The role of social media companies in the regulation of online hate speech

Chara Bakalis, Principal Lecturer in Law, Oxford Brookes University

cbakalis@brookes.ac.uk

Julia Hornle, Professor in Internet Law, Queen Mary University, London

j.hornle@qmul.ac.uk

The role of social media companies in the regulation of online hate speech

Chara Bakalis, Principal Lecturer in Law, Oxford Brookes University

Julia Hornle, Professor in Internet Law, Queen Mary University, London

Abstract

This article is about online hate speech propagated via platforms operated by social media companies (SMCs). It examines the options open to states in forcing SMCs to take responsibility for the hateful content that appears on their sites. It examines the technological and legal context for imposing legal obligations on SMCs, and analyses initiatives in Germany, the UK, the EU and elsewhere. It argues that whilst SMCs can play a role in controlling online hate speech, there are limitations to what they can achieve.

Key words: hate speech, hate crime, internet regulation, social media regulation, online hate speech, cyberhate

1. Introduction

This article is about online hate propagated via platforms operated by social media companies (SMCs), and it examines the options open to states in forcing SMCs to take responsibility for the hateful content that appears on their sites. The focus will be on the US and Europe. The article explains the dilemma we face if SMCs are to be held responsible for user-generated content, particularly with respect to balancing freedom of expression with the need to offer protection from hate speech. However, this dilemma is not examined through a human rights law analysis by balancing specific freedom of expression restrictions with harms. Instead we examine the legal obligations and responsibilities imposed on social media companies as internet intermediaries in the wake of recent legislative initiatives in some EU Member States. In particular, we chart how the approach has changed from specific notice and take-down obligations to greater responsibilities of proactive measures. We contrast this "European" approach with the US approach under the first Amendment which would prohibit the imposition of such responsibilities.

This article argues that regulation in Europe is moving away from giving SMCs as internet intermediaries immunity from liability for illegal hate speech towards a new approach forcing them to take responsibility for user-generated content, and imposing a range of obligations on them to proactively moderate and manage content on their sites. Furthermore, we argue that as a matter of principle, this approach can be made compliant with freedom of expression obligations under the European Convention on Human Rights (ECHR) in a way that might not be possible under freedom of speech rules under the US First Amendment. However, we point to the concern that some of the pro-active measures, to the extent that they automate content moderation and management, do create particular concerns over freedom of expression, and therefore require particular attention.

The growth of online hate has been exponential over the last few years (O'Regan, 2018). Although we do not have official statistics that can give us an accurate picture of the actual amount of online hate, several recent studies have found alarming levels of abuse. For example, the Anti-Defamation League (2019) found that 37% of Americans had suffered online harassment, and that a third of these cases were as a result of the target's protected characteristic such are race, religion, gender identity, sexual orientation or disability. A report by Amnesty International (2019) found that an abusive or problematic tweet was sent to a female politician every thirty seconds and that black women were 84% more likely than white women to receive abusive tweets. Meanwhile, a Canadian survey found that 60% of Canadians had viewed hate speech online (Association for Canadian Studies, 2019).

There are a number of factors, however, that make the regulation of online hate particularly difficult (Bakalis, 2017; O'Regan, 2018). For example, the sheer scale of the amount of online hate, the pseudonymity afforded by the internet, and jurisdictional issues when the perpetrators of hate and their victims do not live in the same country make this a very difficult area to police. Particularly controversial is the issue of free speech which creates difficulties in introducing legislation to prohibit this sort of material. In spite of these difficulties, governments across the world are coming under increasing pressure to do something about the problem, particularly as it is becoming progressively more obvious that this is a particular problem for minority groups.

More recently, the focus has shifted onto social media providers and their responsibility for contributing to the dissemination of online hate speech (Cohen-Almagor, 2015; Laidlaw, 2015). Politicians and social activists have called on SMCs to "do more" to prevent the spread of hate speech, abuse and extremist content on their platforms. While the discussion in the early 2000s mainly focused on the question of technological innovation and immunity for intermediaries, and a narrow tailored notice and take-down obligation, recently the debate has called for greater SMC responsibility, and concomitant with this, pro-active and much more extensive obligations to manage and monitor content (Frosio, 2018).

The focus on SMCs has come about because of the growing realisation that policing online hate by law enforcers is virtually impossible because of the sheer amount of hate that appears online, and the recognition that SMCs are, therefore, much better placed to deal with this because they have a degree of technical control over their platforms.

However, placing the responsibility on social media providers brings with it its own problems, and is not necessarily the quick, easy and cheap solution that politicians may hope for, particularly in relation to any proposal which aims to automate the process of removing hateful material.

Currently, in the US and Europe, SMCs are operating in a sphere that was predicated on the ideals of freedom from governmental regulation and laws. Since the early days of the internet and the world-wide-web, there was an acknowledgment that platform providers cannot be treated akin to offline publishers of the information they allowed to appear on their platforms, as they lack control over the content itself (Murray, 2016; Bridy, 2018). The Communications Decency Act 1996 (CDA) was enacted in the US in order to give protection to service providers from being treated as publishers or distributers of the content they hosted. This was followed in 2000 at the EU level by the E-Commerce Directive 2000//31/EC. While the CDA gives absolute immunity to publishers (other than immunity from Federal criminal law), Article 14 of the E-commerce Directive bases immunity for hosting providers on a knowledge standard. Thus SMCs are shielded from liability only if they do not know (for example through notification or constructive knowledge) that they are hosting illegal content. However, the E-Commerce Directive states that no general obligation can be imposed on SMCs to monitor their platforms.

In this context, the main way in which the pressure on SMCs has manifested itself has been in the form of voluntary codes of conduct such as that set up by the Working Group on Cyberhate convened by the Anti-Defamation League in the US, and the EU Voluntary Code of Conduct (2016). However, in the last two years political pressure has been growing and SMCs have been called upon to "do more". Consequently, we have seen legislative initiatives moving in two directions: one type of legislation imposes standards for the speed and quality of notice and take-down, and the second type of initiatives have moved from mere take-down obligations to imposing a range of pro-active measures.

Thus, in Europe, the tide appears to be turning, and governments are actively rethinking regulation. For example in Germany, politicians were impatient with the apparent lack of action by SMCs in taking down content that is illegal according to German law. As a consequence, the Network Law Enforcement Act[1] was enacted in 2017 which seeks to impose a legal obligation on internet platform providers to act swiftly to remove hateful material from the internet. The French Parliament is also currently considering legislation that would mirror that of the German law. Initiatives have also been taken in the UK and at the EU level where greater responsibility on SMCs is envisaged, and which will be discussed further below.

It will be argued that whilst there may be a good case for imposing some of the burden for the regulation of cyberhate on platform providers, in reality they are limited in what they can do. We also have to be careful that any law requiring SMCs to remove or block certain types of online speech does not unintentionally confer on them too much power over what can and cannot be said online. Instead, what is needed is a proper discussion about the regulation of cyberhate, an acceptance that SMCs are limited in what they can do, and the acknowledgment that in fact a multi-faceted approach is required.

[1] Netzwerkdurchsuchungsgesetz, NetzDG

The first section of this article will outline what technological possibilities are open to SMCs to control hateful content. It will be argued that they are better placed to remove online hate than the police, but that there are real limitations to their ability to do this. There are also problems with requiring SMCs to proactively monitor content. In particular the use of automated content moderation and reliance on private regulatory regimes may mean that perfectly legal content is taken down. The second section will locate the regulatory options within free speech concerns and will argue that the approach adopted by each country needs to reflect the legal norms of each jurisdiction. The final section will look at current developments in a number of European countries and at the EU level and place them in the context of freedom of speech.

To preface this discussion, two definitions need to be made at the outset.

The definition of hate speech is contested but, for the purposes of the argument in this article, it will include content which is illegal under legislation aimed at outlawing speech that incites violence, hatred or discrimination against named groups. We take 'hate speech' to refer to a narrow category of material that is illegal under the law, thus distinguishing it from material that might express hateful content, but which is not in fact illegal. It is also important to distinguish hate speech from general hate crime provisions which, at their most simplistic, can be defined as crimes that deal with behaviour that is already recognised as criminal under the law (such as assault), but which are aggravated because of the perceived hostility of the perpetrator against the victim based on their affiliation to a particular group. By contrast, hate speech provisions are ones which criminalise *speech* on the basis of its hateful content against certain groups.

