The first relates to a conception of what it means to be human, and what that implies. The key proposition is that we should respect (and seek to foster) the autonomy of human beings. The protection of individual autonomy requires us to

---

[18] UNTS (n 14).

[19] International Convention on the Elimination of All Forms of Racial Discrimination (adopted 21 December 1965 UNGA Res 2106 (XX); entered into force 4 January 1969) 660 UNTS 195 (CERD). Article 4 CERD provides that:

> States Parties condemn all propaganda and all organizations which are based on ideas or theories of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form, and undertake to adopt immediate and positive measures designed to eradicate all incitement to, or acts of, such discrimination and, to this end, with due regard to the principles embodied in the Universal Declaration of Human Rights and the rights expressly set forth in article 5 of this Convention, inter alia:
> (a) Shall declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin, and also the provision of any assistance to racist activities, including the financing thereof; […]

[20] The relevant portion of the reservation states: 'That the Constitution and laws of the United States contain extensive protections of individual freedom of speech, expression and association. Accordingly, the United States does not accept any obligation under this Convention, in particular under articles 4 and 7, to restrict those rights, through the adoption of legislation or any other measures, to the extent that they are protected by the Constitution and laws of the United States.' UNTS, 'International Convention on the Elimination of All Forms of Racial Discrimination' (Chapter IV, Human Rights) < https://treaties.un.org/Pages/ViewDetails.aspx?src=IND&mtdsg_no=IV-2&chapter=4&clang=_en> accessed 7 November 2018.

protect the right of all to express themselves freely and their right to have access to the thoughts and ideas of others. Protecting free speech is therefore seen as enabling the self-fulfillment of individual members of a society. This principle underpins our commitment to arguably all forms of expression, including artistic and academic freedom.

The second reason relates to the importance of freedom of speech for democracy. In order for self-government to flourish, for individuals to make good democratic decisions, the argument goes, they need to be able to have open discussions about matters of public importance.[21] Some theorists, C Edwin Baker, for example, have persuasively suggested that the first and second types of reason are related, that we value democracy and freedom of expression because we value the autonomy of human beings and their capacity for self-government.[22]

In developing his argument that we protect freedom of speech in order to protect a system of democratic government, Robert Post draws on Kelsen's distinction between 'autonomous' and 'heteronomous' forms of government. Democratic societies are autonomous because their laws are made (at least indirectly) by the people to whom they apply, whilst under heteronomous forms of government, the laws are not made by those to whom they apply.[23] Heteronomous forms of government therefore lack democratic legitimacy because they are not based on the will of the people.

Autonomous forms of government on the other hand are based on the principle of self-determination and, according to Kelsen, require 'a running discussion between majority and minority, through free consideration of arguments for and against a certain regulation of a subject matter.'[24] And by protecting freedom of speech we make this 'running discussion' possible.

Ronald Dworkin adopted a similar argument when he argued that 'it is illegitimate for governments to impose a collective … decision on dissenting

[21] See for example A Meiklejohn, *Freedom of Speech and its Relation to Self-Government* (Harper and Bros 1948) 26–7; J Weinstein, 'Extreme Speech, Public Order and Democracy: Lessons from *The Masses'* in I Hare and J Weinstein (eds), *Extreme Speech and Democracy* (OUP 2009) 26ff.
[22] See CE Baker, 'Autonomy and Hate Speech' in I Hare and J Weinstein (eds) (n 21) 139, 146.
[23] See R Post, 'Racist Speech, Democracy and the First Amendment' (1991) 32 William and Mary Law Review 267, 280.
[24] ibid 281.

individuals, using the coercive powers of the state, unless that decision has been taken in a manner that respects each individual's status as a free and equal member of the community.'[25]

There can be no doubt that democracies require freedom of speech to be protected in order to ensure the public discussion that will enable them to work properly. But there are two caveats that we should note about the principle.

The first is that not all speech is part of the democratic project, only some of it is. Robert Post accepts (along with the US Supreme Court)[26] that this rationale requires public discourse to be protected to secure democracy. The corollary is that forms of speech that cannot be classed as 'public discourse' cannot draw on this rationale for protection. Drawing the line between public discourse and other forms of speech is a delicate task, and one that judges do not always perform successfully. Indeed, Robert Post in speaking of the jurisprudence of the United States Supreme Court in this regard observed that 'in contemporary doctrine, …, this distinction is notoriously ill-conceived and unreliable. In fact it is commonly accepted that the Court's efforts in this direction have resulted in a dreadful mess.'[27] It is an open question whether this line can ever be drawn in a manner that will not be contested.

The second rider is that the principle of free speech is only one of the normative principles that informs our conception of democracy. There are others, and perhaps the most important in this context, is the principle that all human beings are worthy of equal respect. How these two principles should be accommodated within free speech doctrine is a question that lies at the heart of how we approach hate speech. But before exploring that question more fully, I want to turn briefly to the third reason that is said to underpin our commitment to freedom of speech.

The third reason for protecting free speech, even speech with which we disagree, is that it enables our discovering that established truths are in fact false and therefore fosters the possibility of new truths. This epistemic claim draws on the work

---

[25] See R Dworkin, 'Foreword' to I Hare and J Weinstein (eds) (n 21) vii. See also Dworkin's 'Reply to Jeremy Waldron' in M Herz and P Molnar (eds), *The Content and Context of Hate Speech: Rethinking Regulation and Responses* (CUP 2012) 341.

