Article

Ethnicities

Ethnicities
2018, Vol. 18(3) 297–326

DOI: 10.1177/1468796817709846
journals.sagepub.com/home/etn

SAGE

What is so special about online (as compared to offline) hate speech?

Alexander Brown
University of East Anglia, UK

Abstract
There is a growing body of literature on whether or not *online* hate speech, or cyberhate, might be special compared to *offline* hate speech. This article aims to both critique and augment that literature by emphasising a distinctive feature of the Internet and of cyberhate that, unlike other features, such as ease of access, size of audience, and anonymity, is often overlooked: namely, instantaneousness. This article also asks whether there is anything special about online (as compared to offline) hate speech that might warrant governments and intergovernmental organisations contracting out, so to speak, the responsibility for tackling online hate speech to the very Internet companies which provide the websites and services that hate speakers utilise.

Keywords
Cyberhate, hate speech, free speech, Internet companies, regulation

Introduction

There is a growing body of literature which, not merely documents the variety and extent of hate speech on the Internet, but also considers whether or not online hate speech, or cyberhate, might be different—or special if that is the right word—compared to offline hate speech (Citron, 2014; Citron and Norton, 2011; Cohen-Almagor, 2015; Delgado and Stefancic, 2014; Perry and Olsson, 2009; Tsesis, 2001).[1] Part of the impetus behind this literature is also to better understand the distinctive challenges of combating cyberhate. This article aims to both critique and augment that literature by emphasising a distinctive feature of the Internet and of cyberhate that, unlike other features, such as ease of access, size of audience, and anonymity, is often overlooked: namely, instantaneousness. I argue that the instant

Corresponding author:
Alexander Brown, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK.
Email: alexander.c.brown@uea.ac.uk

nature of online communication encourages forms of cyberhate that are more spontaneous and, therefore, unconsidered.

In addition to this, I seek to address a related set of questions that have also received less attention than they deserve. What, if anything, is different about the regulation of online (as compared to offline) hate speech? And, what, if anything, is different about free speech objections to the regulation of online (as compared to offline) hate speech? Focusing on the case of England and Wales, I argue that, when it comes to hate speech that does *not* fall under the scope of existing laws (offences) restricting the use of hate speech (i.e. the stirring up hatred offences[2] and certain aggravated public order and harassment offences[3])—such as hate speech that takes the form of negative stereotyping, group defamation (*senso stricto*), or even Holocaust denial—there are similarities between the ways governments have sought to outsource the regulation of online hate speech to Internet companies and the way governments have previously outsourced the regulation of offline hate speech to traditional media companies, including TV and radio broadcasters as well as newspapers and magazines. In both instances this reflects concerns over free speech that point toward governments not being directly involved in the regulation of hate speech (provided that the speech does *not* fall under the scope of existing hate speech laws). Nevertheless, one point of differentiation between the extra-governmental regulation of online and offline hate speech is the extent to which social networking websites like Facebook, Internet messaging services like Twitter, and video-sharing platforms like YouTube, for example, do not typically adopt the prior self-restraint employed by traditional media companies whose application of editorial guidelines and codes of conduct can be used to prevent the broadcast or publication of hate speech before it is ever broadcast or printed.

What's special about online hate speech?

I begin by running through some familiar features of the Internet that could potentially mark out online hate speech as *different* from offline hate speech.

Anonymity

One of the supposed advantages of the Internet as a medium for communication is that people are not compelled to reveal aspects of their offline identity unless they wish to do so. It has been suggested that the anonymity of the Internet can provide opportunities for freer speech because people can say what they think without fear that other people will react or respond unfavourably simply because of the colour of their skin, their sexual orientation, or even their gender identity, for instance (Graham, 1999: 143). This cuts both ways, however. For, there is also evidence to suggest that the Internet disinhibits speakers to say things they would not otherwise say, face-to-face (Suler, 2004). There are different strands to this cyber-psychological phenomenon, but one is that anonymity—even perceived anonymity—can embolden people to be more outrageous, obnoxious, or hateful in what they say