Social media have been defined (Boyd & Ellison, 2007) as "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and transverse their list of connections and those made by others within the system."

We define SMCs for the purpose of this Article as providers of a platform environment which allows users to upload content ("user-generated content") in order to share and communicate this content with other users (whether they are a restricted group of contacts, everyone registered on the platform, or more generally with users searching content online). This definition includes major providers such as Facebook, Instagram, Snapchat, WhatsApp, Twitter and Youtube, but also smaller providers. We acknowledge that there is a need to define more clearly between, on the one hand, messaging services, whose main purpose is communication among a limited circle of private users, and on the other hand, content-sharing services whose main purpose is the sharing and distribution of multi-media content which originates with users of the service. We adopt this wide definition not because of a normative argument about the scope of legal regulation, but for the reason that some of the laws and regulations discussed below have such a wide scope, and for the reason that this article canvasses the issues widely. The focus of this article is not a classification of different types of services and their functions, although we acknowledge that more research and conceptualisation is needed for the regulatory debate.

What defines SMCs is that their users entirely determine the user-generated content, but the SMCs are in control of arranging the content through metadata and online profiling, and they control the methods and format of the users' interaction (for example by having a wall with posts or "likes" of content).

2. What *Can* SMCs Do to Prevent Online Hate Being Disseminated?

In order to be in a position to evaluate the recent initiatives in the field of online hate and platform liability, it is important first to outline what exactly SMCs *can* do from a technological and logistical point of view to curb the mass of online hate on their sites. By delineating the parameters of what SMCs can do, it will be seen that there are real limitations to, and dangers *inherent* in, the technology available to them that are sometimes ignored by politicians when they demand that SMCs should "do more" to control online hate.

In answering this question, it is crucial at the outset to make a distinction between re-active content moderation and pro-active moderation. With reactive moderation, content is only taken down after a complaint has been made to the SMC of a potential breach of community guidelines or law. This is also known as notice and take-down. Pro-active content moderation is where material is prevented from being posted, and *before* it has been notified by anyone.

2.1 Re-active Notice and Take-Down

Notice and Take-Down is essentially a reactive form of content moderation whereby SMCs react to notification by users, or organisations they work with, and take content down, or close accounts, groups or channels. Given that SMCs may be liable under the applicable national law unless they take down material expeditiously once they have been notified by their users, a number of SMCs, and in particular the tech giants, have set up notice and take-down systems and procedures.

Facebook has stated publicly that it had 7,500 reviewers in 2018 and had plans to double the members of staff working on safety and security to 20,000 before the end of 2018 (US Senate Committee, 2018). Youtube has been running a "flagging system", whereby content flagged by users is reviewed. In addition, Youtube has developed a “trusted flagger programme” which is a community of trusted users who have a track record of flagging content accurately, according to Youtube’s content guidelines (US Senate Committee, 2018). Youtube has described its trusted flaggers as organizations with specialist expertise, for example, in hate speech and terrorism, and that it expanded its trusted flagger programme by an additional 50 NGOs during 2017. It stated that it would have 10,000 persons working to fight content which violates Youtube content guidelines in 2018, and that it removed 70% of violent extremism videos within eight hours of uploading (US Senate Committee, 2018).

One of the main challenges of notice and take-down is that for some types of content and communications this mechanism is too slow, even if take-down takes place within a few hours of notification. For example, on Twitter conversations develop and escalate quickly, and the impact of the content occurs very soon after the tweet has been published - practically in real time (O’Regan, 2018). Hence, for Twitter, the action it takes on notification is to close accounts, and it stated that until January 2018 it had closed 1.1 million accounts which it had classified as terrorist accounts. However Twitter also admitted that around 5% of its approximately 300 million accounts are fake accounts, many of which are automated bot accounts (US Senate Committee, 2018), which turns the take-down process into a constant fight against the hydra monster of ancient Greek mythology.

But even assuming that SMCs take down content quickly and according to clear guidelines - say within one hour or 24 hours as has been suggested in the EU Proposal for terrorist content or as is the case under the German Network Law Enforcement Act - its negative impact may nevertheless already have been considerable as 1000s, if not 100,000s of users may have seen the content and it may already have been copied, reposted or retweeted to other corners of the internet, including smaller SMCs and hosting companies or companies who refuse to take action against illegal content (Commission Staff Working

Document, 2018). In particular certain SMC applications, such as Facebook Live (live streaming of video) have led to irreversible online harm at the instant that the content is published. For example the filming of the terrible terrorist attack on mosques in Christchurch in New Zealand, which, even though it was taken down within an hour of upload, had already been viewed 4,000 times before it was removed (BBC News, 2019). The Prime Minister of New Zealand, Jacinda Ardern, has stated that she wants the Facebook Live facility to be changed, for example by incorporating a delay before the stream goes live, and she is urging G7 countries to take action to mandate such a restriction (BBC News (2), 2019).

Furthermore, an investigation by the German newspaper, Süddeutsche Zeitung (SZ) back in 2016-7 revealed how these content moderation systems work in practice: Facebook for example receives 46 million take-down requests a week, which means that its 4000 reviewers have about eight seconds to make a decision whether to take content down (Süddeutsche Report, 2016). The reviews are outsourced to several service companies across the globe, where the reviewers usually work just above the minimum wage, and after two weeks of training have to review around 2,000-3,000 pieces of content a day, some of which is so heinous that it leaves them traumatised with the outsourced service company providing little in terms of psychological support (Guardian News, 2018). Some of the content moderators in their interviews with the SZ admitted that the time pressure and the nature of the materials is such that they have given up looking at the pictures properly. There are, therefore, serious questions to be asked about the protection of the employees of the outsourced services, and also about the quality of the notice and take-down process and decision-making (The Cleaners Documentary, 2018).

Considering how complex and context-specific the assessment is, and considering that editorial decisions require careful deliberation, the take-down process has rightly been criticized even though the major SMCs have employed more staff and are working on improving their processes (US Senate Committee, 2018).

Moreover, there are questions about the transparency of the internal rules and guidelines made by the SMC. These internal guidelines, according to which moderators take content down, were originally secret, but were leaked by the UK newspaper, the Guardian, in 2017. In 2018, Facebook published its community standards in response to that leak (Guardian News, 2018). Such transparency has also been demanded by the EU Commission's Communication on Tackling Illegal Content Online (EU Commission, 2017). Therefore, while notice and take-down is established as a mechanism, it continues to involve substantial challenges. However, even greater challenges are inherent in pro-active content moderation by SMCs, which we turn to next.

2.2 Pro-active Prevention of Dissemination

The second way in which SMCs can control their user-generated content is through pro-active prevention or dissemination where content is blocked as it is being posted and before it has been notified by anyone. Frequently, technology aids this process and the temptation for politicians is to call for automation through the use of artificial intelligence.

Owing to the sheer quantity of information posted and uploaded by users on social media every second (Youtube famously quoted the hours of videos uploaded every second as being 400 hours) and every day (there are an estimated 5 million tweets every day), it is impossible to monitor content or exercise editorial responsibility on a manual basis. Therefore, the only realistic way in which content can be proactively removed is through the use of artificial intelligence which filters material at super-human speeds and identifies and blocks material that is deemed unacceptable by community standards, or illegal under national laws.

Algorithmic tools used for content filtering are extremely limited in what they can identify as the subject matter of a video, an image or a text, and are by themselves not yet matured to distinguish between lawful and illegal content (Ammar, 2019; UN Special Rapporteur, 2018). An illustration of this is that automated tools have difficulties distinguishing between an image of a medical operation and that of an execution, or to distinguish between different meanings of the same word (think for example of the Russian feminist activist band "Pussy Riot"). Furthermore algorithmic tools currently cannot understand the context of the information before them, and therefore find it hard to pick up on parody, satire, irony or jokes. They also cannot recognise the context where a user actively and explicitly criticizes an image or where they have reposted a quote. Because of this, there is a risk that automated tools may lead to the removal of counter-speech aimed at hate or even terrorist speech in a counter-productive way (Frosio, 2018). Since the legality or illegality of speech frequently turns on context, this makes automating the legal assessment extremely challenging, if not impossible. Moreover, for some types of speech the legal assessment depends on whether the information is factually true or not, so that extraneous information must be sought before a decision can be made. Finally, there is a risk that if content is taken down by automated tools without human review, important evidence of crime or items of news reporting are made unavailable to investigators or security services. Therefore, automated detection of new illegal content and its classification at present requires human verification (Commission Staff Working Document, 2018).