[26] See *The Hustler Magazine v Falwell* 485 US 46, 54 (1988), and see Robert Post's discussion of the decision in 'The Constitutional Concept of Public Discourse – Outrageous Opinion, Democratic Deliberation and *Hustler Magazine v Falwell*' (1990) 103 Harvard LR 601.

[27] See Post (n 26) 667.

of both JS Mill[28] and Milton[29] and is often referred to as Mill's argument from fallibility. It underpins the proposition that free speech enables 'a marketplace of ideas' as articulated in the famous dissent of Judge Holmes in *Abrams v US* in which he said (and I paraphrase) that the best test of truth is its power to get itself accepted in the competition of the market.[30] It is this principle that informs the argument that the cure for bad speech is more speech.[31]

The argument from fallibility has been subjected to powerful critique by contemporary scholars.[32] Frederick Schauer, for example,[33] has argued that one of the principal consequences of allowing the expression of falsehoods is that it may increase the number of people who believe in those falsehoods. Mill appears to assume that truth will eventually win out, but as Schauer observes modern social science suggests that the objective truth of a proposition may well not be the most significant factor in determining whether it comes to be accepted as true. And that one cannot therefore be certain that permitting facts that are untrue to be protected by freedom of speech will necessarily assist in the discovery and acceptance of truth.

Goldman and Cox have also convincingly explored the marketplace of ideas metaphor. They argue that the proposition that a blanket protection freedom of speech will create a 'speech' marketplace that will ensure that objectively true propositions come to be accepted as true does not hold.[34] They explain that the classical economic theory of markets suggests that free and competitive markets are the best way to organise the production and consumption of goods. And they provide a range of reasons why this theory cannot simply be replicated in the world of free speech. For example, classical economic theory proposes that what markets will produce is what consumers want, but say, Goldman and Cox, there is no evidence that consumers of

---

[28] See JS Mill, *On Liberty* (1859).
[29] See J Milton, *Areopagitica: A Speech for the Liberty of Unlicensed Printing* (1644).
[30] 250 US 616, 630 (1919).
[31] See F Schauer, 'Social Epistemology, Holocaust Denial and the Post-Millian Calculus' in M Herz and P Molnar (eds) (n 25) 129, 131.
[32] See, for example, Schauer (n 31) 129; A Goldman and J Cox, 'Speech, Truth and the Free Market for Ideas' (1996) 2 Legal Theory 1; J Waldron, *The Harm in Hate Speech* (Harvard UP 2012) 155-6; V Blasi, 'Holmes and the Marketplace of Ideas' [2004] Supreme Court Review 1; D Bush, 'The Marketplace of Ideas: Is Judge Posner chasing Don Quixote's Windmills?' (2000) 32 Arizona State Law Journal 1107; P Brietzke, 'How and Why the Marketplace of Ideas Fails' (1997) 31 Valparaiso University Law Review 951.
[33] See Schauer (n 31).
[34] See Goldman and Cox (n 32).

speech want truth, or that they want truth alone. Indeed, they also point out that consumers of speech will often look for speech messages that confirm their existing beliefs, confirmation bias, rather than look for messages that will challenge them.[35]

This third reason for the protection of freedom of speech is thus less cogent than the first two. And it is the first two reasons (that we value speech because it fosters self-fulfillment and because it makes democracy possible) that I think we should bear in mind when we think about how to approach hate speech.

3. *Why We Prohibit Hate Speech*

I turn now to ask what are the reasons that inform a decision to prohibit hate speech. There are several reasons that are suggested. The first is the need to maintain public order, one of the duties of the modern state. In some countries, that this is an important purpose of the hate speech prohibition appears from the terms of the hate speech prohibition itself. So one of the elements of the prohibited conduct is that it is likely to lead to a 'breach of the peace'[36] or a breach of public order.

The risk of imminent violence, which would result in a breach of public order, is the principle established by the US Supreme Court in the leading case of *Brandenburg v Ohio*.[37] Brandenburg was a member of the Ku Klux Klan who had been convicted of an offence under the Ohio Criminal Syndicalism statute on the basis, as described by the Supreme Court, of a 'film [that] showed 12 hooded figures, some of whom carried firearms. They were gathered around a large wooden cross, which they burned. No one was present other than the participants and the newsmen who made the film. Most of the words uttered during the scene were incomprehensible when the film was projected, but scattered phrases could be understood that were derogatory of Negroes and, in one instance, of Jews.'[38] The Ohio statute provided that persons who 'advocate or teach the duty, necessity, or propriety' of violence 'as a means of accomplishing industrial or political reform' commit the offence of criminal syndicalism.

The Supreme Court held that 'the principle that the constitutional guarantees of free speech and free press do not permit a State to forbid or proscribe advocacy of

[35] ibid 31.
[36] See, for example, para 219(1) of the Canadian Criminal Code, 1985.
[37] *Brandenburg v Ohio* 395 US 444 (1969).
[38] ibid 444–5.

the use of force or of law violation except where such advocacy is directed to inciting or producing imminent lawless action and is likely to incite or produce such action.'[39]

In *Brandenburg*, the US Supreme Court carved a very narrow exception to the protection of free speech on the basis of public order concerns. It is possible, of course, to carve broader exceptions to free speech on the basis of public order concerns by prohibiting speech 'that is likely to lead to a breach of the peace' without stipulating the likelihood of imminent violence.