Despite these shortcomings, the large SMCs have invested in automated content recognition and blocking technology, and so have governments, particularly in the context of material relating to terrorism (Wired, 2018).

Facebook has stated that it proactively uses algorithms for text-based machine learning and hashes for matching images which have been previously identified as illegal online extremism (US Senate Committee, 2019). Once a terrorist video or image has been identified as illegal, such known content is taken down within one hour.

It also stated that this automated technology proactively discovers more than 99% of Al Quaeda and IS propaganda material online before it was notified to Facebook (US Senate Committee, 2019). Likewise, Youtube stated that it has invested in machine learning technologies and uses a classification system which pro-actively flags videos for human review as potentially extremist hate speech, and that this has enabled Youtube to remove nearly five times as many videos. Youtube also uses image-matching techniques which prevent the re-upload of extremist videos (US Senate Committee, 2019). Similar to the figures quoted by Facebook, Youtube stated that 98% of videos taken down were initially identified by algorithms, not notification. Twitter also stated that it has developed technology automating the recognition of terrorist accounts before they are reviewed by a human reviewer, and that in 2017, 90% of terrorist accounts were identified by these automated tools and 75% of these accounts were closed before anything was tweeted from them (US Senate Committee, 2019). The combination of artificial intelligence and human review has increased the quantity of illegal content removed and has sped up the process.

Facebook, Youtube and Twitter have also invested in counter-speech initiatives such as the ‘Peer-to-Peer Challenging Extremism Programme’ and the ‘Creators for Change Programme’. These initiatives address the filter-bubble silo problem whereby website algorithms target content to users based on behavioural online profiling and leads to users being caught in content which is highly selective and isolating (Pariser, 2011). They specifically target content critical of violent extremism and containing counter-narratives to users who seem interested in violent extremism and terrorist content (US Senate Committee, 2019).

However, a number of services exist (such as Telegram) which offer encryption to their users, which means that they cannot deploy automated content monitoring. This makes the pro-active detection of illegal content impossible. Moreover automated monitoring is challenged by the fact that terrorist organisations and organisations which spread online hate change their tactics and online strategies in such a way that it is more difficult to automatically recognise such content as online hate or terrorist content.

Finally, Facebook, Youtube and Twitter have a shared database of hashes of known terrorist images and videos, which they use to filter uploads, thus preventing this content to be spread across their platforms. This technology creates a unique hash function or digital fingerprint against which other images and videos can be compared. Digital fingerprinting was first used for this purpose in the context of images of child abuse by the National Center for Missing and Exploited Children which uses a technology called PhotoDNA to find known images of abused children. It has also been used for preventing the dissemination of images or music videos (for example Youtube's Content Id) that infringe copyright. Now SMCs are deploying this technology to ensure extremist images and videos (for example beheadings or propaganda lectures) are removed and stay down (US Senate Committee, 2019).

Thus, whilst AI can certainly be of some help in pro-actively identifying and removing illegal hate speech, it is clear that there are real questions over whether and how this kind of proactive content monitoring can be or should be achieved. There are two main arguments against automated, pro-active filtering without human assessment. First, there is the argument that this type of pro-active filtering is ineffective, or even, in some instances, counter-productive (Ammar, 2019) . Such filtering may be counter-productive as it would remove content uploaded to steer would-be-terrorists away from extremist content. Secondly, there is the danger of the removal of material that is legal, and thereby infringing freedom of expression through excessive censorship.

Summing up, it is clear that that there is a huge quantity of heinous content about which users complain, which SMCs have enabled, and which they now find difficult to control. There is a financial burden attached to this, but for the largest SMCs at least, it seems only fair that they plough more of their huge profits into protecting both users and their moderators, and that they share some of the resources with smaller or not-for-profit SMCs.

The call for SMCs to take greater responsibility for policing their platforms has had some success in improving both notice and take-down, and has also led to investment in automated tools which can ensure the stay-down of images and videos, and can pro-actively detect online hate, and in particular terrorist content. This has sped up detection. However, it is equally clear that artificial intelligence tools cannot completely automate detection and prevent the upload of illegal online hate in the foreseeable future, because of the context sensitivity of such materials and changing strategies of groups propagating such materials. While it is equally clear that using such tools may speed up the process of detection and action against known types of content, actual removal nevertheless requires human review in many cases. There is, therefore, still a question mark over whether it is indeed possible to gain control over the sheer overwhelming quantity of hate materials, and the use of sophisticated technology such as bots posting such material, or encryption by groups propagating hate.

Politically it is convenient to call for artificial intelligence, machine learning and other technology to solve the problem, but the danger is indeed that this call brushes under the carpet the real complexity of the issues involved, including undermining freedom of expression and the question to what extent removal of content may have unintended, counter-productive side effects. Therefore it is important to keep in mind the need to build in safeguards, such as demanding that content automatically detected is reviewed by a human reviewer, and demanding quality standards as to the training and support of such human reviewers.

3. Putting the Regulatory Approaches into the Context of Internet Free Speech

Having examined the technological capabilities of SMCs, this next section will situate the regulatory approaches into the context of free speech. Regulation of online hate has often been opposed because of concerns about free speech. This section will examine these concerns and will argue that the approach adopted by a state should be guided by its own cultural and legal stance on hate speech as well as by broader questions over internet regulation. This is an important insight as the debate in this area has largely been driven by US First Amendment considerations. This has distorted and derailed the debate, particularly in European countries which have established hate speech laws that do not align with the US approach on hate speech. It is crucial that the US-bias in the debate is recognised in order for the discussion in this area to develop and evolve in a way that is more consistent with the cultural and legal norms of each individual country or region.

Until relatively recently, the default position in relation to hate speech has been to avoid enacting any binding legal obligations on SMCs to remove hateful material from their platforms. This default position has been based partly on the concept of 'cyberlibertarianism' which is the school of thought that believes that our concepts of traditional state sovereignty do not work in the virtual world, and so regulation of the internet is impossible and futile (Johnson and Post, 1996). But it is also partly shaped by the US First Amendment view of the issue which does not necessarily fit with the legal norms and culture elsewhere in the world (Belliveau, 2018).

In the US, freedom of speech is guaranteed under the First Amendment of the Constitution which states that 'Congress shall make no law…abridging the freedom of speech…'. (First Amendment). Supreme Court jurisprudence has finessed and delineated the parameters of the right to free speech. In relation to hate speech, the Supreme Court has ruled in a number of cases that hate speech is protected free speech, and states can only prohibit speech if it incites 'imminent lawless action' (*Brandenburg v Ohio*, (1969)). In the case of *R.A.V. v. City of St Paul* (1992), the Supreme Court confirmed that any rules which prohibit the content of speech (such as hate speech) are unconstitutional. It still remains possible, for states to prohibit speech if it constitutes 'fighting words' and thus incites violence. It is, however, unconstitutional for states to create laws that prohibit any speech based purely on its hateful content.

In addition to this, the Supreme Court in *ACLU v Reno* (1997) made it clear that internet forums and internet communication would not be subject to regulation in the same way as the mass media. This case struck down as unconstitutional elements of the CDA which tried to limit the type of material that could appear on the internet to that which was 'decent' because to do so interfered with First Amendment rights. The more recent decision of *Packingham v North Carolina* (2017) confirms that SMCs are viewed as a protected area for free speech. Whilst this decision has been criticised (Citron and Richards, 2018), it remains the law that SMCs cannot be subjected to legislation which purports to limit speech, and thus infringe First Amendment rights.

From a US standpoint, the question of whether to require SMCs to remove hate speech is fairly straightforward. As the US does not have hate speech laws, coupled with the CDA provisions which grant SMCs immunity from liability as confirmed by *Packingham v Carolina* (2017), for the US government to refuse to impose a requirement on SMCs to remove hate speech, tallies perfectly with the US legal approach to free speech (Belliveau, 2018). Although some US academics have put forward arguments in favour of hate speech restrictions (Belliveau 2018, Waldron, 2012) and in favour of online regulation (Bridy, 2018, Keats Citron and Wittes 2017), the situation remains that under current laws, governmental regulation of online hate speech is unlawful. Given that most major SMCs are originally based in Silicon Valley, it stands to reason that US cultural and legal assumptions about free speech will predominate. This is why the starting point for most debates on regulating online speech has been framed by free speech concerns.

However, the starting point from a European perspective is different. Under Article 10 of the European Convention on Human Rights (ECHR), it is possible for a State to create a law that imposes a limit on our freedom of expression so long as under Article 10(2) this law is:

… necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others …(Art 10, ECHR).

In relation to hate speech, the European Court of Human Rights (ECrtHR) has developed a body of case law which outlines to what extent States can deviate from the basic principle of freedom of expression. There is a line of cases that has advanced a relatively low level of protection for expression that has incited hatred against minorities (*Pavel Ivanov v Russia*, 2007) and gives States wide discretion when it comes to criminalising or prohibiting such behaviour. Although there has been criticism of the ECrtHR's approach because it appears to give less protection to some minorities compared to others, the basic point - that hate speech laws are *prima facie* legitimate - still stands.

Whilst the ECHR does not set out a definition of hate speech, and neither does it compel the enactment of hate speech laws, it has gone as far as recommending that signatory countries review their domestic legislation to ensure that it complies with the need for hate speech provisions, and urges signatories to ratify the International Convention on the Elimination of all Forms of Racial Discrimination, which under Article 4 requires countries to outlaw speech that aims to incite racial hatred (Council of Europe, Committee of Ministers, Recommendation on Hate Speech, 1997). In relation to online hate speech, the Council of Europe's Additional Protocol to the Convention on Cybercrime concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems, goes further than the ECHR which 'permits' hate speech laws, by imposing an obligation on signatories to create laws specifically to combat xenophobia and racism generated through computer systems.

Insofar as international human rights frameworks are concerned, freedom of speech is protected by Article 19 of the Universal Declaration of Human Rights and under Article 19 of the International Covenant on Civil and Political Rights (ICCPR). Freedom of speech is fundamental according to these frameworks, but not absolute, and limitations to freedom of speech are articulated under Article 19(3). In addition to these limitations, Article 20(2) of the ICCPR requires that any 'advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility or violence" must be prohibited by law. Thus freedom of speech is firmly part of international human rights law and set the outer limits of hate speech laws, but the wide acknowledgment of freedom of speech hides huge discrepancies in how free speech is balanced with hate speech (O'Regan, 2018). The USA in particular has entered a reservation with regard to Article 20 (2).

Most European countries have evolved their own set of hate speech laws that are the result of their own political and cultural history. As such, hate speech laws will vary from country to country such as in terms of which groups are protected by the laws, or how the 'hate' is manifested (O'Regan, 2018). Nevertheless, there is a common core to these offences in that they attempt in some way to outlaw speech that incites violence, hatred or discrimination against named groups. Such hate speech offences, therefore, put the onus on European states to enforce them. Since it is frequently impossible to locate and prosecute the speaker of the information, this immediately raises the question whether SMCs as gatekeepers should be liable. SMCs in the EU may be liable under relevant criminal laws if they have knowledge of such speech and omit to take any action under principles of accessorial liability (Coe, 2015).

If we take the US position on hate speech as our starting point, this means that there is a justifiable assumption that the law should not compel SMCs to do anything about this material as to do so would impose unjustifiable restrictions on free speech. Thus, if an argument is to be made for regulation to occur, we would need to put forward a very good explanation for why the material is harmful, and why it needs to be criminalised. Whilst some academics have engaged philosophically with this question (for example, Waldron, 2012) and researchers have tried to show the harm caused by online hate (for example Awan and Zempi 2015, 2016), it can be difficult to prove categorically a causal link between online and offline hate crime other than in the most extreme cases such as in terrorism-related situations.

By contrast, if the debate were framed more from a European perspective, the question posed would be fundamentally different. Given that the material concerned is already illegal under national laws, the question then becomes why SMCs should *not* be compelled to remove it.

This has led many to raise the question of whether SMCs should take greater responsibility for content on their platforms. For example, a report by the UK Parliament has called for a special category of responsibility to be created under UK law which would see SMCs defined as something between mere ‘platforms’ and publishers, thus presumably envisaging greater responsibility than mere notice and take-down (House of Commons, ‘Disinformation and Fake News final report’, 2019). This recommendation has now been followed up in the UK Government ‘Online Harms’ White Paper. This has outlined plans to impose a duty of care on SMCs to protect their users from harm (DCMS and Home Department, 2019). The appropriate legal status of SMCs has been explored in detail elsewhere (for example, Bridy 2018, Klonick 2018). However, it is clear that the status quo is being challenged and in ways that do not automatically result in infringements to freedom of expression.

It is important to recognise this difference between the US approach to regulation, and what could be broadly referred to as the European approach. Failing to do so can mean that two important points are lost in the debate. The first is that whether or not a state can legitimately impose legal obligations on SMCs to remove illegal hate material will depend on its approach to free speech as a constitutional right more generally, and hate speech more specifically. Therefore, to oppose regulation purely on the basis of freedom of speech is driven largely by US First Amendment concerns and does not recognise the varying approaches to hate speech across the world.

The second point that is often lost in the debate because of the emphasis on free speech is that there is a crucial difference between regulating legal speech and regulating illegal speech. Any discussion about regulation needs to pay close attention to what is considered illegal hate speech under the law, and cannot be based purely on what might be deemed to be 'unacceptable' content, but which may be entirely legal, and which it would not be legitimate to expect SMCs to remove.

Thus, this section has shown that whilst freedom of expression concerns are legitimate, where hate speech laws already exist, imposing an obligation on SMCs to take more responsibility for content on their site is not controversial as a general principle. However, how this is implemented in practice is crucial. The next section will examine some of the ways currently being used to do this, or where proposals have been put forward to impose greater responsibility on SMCs.

4. Options for Regulation

So far, we have shown that SMCs are in a position to exert some control over the material on their platforms. We have also shown that whilst freedom of expression concerns are legitimate, where hate speech laws already exist, imposing an obligation on SMCs to accept more responsibility for their site is not controversial as a general principle. However, how this is implemented in practice is crucial, and overly broad provisions, or ones that do not sufficiently oversee the moderation process, could lead to too much legal material being removed.

In this next section, we will examine two ways in which regulation of hate speech can occur. The first is through self-regulation, and the second is through top-down regulation with an element of co-regulation (Finck, 2018). It will be argued that self-regulation is problematic and not the appropriate way forward. A better approach is through top-down regulation, such as in Germany and the UK, but in order for this to be successful, it has to be done in such a way that there are appropriate protections in place for freedom of expression. We will also analyse the EU approach to regulation which appears to be moving towards a pro-active filtering model which will require SMCs to use automation, at least to an extent, in order to keep their platforms safe. This approach is mirrored, in part, by the UK proposals in this area, and suggests that this is the direction in which regulation is moving. This too will bring challenges from a freedom of speech point of view that will require particular attention to be paid to the balancing of the different interests in this area.

a) Self-Regulation As A Public Relations Exercise And Its Impact On Free Speech- Really A Softer Option?

To begin with, SMCs were reluctant to police the material that appears on their platforms because this interfered with their business model and the concept of net neutrality. However, as it became increasingly clear that their users were concerned by the level of hate that appears on these platforms, SMCs could see that there were business advantages to being seen to take the problem seriously. Even in the US where freedom of expression concerns are paramount from a legal point of view, research by the Pew Research Centre suggests that 80% of respondents are firmly in favour of SMCs taking responsibility for preventing abuse online, whilst more than half of respondents said that it was more important that SMCs created a welcoming environment than for people to have the right to say what they want online (Pew Research Centre, 2017). From the SMCs point of view, there has, therefore, been a very clear business case for creating their own rules in relation to what material appears online (Frosio, 2018). As a result, SMCs, such as Facebook and Twitter, have published on their websites acceptable use policies and guidelines which are, essentially, self-regulatory tools to govern "objectionable content". An additional reason why SMCs have been keen to regulate is because it was seen as a way of avoiding governmental interference with their business structures.