Eric Heinze has argued that whether there is a causal link between hate speech and a threat to public order cannot be answered in the same manner across all democracies. He suggests that 'long-standing, stable and prosperous democracies' (LSPDs) do not produce atrocities, and therefore a prohibition on hate speech is not necessary to preserve public order in such democracies.[40] He accepts that in non-LSPDs, hate speech bans may be necessary in order to preserve public order. Heinze's argument that prohibitions on hate speech may be more important to preserving public order in new and deeply divided democracies has intuitive appeal. If he is right then it suggests that there is not a 'one size fits all' approach to the manner in which hate speech should be regulated across democracies. Heinze also acknowledges that LSPDs are a relatively recent form of democracy that emerged after the end of the Second World War. Contemporary events might be understood to be raising the question how permanent the status of LSPD may prove, and whether the assertion that they will not be damaged by hate speech will continue to hold. Regardless of how resilient LSPDs may prove in the long run, Heinze's evidence establishes that whether a prohibition on hate speech will be necessary or appropriate to preserve public order in any particular democracy is not a question that can be answered with certainty forever and will need to be reviewed in every democracy from time to time.

In addition to the preservation of public order, there is another important argument that underpins prohibitions on hate speech. Jeremy Waldron has argued that

---

[39] ibid 447.

[40] See E Heinze, *Hate Speech and Democratic Citizenship* (OUP 2016) 69. Heinze relies on *The Economist*'s annual *Democracy Index* and identifies 25 democracies that 'maintain sufficient legal, institutional, educational, and material resources to admit all viewpoints into public discourse, yet remain adequately equipped to protect vulnerable groups from violence or discrimination'. The 25 democracies are Norway, Sweden, Iceland, Denmark, New Zealand, Australia, Switzerland, Canada, Finland, Luxembourg, the Netherlands, Ireland, Austria, the United Kingdom, Germany, Malta, Uruguay, Mauritius, the USA, Japan, the Czech Republic, South Korea, Belgium, Costa Rica, and Spain.

there is a public good of inclusiveness that our society should protect. The principle of inclusiveness is based on a recognition that although we may differ on grounds of race, religion, sexual orientation, language, we are all nevertheless 'embarked on a grand experiment of living and working together despite these sorts of differences' and that each group must remember that society is not just for them, but for them too, along with others.[41] Similarly, that recognition requires that everyone in society should know that their social standing as members of the society entitles them to be treated as equals, which is what Waldron calls their dignity.[42] He is careful to distinguish this basis for the justification of a prohibition on hate speech from the principle that hate speech is there to prevent people being offended. The distinction is between the objective aspect of the standing of a person or group in society, and subjective aspects of feelings, such as hurt, shock, and anger.[43]

The principle that Waldron is arguing for seems closely related to the idea that all members of a society are entitled to be treated as of equal worth. This is an important principle that informs constitutional equality guarantees and the protection of human rights generally. It provides a different sort of counterweight to free speech to that provided by public order concerns and suggests that the prohibition on hate speech is based not so much on the preservation of the peace, but on our normative commitment to treating each person with respect. One can find hate speech provisions that seem to adopt this principle too. So for example Article 266b of the Danish Penal Code provides that 'whoever publicly ... makes statements or other pronouncements, by which a group of persons is threatened, derided or degraded because of their race, colour of skin, national or ethnic background, faith or sexual orientation, will be punished …'

I have mentioned that Waldron distinguishes his argument from a proposition that we prohibit hate speech because it may give offence, but there is a slightly different normative proposition which might also inform a hate speech prohibition and that is the proposition that we prohibit hate speech where the speaker intends to harm or hurt the group about which she or he speaks malignly. That is an idea that informs section 10 of the Promotion of Equality and Prevention of Unfair

[41] J Waldron, *The Harm in Hate Speech* (Harvard UP 2012) 4.
[42] ibid 5.
[43] ibid 106.

Discrimination Act (Act No 4 of 2000) in South Africa, which prohibits speech based on a prohibited ground (that is race, gender, sexual orientation, age, etc.) that 'could reasonably be construed to demonstrate a clear intention to be hurtful; be harmful or to incite harm; or promote or propagate hatred.' Criminal prohibitions of this sort, based on an intention to be hurtful or incite harm, are closer to our understanding of the purpose of criminal law which is to prevent intentional harm by one person to another. The criminal standard of proof, coupled with the need to show the intent to harm, may mean that successful convictions are not that easy to obtain, and may be one way to balance the relationship between freedom of speech and a public commitment to inclusiveness.

There are thus several reasons that underpin the protection of freedom of speech, on the one hand, and also several reasons that inform prohibitions of hate speech on the other. Few would argue that there is only one right way to accommodate these conflicting values. As Eric Heinze has suggested, social and political context will be important in determining how that accommodation should be reached, as will history. It is not surprising that Germany outlaws holocaust denial, nor is it surprising that South Africa has firm rules prohibiting hate speech. Reasonable disagreement across and within our societies on how these competing values should be accommodated is unavoidable. And therefore disagreement on the precise formulation and reach of prohibitions on hate speech is inevitable. The likelihood of reasonable disagreement in how to accommodate the principles of free speech with the prohibition of hate speech is one of the key challenges that needs to be considered in deciding what to do about hate speech online.