There have also been initiatives both in the US and in Europe to set up voluntary codes of conduct that SMCs sign up to, and which encourage them to remove unlawful material. In the US, the Working Group on Cyberhate was convened by the American Defamation League (ADL) to look into developing the most effective responses to online hate and bigotry (ADL, 2016). They have produced a Best Practices report which tech companies are urged to voluntarily adopt. As we have seen above, the EU has also published its own voluntary code of conduct which it periodically evaluates in order to test the efficacy of self-regulation.

Whilst SMCs are not state entities, and so therefore not subject to First Amendment restrictions in the US, there are concerns that the size and dominance of these websites, as well as the central role they play in forming public opinion, effectively means that they control citizens' access to speech and so if they block material that is not illegal according to the law, they are creating censorship through the back door. This has raised serious concerns in the US, particularly amongst free speech advocates and internet libertarians who have strong beliefs in the importance of a free and neutral internet. They worry that permitting SMCs to block material at will prevents freedom of expression and curbs innovation (see for example discussion in Citron and Richards, 2018). This has led to attempts to impose network neutrality on SMCs through the Open Internet Order 2010 that purported to prohibit SMCs from blocking any material that passes through their website. However, in the landmark case of *Verizon Communications Inc. v. Federal Communications Commission* (2014) the Court of Appeal invalidated certain aspects of the Order and effectively ruled that SMCs could block material on their websites. Whilst *Verizon* does now allow SMCs legally to apply their community standards, a debate continues to rage in the US over whether network neutrality should also apply to them. More recently, the Democrats have introduced the 'Save the Internet Act 2019' in a bid to restore aspects of the Order. At the time of writing, this has successfully been passed by the House of Representatives but awaits its fate in the Senate.

This issue is compounded by two further problematic aspects of these guidelines. Whilst SMCs, such as Facebook, are willing to adapt their moderation process at the regional level in order to include material that happens to be illegal in a particular country (Klonick, 2018), their terms of service which apply to material which is not necessarily illegal, take effect globally. To the extent that the balance is made in the US headquarters of the companies concerned, there is an allegation of US dominance. For example, Facebook seems to be more obsessed with nudity than depictions of extreme violence, a critique which was made in the wake of it taking down an iconic image of a nine-year old girl running away from a napalm attack during the Vietnam War for the reason that it showed "fully nude genitalia". Secondly, while the guidelines have now been published, they contain, by necessity, general principles which are abstract and whose application in a particular case are opaque.

Thus, whilst voluntary codes might be viewed as a cheap and fairly easy solution to the problem of online hate, and one which avoids fully fledged legislation, it does give SMCs a great deal of power over what appears online. Whilst freedom of speech advocates have been very critical of proposals in favour of governmental regulation of SMCs, this seems to miss the point that without governmental oversight, self-regulation by the SMCs themselves runs the risk of over-moderation. This potentially poses a bigger risk to freedom of expression than a well thought-out regulatory framework which would limit content removal to material that is unlawful. Furthermore, self-regulation by itself is also not entirely suitable because platform providers are motivated by their own financial and business interests, and thus self-regulation lacks transparency and legitimacy insofar as the public interest is concerned.

b) Legislative Interventions

As a result of some of the problems associated with self-regulation, some countries have opted for a legislative approach. Germany was the first country to impose fines on SMCs for failing to remove illegal material quickly enough (Frosio, 2018). The French Parliament has followed suit, and at the time of writing, the French lower house of Parliament has voted in favour of introducing similar provisions in France whereby SMCs will be fined for not removing flagged content within 24 hours. Austria has opted not to hold SMCs responsible for the content that appears on their platforms, but instead proposes to impose an obligation on them to verify their users' identity so that they can be traced if they post hate speech anonymously. More recently, the UK has put forward proposals for a systematic approach to the regulation of SMCs that differs from the approach adopted in Germany and France. This section will consider the German and UK approaches in more detail.

In April 2019, the UK Government published a White Paper setting out its intention to introduce a legislative framework for minimising the dissemination of 'online harms' on social media. The White Paper deals with a broad spectrum of 'online harms' including pornography, terrorist content and child sexual exploitation. Hate speech is not included in the list of online harms, although it can be assumed that the paper has included this with 'hate crime' which is within the ambit of the proposals (UK White Paper, 2019). The current UK Government has debated for a while as to how to tackle 'harmful' content on social media sites. The White Paper proposes to require technology firms to sign up to a number of Codes of Practice, which impose obligations on SMCs to police content on their site. It is also proposed that a new statutory duty of care will be imposed on SMCs, and that a new regulator will be created. This regulator will have the power to fine and issue sanctions against senior executives, and the power to disrupt through the obligations imposed on ancillary services such as search engines and payment providers, and to order blocking at internet access level (UK White Paper, 2019). Thus the UK White Paper goes far beyond a notice and take-down obligation for SMCs as hosting providers and will impose a variety of obligations both on SMCs themselves as well as third parties. SMCs themselves will have an obligation to take pro-active measures to police their sites by using automated filtering and content recognition technologies. Both the vagueness of the regulations imposed by the regulator and the breadth of the scope of measures and the fact that these measures will apply not only to illegal content, but also to "unacceptable" content causes great concern about freedom of expression. Although the White Paper does mention safeguards such as transparency, accountability and complaints procedures, these may not be sufficient.

Whilst it is not surprising in the current climate that the UK government is seeking to impose legal obligations on SMCs to ensure that illegal content does not appear on their sites, it is concerning that the White Paper is not precise in its treatment of hate speech. To begin with, it is particularly problematic that 'hate speech' is, we assume, simply subsumed into the category of 'hate crime' without any recognition of the different issues relating to the two in this context. Whilst 'hate crime' can be used as a broad category that can include 'hate speech', it is important to understand that in this context, 'hate speech' is different to other 'hate crimes' in one important respect. Hate speech is characterised by the fact that it makes certain types of *speech* illegal based on its *content* , whereas, generally speaking, other types of hate crime deal with *behaviour* that is already illegal (such as assault or criminal damage), but which is aggravated on the basis that the perpetrator was motivated by or demonstrated hostility towards a protected characteristic. This means that freedom of speech concerns are central to any treatment of 'hate speech' offences, whereas of less concern in relation to other types of hate crime. As such, in order to ensure that our freedom of expression is properly protected, SMCs would need very clear guidance on what material they can remove and what material they should not remove. This issue is compounded by the fact that the statutory duty of care envisaged under the White Paper, would not only apply in respect of content that is illegal under UK laws, but also to "unacceptable content" that is offensive but legal. The use of such vague terminology does little to assuage any fears that SMCs will find it more expedient to over-moderate in order to be sure they satisfy their duty of care, than to under-moderate and risk breaking the law.

While the UK proposals could be termed ambitious and all-encompassing, by contrast, the German Act focuses specifically on improving the speed and efficiency of notice and take-down and is therefore far more limited in scope. The German legislative proposals do not impose any obligations to pro-actively filter content as to do so was seen as contrary to the hosting immunity contained in Article 14 and the prohibition on a general obligation to monitor in Article 15 of the E-commerce Directive 2000/31/EC. Arguably the obligations of SMCs to monitor online content as imposed by the duty of care in the UK White Paper may conflict with Articles 14 and 15 (which may or may not apply to the UK by the time the legislation is enacted).

In Germany, like the UK, politicians have blamed social media providers for contributing to the dissemination of hate speech online and have called on them to "do more" to prevent the spread of hate speech, abuse and extremist content on their platforms. In this vein, the German Minister of Justice, Heiko Maas, published a draft Bill on 17 March 2017 which was passed on 1 September 2017 and came into force on 1 October 2017 (Network Law Enforcement Act 2017). In the speech introducing the Bill to Parliament, the German Minister said: "self-regulation by the relevant companies has had some success, but has been insufficient. New figures show: not enough criminal content is taken down and the processes are too slow. The biggest problem remains that social networks do not take seriously the complaints of their own users. Therefore it is clear to us that we have to increase the pressure on social networks. …." (Maas Speech, 2017).

This Act obliges SMCs with a user base of at least two million users in Germany to take down content infringing a list of certain provisions of the German Criminal Code within 24 hours (for obviously infringing content) or seven days (where infringement is not immediately obvious), and provide an accessible and efficient notice and take-down procedure for German users, failing which companies may be fined up to 50 million euros (Guggenberger, 2017; Frosio, 2018). This proposal was motivated by the perception of an unacceptable avalanche in hate crime, online abuse and fake news not being countered effectively by SMCs. The Act also introduced bi-annual reporting obligations on SMCs to enhance transparency about user complaints and take-downs, and to put in place a complaints procedure where users can complain about content which has not been taken down. In 2018, the independent complaints body received 8617 cases, but found only 3096 justified as content which should be taken down (36%). Only two percent of the cases of illegal content (62) related to racist online hate (Eco Annual Report, 2019). Thus transparency is one of the standards imposed by newer forms of regulation. However the German legislation does not force SMCs to provide granular reports on the type of speech which has been removed which would be required to assess the operation of the Act in practice (O'Regan, 2018).

Whilst so far, SMCs have appeared to be judicious in their application of the law (CEPS Report, 2018), the fact remains that the primary obligation of the SMCs is to remove material rather than to protect freedom of expression. The law itself does not highlight the importance of freedom of expression, and there appears to be no penalty imposed on SMCs if they over-moderate.

The absence of clear protection for freedom of expression, both in the German Law and in the UK White Paper leaves those attempts open to criticism from freedom of speech advocates. It is possible both to impose a legal obligation on SMCs to remove illegal material and to protect freedom of speech. However, neither attempt analysed here has done so with the necessary rigour and force.

c) EU Law Shifting Away from Intermediary Immunity by Imposing Technological Monitoring Obligations

As has already been observed, until about 2016 the main approach for dealing with illegal content on SMCs' sites was reliance on self-regulatory Codes of Conduct. At the EU level, this manifested itself in the EU Code of Conduct on countering illegal hate speech online which was initiated by the Commission and which was initially joined by Facebook, Microsoft, Youtube and Twitter, with more SMCs joining in 2018. The EU Commission claims in its 4th Monitoring Round of the operation of this self-regulatory Code of Conduct that 89% of content flagged/reported was reviewed within 24 hours and that 72% of content alleged by users and relevant organisations to be illegal hate speech was actually removed (EU Commissioner for Justice, Consumers and Gender Equality, 2019). More specifically, Youtube removed 85% of such content, Facebook 82%, but Twitter only 44%. As to feedback to users and transparency, on average 65% of user notification received feedback from the relevant SMC: Facebook 93%, Twitter 60% and Youtube only in 25% of notifications. The reason for this may be that Youtube is placing reliance on its trusted flaggers programme to which it provides feedback, but not to normal users. Google+ does not provide any feedback in response to notifications (EU Commissioner for Justice, Consumers and Gender Equality, 2019).

However, the recent spate of terror attacks within the EU has changed the purely self-regulatory, laissez-faire approach, and this change is beginning to be reflected in EU instruments countering illegal content. While EU law prevents Member States from imposing liability for illegal content on SMCs before they have actual or constructive knowledge of illegal content on their sites, authorities or courts can order intermediaries to prevent or terminate an infringement or establish a procedure for removing or disabling access to information according to Article 15 of the E-commerce Directive 2000/31/EC. Thus while the starting point is a general immunity for internet intermediaries, various EU legal instruments have recently qualified this immunity. As a result, while initially EU instruments in this area advocated self-regulation and abstaining from the use of automated detection tools, this approach is now changing with a move towards regulatory measures and the use of (at least partially) automated content moderation.

The Counter-Terrorism Directive (2017) imposes an obligation on EU Member States to ensure the prompt take-down of 'online content constituting a public provocation to commit a terrorist offence' (Article 21(1)), and where this is not possible, they may provide for internet access blocking of such content (Article 21(3)) subject to transparent procedures and adequate safeguards (Article 21(3)). The Directive explicitly does not impose an obligation to seek out prohibited content, for example through automated means using artificial intelligence, but leaves the active policing of their platforms to SMCs through self-regulation. It also limits states' legal intervention to ensuring take-down occurs (Recitals 22-

23). This aligns with the EU approach to online media regulation in the latest reiteration of the Audio-Visual Media Services Directive (AVMS) (EU Directive, 2018), which included video-sharing platforms for the first time within the scope of regulation. SMCs are included in the category of video-sharing platforms if the sharing of videos is not merely an ancillary or minor part of the functionality they offer (Recital 5, Art 1 (1) (aa)).

The AVMS Directive envisages and encourages the drawing up of Codes of Conduct by the video-sharing platforms (Art 4a (1) and (2)), but it advocates a co-regulatory approach, beyond the self-regulatory approach. Member States must establish (a) regulator(s) to assess the measures taken by the video-sharing platforms themselves (Art 28b (5)).

Firstly, Article 28b stipulates that EU Member States must take positive measures to ensure protection from three types of content. Secondly, the general public must be protected from user-generated videos and advertising that contains incitement to violence or hatred against a protected group (Art 21 and Art 28(b) of the EU Charter of Fundamental Rights). Thirdly, the general public must additionally be protected from three types of content prohibited in EU criminal law instruments contained in user-generated videos and advertisements: (1) public provocation to commit a terrorist offence (Counter-Terrorism Directive, 2017), (2) child pornography (Directive on Combatting the Sexual Abuse and Sexual Exploitation of Children 2011) and (3) offences related to racism and xenophobia (Council Framework Decision on Racism and Xenophobia 2008). Member States may impose stricter measures, additionally regulating other types of content, thus the AVMS Directive does not fully harmonise the standards in this area.

Thus, the EU Member States, once the implementation deadline for the AVMS Directive has passed on 19th September 2020, have to take *regulatory* measures to curb online hate speech, terrorist content and child sex abuse material on video-sharing services. The

AVMS Directive does not stipulate the precise nature of the measures to be taken by the Member States, but sets out the general principles for taking such measures which are similar to the principles set out in the UK White Paper. First of all, Member States should adopt a risk-based approach, being informed by the nature of the content and its harmfulness, the intended audience to be protected, as well as the interests of the video-sharing platform, the users who have uploaded the content and the public interest. Furthermore the AVMS Directive adopts a practical and proportionate approach which takes into account the size of the video-sharing platform and the nature of its service. Interestingly the AVMS Directive states that the measures should not comprise "ex-ante control measures" or "upload-filtering of content" in breach of the prohibition on the imposition of general monitoring obligations on hosting services (E-commerce Directive, Article 15(1)). In other words automated tools based on artificial intelligence must not be implemented in such a way that they lead to the automated, overbroad filtering of content and general monitoring of all content. This means that such tools must be supplemented by human review and the measures themselves must be specific and targeted, in accordance with Article 14 (3) of the E-commerce Directive which permits specific orders by administrative authorities or courts to terminate or prevent an infringement and which also permits procedures "governing the removal or disabling of access to information". Under Article 28b, the AVMS Directive lists the measures which video-sharing platforms must implement by way of co-regulation, such as prohibiting the three types of content in their terms and conditions, providing for users the opportunity to report and flag such illegal content, providing transparent information as to what the SMC has done with content reported or flagged, providing age-verification mechanisms for content harmful to children, implementing content rating systems. parental control systems for content harmful to children, complaints handling measures and measures to improve digital literacy. Furthermore the AVMS Directive envisages the use of alternative dispute resolution mechanisms. Finally the AVMS Directive envisages further Codes of Conduct in respect of hate speech. While the UK may not be part of the EU in 2020, the UK White Paper strongly aligns with the approach in the AVMS Directive.

A similar co-regulatory approach (Codes of Conduct coupled with an obligation to implement these standards by SMCs) has been adopted in respect of copyright infringement and this also means that SMCs have to use technological solutions to prevent copyright infringement (such as the prevention of re-uploading of proscribed content previously found through Youtube's content id, or the closure of accounts) in the recent revision of the EU Copyright Directive, which has been similarly controversial (Reynolds, 2019). This Copyright Directive also forces SMC to take on more responsibility in respect of law infringements by user-uploaded content.