4. *Some Distinctive Characteristics of Digital Publication*

In thinking about how to deal with hate speech online, it is also important to note that there are several aspects of digital publication of information and communication that distinguish it from traditional forms of publication which may be relevant for thinking about how we might address online hate speech.

The first difference is the sheer scale of digital publication, something I have mentioned already – the quantity of digital publication dwarfs all forms of publication that have preceded it. The scale of internet publication will require any regulation of hate speech online to be able to manage a very large number of speech acts efficiently and fairly.

The second is the breadth of the publication – publication reaches across the globe, unlike traditional forms of publication that were often limited to one or at most a few jurisdictions, which meant of course that most publications could be regulated in terms of one system of domestic law. The cross-jurisdictional reach of online speech means that systems based in one jurisdiction will not address the reach of online hate speech.

The third is the fact that publication is almost instantaneous, a tweet issued by @therealDonaldTrump reaches the devices of his 47 million followers within seconds. This characteristic of online speech means that most mechanisms to constrain online hate speech will only take effect long after the hate speech has reached a wide audience.

The fourth is that most traditional methods of publishing information insert an editorial decision between author and publication, a decision that is normally taken by a person other than the author. In imposing civil liability for the publication of harmful speech, modern libel or defamation law often seek to constrain the editorial decision, as for example, in the defence of responsible publication. Such constraints are not available in relation to self-published online speech. In addition, it can probably be assumed that the editorial policy of many publications will not permit the publication of hate speech and that the insertion of an editorial decision applying that policy prior to publication will therefore often restrict the publication of hate speech. Such control is again absent in the case of direct author publication or posting on online platforms.

The fifth characteristic is that – at least for the moment – a very small group of Internet platforms or intermediaries, for want of a better description, host a very substantial portion of all Internet speech. This characteristic means that if the online intermediaries are held responsible in an effective manner for ensuring that their platforms are not used for hate speech, much online hate speech might be reached. Of course, the online market shifts fast and it may be that the dominance of a small group of platforms may wane.

I turn now to consider the manner in which hate speech online is being addressed in four major jurisdictions: the USA, the UK, the European Union and Germany. Each jurisdiction approaches the question differently and in a sense this provides us with an opportunity to reflect carefully on how we might approach the problem.

5. *Four Models for the Regulation of Hate Speech Online*

A. *The United States*

The USA, of course, has a long tradition of protecting freedom of speech under the First Amendment to the US Constitution. The approach in the USA differs in some significant ways from the approach adopted in many other jurisdictions.[44] That difference is particularly notable in relation to hate speech. As described above, the US Supreme Court has carved a very narrow exception to the protection of freedom of speech under the First Amendment to prohibit speech that incites imminent violence.[45] The prohibition on hate speech is far narrower than in many other democratic countries. The approach to freedom of speech in the USA is particularly important, because as the home of all of the giant internet intermediaries with global reach, what Timothy Garton Ash calls the private superpowers,[46] it is that approach to free speech which has informed the expectations and attitudes of those intermediaries and of many commentators in the field.

Particularly important for online speech is section 230 of the Communications Decency Act of 1996,[47] which provides what has been called a 'safe harbour' provision: Internet service providers are not to be treated as publishers of information, or speakers, in relation to any content they publish that has been produced by another person. Section 230 thus considers Internet service providers to be intermediaries not publishers and exempts them from the obligations that are imposed upon publishers of speech, like newspapers, broadcasters and publishing houses.[48] Many commentators in the USA consider section 230 to be a core component of protecting free speech.[49]

---

[44] For an analysis see E Heinze, 'Wild-West Cowboys versus Cheese-Eating Surrender Monkeys Some Problems in Comparative Approaches to Hate Speech' in Hare and Weinstein (eds) (n 21).
[45] See *Brandenburg v Ohio* (n 37), and accompanying text.
[46] See TG Ash, *Free Speech: Ten principles for a connected world* (Atlantic 2016).
[47] 47 USC s 230 'No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.'
[48] For an analysis of how §230 has played out in the US Courts, see DS Ardia, 'Free Speech Savior or Shield for Scoundrels: An Empirical Study of Intermediary Immunity under §230 of the Communications Decency Act' (2010) 43 Loyola of Los Angeles LR 373.
[49] See, the comment by M Ammori, 'The New *New York Times*: Free Speech Lawyering in the Age of Google and Twitter' (2013-2014) 127 Harvard LR 2259, 2264–2264.

There are two exceptions to the exemptions provided to Internet intermediaries by section 230. The first is a provision that relates to the publication on Internet platforms of material that infringes copyright. Under section 202 of the Digital Copyright Millennium Act, Internet service providers are exempted from liability where their platforms are used to breach copyright,[50] but this exemption is not absolute. Internet service providers bear an obligation when notified of an alleged infringement of copyright to take down the allegedly infringing material. If they do not do so, they lose their exemption from liability in relation to that material.

The second exemption is new. It came into force this year. It is an exemption under a piece of legislation known as the SESTA (Stop Enabling Sex Traffickers)/FOSTA (Allow States and Victims to Fight Online Sex Trafficking) law. The law combines two proposed amendments with bipartisan support that that had been separately introduced one in the House of Representatives and the other in the Senate.[51] The effect of the amendment is to clarify that section 230 does not prohibit the enforcement of federal and state criminal and civil law relating to the sexual exploitation of children against Internet intermediaries. It is a narrow exception that has been strongly opposed by free speech and civil liberty organisations, such as the ACLU.

The First Amendment approach, coupled with the immunity from civil and criminal liability for Internet intermediaries as provided by section 230 of the Communications Decency Act of 1996, has been endorsed by many NGOs that seek to promote freedom of speech on the internet. For example, in March 2015 a group of NGOs published the Manila Principles,[52] which propose that intermediaries should be shielded from liability for the publication of third party content. Accordingly, they should not be required to remove or block content without a court order, which must be clear and have been obtained after following due process. Furthermore, laws that restrict content must be necessary, proportionate and require due process.

---

[50] 17 USC §512: this legislation gives effect in the US to two international treaties of the World Intellectual Property Organisation (WIPO), the WIPO Copyright Treaty and the WIPO Performances and Phonograms Treaty. Title II of the Copyright Millennium Act protects online service providers from liability for copyright infringement on certain conditions, notably, where they block access to allegedly infringing materials where they are given notice of the alleged infringement by copyright holders.

[51] FOSTA HR 1865 and SESTA S 1693.

[52] See the *Manila Principles*, <https://www.manilaprinciples.org/principles> accessed on 22 February 2018.

It should be noted however that the approach in the United States does not prohibit the Internet intermediaries from regulating the content on their own platforms as they see fit.[53] Indeed, the major Internet platforms engage in extensive private rulemaking and adjudication of speech norms, with little or no First Amendment control. The consequence is that the Internet platforms can set strong restrictions on forms of speech, even speech that falls within the protection of the First Amendment.

B. The United Kingdom

In the United Kingdom, the main technique currently employed by the state for the regulation of hate speech is the criminal law. There is a long tradition of rendering various forms of speech subject to criminal prosecution. At common law there were four forms of criminal libel: seditious libel, obscene libel, criminal defamatory libel and blasphemous libel. The crimes of blasphemy and blasphemous libel were abolished in 2008, and the crimes of seditious libel, obscene libel and criminal defamatory libel are now restricted in scope and rarely give rise to prosecutions.

Although here has been a decline in the relevance of the old common law criminal offences, a series of modern statutory offences has been created. The Crime and Disorder Act establishes aggravated offences of threatening, abusive or insulting conduct on racial or religious grounds[54] and the Public Order Act establishes offences in respect of conduct that is intended or likely to stir up racial hatred, [55] or hatred on grounds religion [56] and sexual orientation. In addition, Part 1 of the Malicious Communications Act 1988 makes it an offence to send 'indecent and grossly offensive' communications with the intention of causing distress or anxiety, and section 127 of the Communications Act 2003 prohibits sending grossly offensive or obscene messages on public electronic communications networks.

In a report in 2014, the Law Commission of England and Wales noted that the 'aggravated crimes' under the Crime and Disorder Act 1998 relate only to hostility on the grounds of race and religion, and not to other protected characteristics, while the offences under the Public Order Act are concerned with conduct that stirs up hatred

[53] Ammori (n 48) 2273; M Heins, 'The Brave New World of Social Media Censorship' (2013-2014) 127 Harvard LR 325, 328.
[54] See subsections 31 and 32 of the Crime and Disorder Act 1998.
[55] See sections 18, 19 and 21 of the Public Order Act 1986.
[56] See sections 29B, 29C, 29D and 29E of the Public Order Act 1986.

on the grounds of race, religion or sexual discrimination. The Law Commission observed that it was 'undesirable' for the aggravated offences not to apply to all protected characteristics.[57] Nevertheless the Law Commission did not recommend the immediate extension of the offences to cover other protected characteristics, because of serious concerns that had been raised regarding the aggravated offences.[58] The Law Commission therefore recommended a full-scale review of the aggravated offences to establish whether the offences should be retained, amended, extended or repealed.[59] The Law Commission also did not recommend the extension of the 'stirring up' offences under the Public Order Act, noting that it was unlikely that even if the offences were extended to other protected grounds, few prosecutions were likely to result,[60] because there are few prosecutions under the existing 'stirring up' offences and so it was unlikely that extending the offences would result in successful prosecutions for the new offences. On 18 October 2018, the Law Commission announced that it would undertake a wide-ranging review of hate crime to explore how to make current legislation more effective and whether it was necessary to expand the classes of protected characteristics.

There is no reliable evidence available as to the extent of online hate speech in the United Kingdom, but it does seem to be growing. Evidence before the UK House of Commons Select Committee for Home Affairs in 2016 suggested that there had been a tenfold increase in prosecutions under Part 1 of the Malicious Communications Act between 2004 and 2014.[61] In its most recent set of crime statistics on hate crime, which includes a range of violent crimes motivated by hatred, the Home Office included an appendix containing some experimental statics concerning reports of hate crime online – the first time such statistics had been produced.[62] Because the statistics are experimental, and probably incomplete, the report suggests that they be treated with caution. It records that approximately 2% of all hate crime reported had an

[57] See Law Commission, *Hate Crime: Should the Current Offences be Extended?* (Law Com No 348, 2014) para 1.53.
[58] ibid para 1.54.
[59] ibid para 1.61.
[60] ibid para 1.68.
[61] See Written Evidence submitted by the Law Commission of England and Wales to the House of Commons Select Committee on Home Affairs, House of Commons Select Committee on Home Affairs, *14th Report Session Hate Crime: abuse, hate and extremism online* (2017, HC 609) paras 2.4 and *e* 8.
[62] See Home Office, *Hate Crime, England and Wales 2016/2017* (Statistical Bulletin 17/17), Appendix B, 'Experimental statistics: online hate crime'.