Finally, the EU has issued several instruments on measures to effectively tackle illegal content online. The EU Commission's Communication (2017) on tackling illegal content online outlines the Commission's thinking in respect of achieving enhanced responsibility of online platforms for illegal content such as incitement to terrorism, xenophobic and racist speech, and child sex abuse materials and responds to the EU Council's political calls for industry to develop "technology and tools to improve the automatic detection and removal of content" (European Commission Communication, 2017). The Communication states that SMCs should take pro-active steps to detect and remove illegal content through automated means, but that this currently requires final vetting through human review (which it calls the "human-in-the-loop" principle). The EU Commission points to the need to ensure notice and stay-down of illegal content, and in particular, the need to prevent re-uploads of the same known content by automatic means.

Moreover, it points to the need for close co-operation between SMCs and law enforcement, but also between law enforcement authorities within the EU to achieve a better co-ordinated response and refers to the EU Internet Referral Unit at Europol as a model of EU co-operation. It points to the greater effectiveness of notice and take-down schemes using trusted flaggers (such as the IRU at Europol) and recommends EU-wide criteria and certification of trusted flagger schemes to prevent abuse of take-down mechanisms and to protect freedom of expression. Furthermore all users should have available convenient and easy-to-use reporting mechanisms. The Communication points to the need to preserve the evidence of criminal activity (and share it with law enforcement). Finally the EU Commission calls for increased transparency about the number and types of notices received, the time it took to respond to the notices, and any actions taken. In addition, the Community Guidelines and procedures for notice and action should be transparent, and the Commission recommends the availability of counter-notices contesting removal of content.

The Communication was followed up with a non-binding EU Commission Recommendation (2018) on measure to effectively tackle illegal content online.

Finally the EU has issued a Regulation for creating a harmonised system of removal orders for online *terrorist* content (EU Regulation on preventing the dissemination of terrorist content online, 2018). The EU Commission Proposal envisages a new removal order for terrorist content (any format, not just videos, but also images and text) on hosting services, including social media. This would apply to all hosting platforms, regardless of their size and introduce co-ordination obligations between the authorities of the Member States and Europol and sets as a standard that terrorist content must be removed by SMCs within one hour. The Proposal also provides that SMCs must use automated detection tools, but envisages safeguards, complaints mechanisms and transparency reporting. In particular Article 9 (2) currently provides that "Safeguards shall consist, in particular, of human oversight and verifications where appropriate and, in any event, where a detailed assessment of the relevant context is required in order to determine whether or not the content is to be considered terrorist content." Finally, it provides for the preservation of content taken down, in order to enable the investigation and prosecution of criminal offences, or if content is found not to be illegal, in order to enable it to be uploaded again.

Thus, it can be seen that both the UK and the EU approach are moving away from notice and take-down, and towards pro-active filtering. This brings with it particular issues in relation to freedom of expression that will need to be at the heart of any such initiatives. Coe points to the social media paradox - the fact that social media open up unprecedented opportunities for the free flow of speech, and thus user empowerment, but that this empowerment is equally dangerous and threatens individual rights and public disorder (Coe, 2015). It is this paradox which calls for the finding of an appropriate balance between the protection of free speech and the prohibition of hate speech. There are huge challenges ahead, and to steer a path that balances the different interests at stake will require compromise and cooperation.

5. Ways of Tackling Hate Speech Other than Hard Law or Self-Regulation

It seems clear, therefore, that whilst platform providers do have at their disposal the technology and money to do something to help combat online hate speech, there are important limitations to the effectiveness of these remedies. Shifting the responsibility to third party intermediaries is a cheap and politically expedient solution, but it is important to recognise that it will not be a panacea.

There are different types and levels of hate speech. The motivation of the maker of the hate speech can range from the unthinking and thoughtless, to the purposeful and intentionally destructive. The impact of the hate speech could be just as serious irrespective of the intention of the offender, however, it may make a difference to how SMCs deal with that behaviour, particularly when dealing with the makers of hate speech who are on the lower end of the spectrum of seriousness (Rowbottom, 2012; Bakalis, 2017).

Researchers have found that people behave differently depending on a variety of factors such as anonymity and incentives for good behaviour (Binns, 2014). For example, there is evidence that those sites which encourage anonymity have far greater incidences of bullying and hate speech (Binns, 2013).

Furthermore one problem, is the "filter bubble", which means that because of the profit-maximising architecture of most social media sites, content is targeted on the basis of profiles of users' interest as this maximises users' engagement with the social media site and therefore advertising revenues (Pariser, 2011) But as a consequence, user groups are segregated into different groups, for example in relation to their political or religious identity. This in turn means that users do not challenge their own views and opinions against those of others which leads to echo chambers and increases the likelihood of users expressing hate. This again is a problem stemming from the architecture of social media sites. The major SMCs, such as Facebook and Twitter, have therefore launched specific counter-speech initiatives to challenge those users who seem to be interested in extremist content.

Other potential ways of discouraging hate speech could be by allowing victims of hate speech to confront the person that has written something about them to explain why what they have said is harmful. Or by encouraging the use of the technology that already exists on SMCs, such as Twitter, which allows users to block material from their view or to silence it without Twitter having to remove the offending material. There are also preventative measures that should be considered, and which could be used to prevent unlawful material appearing in the first place. For instance, codes of conduct can provide clearer guidelines to users about what kind of conduct is considered unacceptable by giving examples of the sorts of speech that can fall foul of the law.

Conclusion

Evidently, SMCs do have the technological know-how to help in the fight against online hate, at least to some extent. However, the rhetoric in relation to regulation of online hate has tended to be dominated by US First Amendment concerns, which do not represent the legal culture in other areas of the world, such as Europe. It is, therefore, legitimate for a state to compel SMCs to remove online hate if to do so aligns with its legal stance on hate speech and with its position on internet regulation more broadly.

However, it is important that any attempt to do so makes clear distinctions between legal material (which should not be removed) and illegal material (which can be removed). Attempts to impose legal responsibilities on SMCs in Germany and the UK, while different in their respective approaches, fall short of this.

Given that the EU position appears to be shifting towards imposing greater responsibility on SMCs, including the potential to require them to act proactively in relation to illegal material, this issue is pressing. As well as some of the conceptual concerns identified in this piece about the difference between legal and illegal hate speech, the issue of proactivity and reactivity and the appropriate legal status of SMCs, we also need to consider whether there are other ways in which SMCs can be forced to act, for example by finding ways to actively discourage hate speech on their platforms through measures with less impact on freedom of expression. Content moderation through technology is also a major concern for free speech as it makes regulation non-transparent, inaccurate and unaccountable.

Bibliography

American Defamation League. (2016). Best Practices for Responding to Cyberhate, https://www.adl.org/best-practices-for-responding-to-cyberhate

American Defamation League. (2019). Online Hate and Harassment: The American Experience. Retrieved from https://www.adl.org/onlineharassment

Ammar, J. (2019). Cyber Gremlin: Social Networking Machine Learning, And The Global War On Al-Qaida and IS-Inspired Terrorism. International Journal of Law and Information Technology 27 (forthcoming).

Amnesty International. (2019). Troll Patrol Findings. Retrieved from https://decoders.amnesty.org/projects/troll-patrol/findings#what_did_we_find_container

Association for Canadian Studies, and Canadian Race Relations Foundation. (2019). Canadians Views on Hate

BBC News (1). (2019). Retrieved from https://www.bbc.co.uk/news/technology-47758455

BBC News (2). (2019). Retrieved from https://www.bbc.co.uk/news/world-asia-48033313

Beliveau, A. (2018). Hate Speech Laws In The United States And The Council Of Europe: The Fine Balance Between Protecting Individual Freedom Of Expression Rights And Preventing The Rise Of Extremism And Radicalization Through Social Media Sites. Suffolk University Law Review, 51 (4), 565-588.

Bakalis, C. (2017). Rethinking cyberhate laws. Information and Communications Technology Law 27: 86.

Binns, A. (2014) Twitter city and Facebook village: Teenage girls' personas and experiences influenced by choice architecture in social networking sites. Journal of Media Practice, 15 (2): 71.

Binns, A. (2013) Facebook's Ugly Sisters: Anonymity and Abuse on Formspring and Ask.fm. Media Education Research Journal, ISSN 2040-4530.

Boyd and Ellison (2007), Social Network Sites: Definition, History and Scholarship,

Journal of Computer-Mediated Communication, 13, 210–30.