'online element', in total 1067 online hate crime offences were reported in the year ending March 2017, a number the report considered to be lower than would be expected. Of the online hate crimes reported, nearly 671 concerned race, 199 sexual orientation, 140 disability and 132 religion (some crimes are flagged as involving more than one ground).

The current legal framework for countering hate speech online is under consideration by Parliament as well. In its report on Hate Crime in April 2017, the House of Commons Select Committee on Home Affairs recommended that the government review 'the entire legislative framework governing online hate speech, harassment and extremism and ensure that the law is up to date'.[63] In addition, the Select Committee noted its concern about the manner in which social media companies were responding to hate speech online. It recommended that social media companies review their community standards and the manner in which they are implemented,[64] and called upon social media companies to publish quarterly reports on their efforts to safeguard users, including what action they had taken to eliminate illegal content, and recommending that if social media companies did not do so, that government should require them to do so.[65]

The work of the Select Committee was interrupted by the General Election in mid-2017, but in October 2017, the House of Commons Select Committee on Home Affairs launched a new inquiry into hate crime and its violent consequences, which is continuing.

C. *The European Union*

The European Union has also been considering its approach to online hate speech for some time and several of its regulatory frameworks affect the question. The first is the e-Commerce Directive,[66] which provides that Internet intermediaries will be exempt from any liability for the content on their platforms if they meet two conditions: if an intermediary becomes aware that it is hosting illegal content, it must remove it

[63] See House of Commons Select Committee on Home Affairs (HC 609) (n 61), Recommendation 15, 24.
[64] ibid Recommendation 10, 23.
[65] ibid Recommendation 13, 23.
[66] See Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market OJ L 178 .

expeditiously or disable access to it; and to be exempted from any liability, Internet intermediaries must have played a neutral and passive role in relation to the illegal content, that is they may have not generated or endorsed the content. The e-Commerce Directive also provides that member states cannot impose obligations upon intermediaries to monitor the content they manage. This obligation restricts the mechanisms available to EU member states in relation to the regulation of online content.

A second key EU provision is the European Council Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law.[67] The Framework Decision seeks to ensure that serious forms of racism and xenophobia constitute criminal offences across the member states. The conduct that is targeted is 'public incitement to violence or hatred directed against a group of persons or a member of such a group defined on the basis of race, colour, descent, religion or belief, or national or ethnic origin' and 'publicly condoning, denying or grossly trivialising crimes of genocide, crimes against humanity and war crimes as defined in the Statute of the International Criminal Court' and 'crimes defined in Article 6 of the Charter of the International Military Tribunal, when the conduct is carried out in a manner likely to incite violence or hatred against such a group or a member of such a group'. The Framework Decision requires that effective, proportionate and dissuasive penalties are imposed for such offences, and mentions imprisonment up to one year. The Framework Decision is not directed only at online hate speech of course but at all forms of speech that are in conflict with its provisions.

Thirdly, the European Commission has issued a Communication on Tackling Illegal Content Online – Towards Greater Responsibility of the Internet Platforms in September 2017.[68] The Communication set out a range of guidelines to ensure that intermediaries 'stepped up' the fight against illegal content online. Six months later, on 1 March 2018, a Recommendation on Tackling Illegal Content Online was published which in the words of the EU 'translated the political commitment of the Communication into a non-binding legal form'.[69] The Recommendation proposes that

[67] See Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law OJ L 328.
[68] See European Commission, *Communication on Tackling Illegal Content Online* (n 8).
[69] See European Commission, *Recommendation on Measures to Effectively Tackle Illegal Content Online* (n 8).

Internet intermediaries establish notice-and-action mechanisms on their websites whereby users can draw the attention of intermediaries to illegal content online so that intermediaries can remove or block access to content that is illegal. The Recommendation provides that the person who posted the content should be informed of the decision by the intermediary and given an opportunity to respond to it.

A further important EU initiative is the EU Code of Conduct on Countering Illegal Content Online. The Code of Conduct was originally consented to by four major Internet platforms (Facebook, YouTube, Twitter and Microsoft) in June 2016. In terms of the Code of Conduct, the platforms agreed to review requests to remove content that is in breach of their community standards or illegal within 24 hours. Since then, four more Internet platforms (Google +, Instagram, Snapchat and Daily Motion) have agreed to implement the Code of Conduct.

Regular reports of the steps taken by the companies are published. The third evaluation of the implementation of the Code of Conduct took place over a six-week period in November and December 2017, and was published in January 2018. The report showed that 2982 notifications were received by the IT companies in the reporting period and that 70% of the content that was the subject of complaint was removed by the companies, with 80% of complaints being dealt with within 24 hours. 511 cases were referred to the criminal justice authorities.[70]

It is interesting to note that the Code of Conduct requires the platforms to remove not only illegal content but also content that is in breach of the platforms' own community standards. Facebook has recently updated its community standards, which provide very detailed descriptions of what is prohibited on the platform.[71] Its community standards state:

> We define hate speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some

[70] European Commission, 'Code of Conduct on countering illegal hate speech online: Results of the 3rd monitoring exercise' (Fact sheet, January 2018)8) <http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=612086> accessed on 30 September 2018.

[71] See the most recent iteration of its community standards here: Facebook, 'Facebook Community Standards' (*Facebook*, 2018) <https://m.facebook.com/communitystandards/hate_speech/> accessed 30 September 2018.

> protections for immigration status. We define "attack" as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation.[72]

The Facebook community standards are implemented by staff members who consider complaints received by Facebook about posts on the platform. By requiring the intermediaries to remove content that is in breach of their own community standards as well as EU law, the EU is placing the obligation to remove hate speech directly onto the intermediaries themselves, and to some extent upon the Internet users, given that it is complaints that initiate the possibility of removal. Consistent with the e-Commerce directive, there is no obligation placed upon the platforms to monitor content uploaded to the platforms for consistency with the law or their own community standards. Their obligation is to provide a complaints procedure and to enforce their standards once complaints are received.

D. *German Netzgesetz*

The German government has introduced an important law to regulate online speech including hate speech, the *Netzwerkdurchsetzungsgesetz* (Network Enforcement Law), or as its long title is: Law for the Improvement of the Legal Regulation of Social Networks. The Act came into force on 1 October 2017 and full operation on 1 January 2018.

The Act does three things. First, like the EU Code of Conduct, the *Netzgesetz* requires Internet platforms[73] to adopt effective and transparent procedures to handle complaints about illegal content being published on their platforms.[74] The process must ensure that complaints are immediately considered and, where the content is found to be illegal, it must be blocked or deleted.

What the law calls 'manifestly illegal' (*offensichtlich rechtswidrig*) content must be removed or blocked within 24 hours of the receipt of the complaint,[75] and

---

[72] ibid.
[73] See s 1 of the Law which defines the bearers of obligations under the Act as providers of for profit telemedia services that operate internet platforms intended to allow users to share content with other users or with the public. Excluded from the operation of the law, are platforms which produce edited content that are the responsibility of the service provider, or platforms that offer individual communications or specific content, and, in effect, platforms that have fewer than two million registered users.
[74] Network Enforcement Law s 3(1).
[75] Network Enforcement Law s 3(2)(2).

other illegal content must be deleted or blocked within seven days, subject to some exceptions. [76] Both complainants and content generators must immediately be informed of the decision on the complaint, and reasons for the decision must be provided.[77] The Act also permits platforms to refer complaints to a recognised self-regulation institution, as long as the platform agrees to accept the decision of that institution.[78]

In determining what content is illegal, the Act relies on a raft of provisions in the Criminal Code.[79] These criminal prohibitions include:

- incitement to hatred against a national, racial, religious or ethnic group, or segments of the population or calls for violent or arbitrary measures against such groups, or assaults on the human dignity of others by insulting, maliciously maligning or defaming segments of the population in a manner capable of disturbing the public peace;[80]
- dissemination of depictions of cruel or inhuman violent acts against human beings in a manner that glorifies or trivialises the act or violates human dignity;[81] and
- defamation of beliefs or religious or ideological organisations in a manner that threatens the public peace.[82]

The German Networks Act thus requires Internet platforms to establish a prompt procedure to investigate complaints that content published on their website may constitute a criminal offence under any of these provisions. If the decision that follows is that the content does constitute unlawful speech, the network must block or delete the content within 24 hours if the speech is 'manifestly unlawful' or otherwise within seven days. Deleted content must be kept for ten weeks.[83] The core obligation under the German Networks Act is to create an effective and transparent process that is always available to users.[84]

---

[76] Network Enforcement Law s 3(2)(3).
[77] Network Enforcement Law 3(2)(5).
[78] Network Enforcement Law ss 3(2)(3)(b) and 3(6).
[79] Network Enforcement Law s 1(3).
[80] German Criminal Code s 130.
[81] German Criminal Code s 131.
[82] German Criminal Code s 166.
[83] Network Enforcement Act s 3(4).
[84] Network Enforcement Act s 3(1).

Secondly, failure to establish and implement a compliant process for handling complaints may lead to the imposition of fines of up to €5 million. Should the German administrative authorities wish to levy an administrative fine on a network because a network has failed to delete or block a particular example of illegal content, the authorities must first obtain an order from a competent court that the content is indeed illegal.[85] The fines are thus primarily aimed at systemic failure to provide a complaints process, rather than at a failure to remove an individual post..

The third key aspect of the Act is its imposition of a reporting duty on networks. Network providers that receive more than 100 complaints of unlawful content in a calendar year must produce a biannual report concerning the handling of the complaints, which must be published in the Federal Gazette and in an accessible manner on their homepages.[86] The reports must include information on the number and type of complaints, as well as the number that result in the deletion or blocking of content, the ground for the deletion or blocking, and the time that elapsed from the lodging of the complaint to the deletion of blocking of content. In addition, reports must contain information concerning the technical and language competence of those handling complaints, as well as the manner in which those handling complaints are trained and supervised.[87] These reports, once they are published, will provide an important basis on which to assess the operation of the Act.