Bridy, A. (2018). Remediating Social Media: A Layer-Conscious Approach. Boston University Journal of Science and Technology Law, 24:193.

Citron, D.K. and Richards, N.M. (2018). Four Principles for Digital Expression (You Won't Believe #3!). Washington University Law Review 95:1353.

Citron, D.K. and Wittes B. (2017). The Internet Will Not Break: Denying Bad Samaritans Section 230 Immunity. Fordham Law Review, 86(2), 401.

Coe, P. (2015) The Social Media Paradox 24 (1) Information & Communications Technology Law, 24(1), 16-40.

Cohen Almagor, R. (2015). Confronting the Internet's Dark Side: Moral and Social Responsibility on the free highway. Cambridge University Press.

Commission Staff Working Document. (2018). Impact Assessment, Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online. SWD(2018) 408 final of 12. September 2018, p.14

Commission Staff Working Document. (2019). Impact Assessment, Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online. SWD(2018) 408 final of 12. September 2019, p.15

Council of Europe. (2002). Additional Protocol to the Convention on Cybercrime concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems. Retrieved from https://rm.coe.int/168008160f

Council of Europe, Committee of Ministers. (1997). Recommendation R (97) 20 of the Committee of Ministers to Members States on "Hate Speech"

Council Framework Decision on Racism and Xenophobia. (2008). 2008/913/JHA of 28th November 2008

Digital, Culture, Media and Sport Committee. (2019). Disinformation and Fake News final report. House of Commons, 8th Report of Session 2017-19, 2019

Eco Annual Report. (2019). Retrieved from https://www.eco.de/wp-content/uploads/2019/03/20190310_Jahresbericht_Beschwerdestelle_2018.pdf and https://www.zeit.de/politik/deutschland/2019-03/netzdg-netzwerkdurchsetzungsgesetz-jahresbericht-eco-beschwerdestelle

Echison, W and Knodt, O. (2018). Germany's NetzDG: A key test for combatting online hate. CEPS Research Report, No. 2018/9.

EU Commission's Communication. (2017). COM(2017) 555 final of 28. September 2017. Retrieved from https://www.isdc.ch/media/1579/2-https___eur-lexeuropa.pdf

EU Commissioner for Justice, Consumers and Gender Equality. (2019). Factsheet. Retrieved from https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en

EU Voluntary Code of Conduct. (2016). Retrieved from https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en

Finck, M. (2018). Digital co-regulation: designing a supranational legal framework for the platform economy. European Law Review, 43(1), 47-68.

Frosio, G. (2018). Why keep a dog and bark yourself? From intermediary liability to re-sponsibility. International Journal of Law and Information Technology, 26 (1), 1-33.

Guardian News (1). (2018). Retrieved from https://www.theguardian.com/technology/2018/sep/24/facebook-moderators-mental-trauma-lawsuit

Guardian News (2) (2018). Facebook releases content moderation guidelines – rules long kept secret. Retrieved from https://www.theguardian.com/technology/2018/apr/24/facebook-releases-content-moderation-guidelines-secret-rules

Guardian Newspaper. (2018). Cambridge Analytica files. Retrieved from https://www.theguardian.com/news/series/cambridge-analytica-files

Guggenberger, N. (2017). Das Netzwerkdurchsetzungsgesetz in der Anwendung. Neue Juristische Wochenschrift, 70 (36), 2577-2582.

Imran, A. and Zempi, I. (2016). The Affinity between Online and Offline anti-Muslim Hate Crime: Dynamics and Impacts. Aggression and Violent Behaviour, 27:1-8.

Imran, A. and Zempi, I. (2015). 'I will Blow your face off' - Virtual and Physical World Anti-Muslim Hate Crime. British Journal of Criminology DOI: 10.1093/bjc/azv122

Johnson, D. and Post, D. (1996). Law and Borders: The Rise of Law in Cyberspace. Stanford Law Review, 48(5), 1367.

Klonick, K. (2018). The New Governors: The People, Rules and Processes Governing Online Speech. Harvard Law Review, Vol.131:1598.

Laidlaw, E. (2015) Regulating Speech in Cyberspace. Cambridge University Press

Lessig, L. (1999). Code and Other Laws of Cyberspace. Basic Books.

Maas Speech. (2017). Retrieved from http://www.bmjv.de/SharedDocs/Artikel/DE/2017/03142017_GE_Rechtsdurchsetzung_Soziale_Netzwerke.html

Murray, A. (2007). The Regulation of Cyberspace: Control in the Online Environment. Routledge.

Murray, A. (2016). Information Technology Law. OUP, 3rd edition.

O'Regan, C. (2018) 'Hate Speech Online: an (Intractable) Contemporary Challenge?' Current Legal Problems, 71 (1), 403-429.

Pariser, E. (2011). The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think. Penguin.

Rainie, L and Anderson, J. (2017). The Fate of Online Trust in the Next Decade. Pew Research Center. Retrieved from http://assets.pewresearch.org/wp-content/uploads/sites/14/2017/08/09163223/PI_2017.08.10_onlineTrustNextDecade_FINAL.pdf [https://perma.cc/YK8P-ABYU]

Reidenberg, J. (1996). Governing Networks and Rule-Marking in Cyberspace. Emory Law Journal, 45, 911.

Reynolds, M. (2019). What is Article 13? The EU's divisive new copyright plan explained. Wired. Retrieved from https://www.wired.co.uk/article/what-is-article-13-article-11-european-directive-on-copyright-explained-meme-ban

Rowbottom, J. (2012). To Rant, Vent and Converse: Protecting Low Level Digital Speech. Cambridge Law Journal, 71:355.

Secretary of State for Digital, Culture, Media & Sport and the Secretary of State for the Home Department. (2019). Online Harms White Paper.

Süddeutsche Report. 2016. Retrieved from https://www.sueddeutsche.de/digital/exklusive-sz-magazin-recherche-inside-facebook-1.3297138

The Cleaners Documentary. (2018). Retrieved from https://www.theverge.com/2018/1/21/16916380/sundance-2018-the-cleaners-movie-review-facebook-google-twitter

UN General Assembly. (2018). Report of the UN Special Rapporteur David Kaye on the Promotion and Protection of the Right to Freedom of Opinion and Expression. A/HRC/38/35

US Senate Committee on Commerce, Science and Transportation. (2018). Hearing on Extremist Propaganda and Social Media. Retrieved from https://www.c-span.org/video/?439849-1/facebook-twitter-youtube-officials-testify-combating-extremism

Waldron, J. (2012). The Harm in Hate Speech. Harvard University Press.

Wired. (2018). Retrieved from https://www.wired.co.uk/article/isis-propaganda-home-office-algorithm-asi

Zuckerberg, M. (2019). Retrieved from https://www.independent.co.uk/news/world/americas/mark-zuckerberg-facebook-regulation-internet-government-washington-post-a8847701.html (original Washington post article is behind a pay-wall).

Cases, Statutes and EU Directives/Regulations

Pavel Ivanov v Russia, Application no. 35222/04 (2007)

Brandenburg v Ohio, 395 U.S. 444 (1969). In the case of *R.A.V. v. City of St Paul*, 505 U.S. 377 (1992),

ACLU v. Reno (1997) - 521 U.S. 844 (1997).

Packingham v North Carolina 137 S. Ct. 1730 (2017).

Verizon Communications Inc. v. Federal Communications Commission 740 F.3d 623 (D.C. Cir. 2014).

Network Law Enforcement Act 2017 http://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/DE/NetzDG.html (in German)

The Counter-Terrorism Directive Directive (EU) 2017/541 on Combating Terrorism of 15th March 2017.

Audio-Visual Media Services Directive (EU) 2018/1808 of 14th November 2018.

EU Charter of Fundamental Rights.

E-commerce Directive 2000/31/EC.

EU Directive Copyright in the Digital Single Market.

EU Regulation on preventing the dissemination of terrorist content online, 2018 Proposal of 19th September 2018, COM/2018/640 final.

Counter-Terrorism Directive (EU) 2017/541.

Combatting the Sexual Abuse and Sexual Exploitation of Children Directive 2011/93/EU

EU Commission (2017) COM(2017) 555 final of 28. September 2017, p. 16.

Commission Recommendation of 1. March 2018, C(2018) 1177 (final).