E. Assessment

These are four different systems for tackling the question of online hate speech. The approach in the United States is to impose no obligation upon Internet platforms to remove online hate speech. This approach permits Internet platforms to host hate speech without legal consequence. However, and somewhat counter-intuitively, the approach in the United States also permits Internet platforms to design their own systems to regulate content online, even where that regulation may result in removing content that may be protected under the First Amendment. Many of the platforms have developed elaborate systems of control, which include limitations on hate

[85] Network Enforcement Act s 4(5).
[86] Network Enforcement Act s 2(1).
[87] Network Enforcement Act s 2(2).

speech. Those systems of control are not subject to any oversight or monitoring and there is accordingly no guarantee that freedom of expression is properly protected by the platforms.

In the United Kingdom, the European Union and Germany, governments are paying more attention to what is posted on line and are seeking to regulate the question in different ways. In the United Kingdom, the key approach at the moment is to impose criminal sanctions on some forms of hate speech, although the range of criminal prohibitions is under review by the Law Commission. It is an open question whether criminal prohibitions, coupled with careful prosecutorial decision-making, will effectively address hate speech online, without unduly restricting freedom of speech. It is clear from the proceedings before the House of Commons Select Committee on Home Affairs that Parliament is considering whether it is appropriate to supplement the existing legal framework with statutory obligations that would impose obligations upon social media platforms to take steps to remove hate speech online.

So far, it is only Germany that has imposed a statutory obligation upon social media companies to establish a mechanism to consider complaints about (amongst other things) hate speech on their platforms and where the complaints are found to have merit, to remove the offending content. The European Union appears to be moving in a similar direction.

There are some clear difficulties with the approach adopted in Germany. Legitimate concerns have been raised that there is a risk that platforms will 'overblock' content, that is, that they will block or delete content that is not unlawful, with deleterious implications for freedom of speech.[88]

It is a pity that the Networks Act does not require Internet platforms to maintain a publicly accessible archive or database of the complaints lodged, the content that gave rise to the complaints, and the decisions on the complaints as well as the reasons for the decisions. Granular analysis of the complaints handling process

[88] See, for example, the response of Human Rights Watch, which argues that the law is overbroad and turns private companies into overzealous censors. Human Rights Watch, 'Germany: Flawed Social Media Law: NetzDG is Wrong Response to Online Abuse' (*Human Rights Watch*, 14 February 2018) <https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law> accessed on 31 October 2018.

would enable an assessment of the operation of the Act and in particular of the quality of the decisions being made in relation to complaints.

There are other risks too. In particular, there is a risk that there may be uneven, arbitrary and even discriminatory decisions on complaints, so that content may be blocked on one platform and not others, or that content by particular groups may be subjected to more stringent blocking and deletion than that of other groups. Again without access to the underlying data relating to the complaints, it will not be possible to determine whether the complaints systems established by the platforms do result in discriminatory or arbitrary decision-making. It is accordingly not surprising that the German legislation has attracted disapproval.

On the other hand, there are reasons why requiring social media platforms to establish effective complaints systems may present a workable way forward: the factors that differentiate online publication from other forms of publication – its scale, breadth, speed, as well as the absence of a third-party editorial eye – all point to the practicability of complaints systems managed by the host platforms themselves. The large platforms are likely to have the capacity to handle the scale, breadth and speed of publication better than an institution established by the state.

But there are real challenges to ensuring that such complaints systems strike an appropriate balance between the protection of freedom of speech and curbing hate speech. Decisions as to what content should be blocked or removed will always be difficult and contested. As explained above, determining where to draw the line in relation to hate speech, involves accommodating the values that inform freedom of speech as well as the values that suggest that we should restrict hate speech.

Given the number of such decisions that will have to be taken on a daily basis, it is likely that the decisions will be somewhat rough and ready, and there will be a real risk that freedom of speech will not be given sufficient weight in the decision-making process.

External scrutiny of the platform's decisions may assist in ensuring that a proper balance is found. There are perhaps two ways in which external scrutiny could take place. First, an appeals system could be established which would enable those dissatisfied by a decision to block (or not to block) to have the decision reviewed. Second, a system for external review could be established in which a random sample of decisions could be scrutinised and assessed. The regular review of an adequate sample basis might serve to improve the quality of the primary decision-making.

In conclusion, it remains open to debate whether any of the four systems reviewed here is striking an appropriate balance between the protection of freedom of expression, on the one hand, and prohibitions on hate speech, on the other. It is not surprising then that the question remains under review in all four of the legal systems – and the question should remain under the close review of scholars as well.

6. *Conclusion*

The extraordinary scale and reach of publication on the Internet has introduced a profound challenge for legal systems that seek to curb hate speech without unduly impairing freedom of speech. In determining how to address hate speech it is necessary to balance the range of normative principles that underpin our commitment to freedom of speech with the purposes which underpin prohibitions on hate speech. The tension that exists between these two goals has been reflected in international human rights law since the 1960s and inevitably there is reasonable disagreement as how best to strike the balance. Accordingly, it is difficult to envisage one set of rules being developed that will be accepted everywhere. Although each of the four jurisdictions reviewed here is paying attention to the question of how to counter hate speech online, it is not yet clear that a satisfactory approach has yet been devised. The mechanisms recently introduced in the EU and Germany constitute novel and interesting attempts to require Internet platforms to prevent the publication of unlawful speech on their platforms, as well as speech that breaches their own community standards. Whether they can be implemented in a manner that does not unduly curb freedom of speech remains to be seen.