than would be the case in real life (Branscomb, 1995: 1642–1643; Citron, 2014: 57, 59–60; Coffey and Woolworth, 2004: 1–14; Cohen-Almagor, 2015: 86–87, 114, 146; Poland, 2016: 22–24). For instance, the perceived anonymity of the Internet may remove fear of being held accountable for cyberhate and may also evince a sense that the normal rules of conduct do not apply; the associated feeling of liberation may drive people to give in to their worst tendencies (Citron, 2014: 58; Delgado and Stefancic, 2014: 322; Kang, 2000: 1135n.16). On the other hand, the perceived anonymity of the Internet may also liberate victims of online hate speech, as well as their supporters or defenders, to engage in counter-speech. That the hate speakers against whom counter-speakers are speaking back do not know who the counter-speakers are could reduce the fear that counter-speakers might otherwise have had about being identified and targeted in their homes or workplaces. But, then, of course, the anonymity might also encourage counter-speakers to engage in their own hate speech attacking the original hate speakers (Coffey and Woolworth, 2004).

However, the anonymity of the Internet is more complex than it may first appear.[4] For example, users of email accounts, Internet search engines, and media streaming applications now find that their different online identities are merged and cross-checked by their computer's operating system, and some parts of this interconnected web of online identities may touch the offline world, such as when people are users of paid Internet applications and services that require them to give details of home addresses, phone numbers and credit card payment details which are verified. And so users cannot sensibly assume that they will not be tracked down by the police if they perform an illegal act of hate speech on a digital forum, social networking website or Internet messaging service, for example. The wider point is that today the police have powers to track down and seize digital evidence, along with other forms of evidence.

Of course, not all hate speech is illegal. Far from it. So why would anonymity matter for speakers who *only* wish to engage in legally permissible acts of hate speech? Why would they choose to perform such acts online rather than offline? One possibility is that face-to-face hate speakers run the risk of being assaulted by the individuals they are verbally abusing or by other people at the scene. Because the online hate speaker is not *physically present*, he or she does not have to worry about an immediate physical backlash. Yet there remains the risk that a hate speaker can be found after the event, following a bit of detective work by the victims of their speech. Perhaps this is where anonymity comes in. Maybe people are more inclined to engage in even legally permissible acts of hate speech, or more inclined to do so on a regular basis, if they can use anonymous online identities which give no clue as to their true identities. Then again, the technical fixes that people might use to engage in anonymous cyberhate are not bulletproof; so there is always the risk of being discovered and then digitally outed. It is possible in the future that potential online hate speakers might become inhibited by stories of digital vigilantes; people who trap hate speakers into revealing who they are or simply hack into the accounts of online hate speakers, and then post their real identities online for all to see.

Notwithstanding these observations, what about online hate speech that does remain anonymous? Even here we must ask: Is anonymous online hate speech actually so very different to anonymous offline hate speech? If person A walks up to person B on the street and calls them a 'Fucking X', where X is a hate slur, and if A and B are strangers, and if A carries on walking, then A is probably anonymous to B in the sense that B does not know who A is and may have little means of finding out who A is. B may have difficulty even remembering what A looked like if the moment is particularly traumatic. So the chances of being able to provide an accurate description of A to the police as a means to identifying A may be limited, as are the chances of B bumping into A again at some time in the future. In any event, committed hate speakers always have the option of wearing masks, balaclavas or scarves to cover their faces, as many members of hate groups have done, both historically and up to the present day, when participating in rallies or demonstrations in public places. So perhaps what really makes online hate speech different is something else.

Invisibility

A second potentially distinctive feature of online hate speech is that there can be a physical distance between speaker and audience, meaning that the speaker can be non-visible or in some sense invisible to the audience and *vice versa*. Putting visual online communication (online videotelephony) aside for a moment, because non-visual online communication lacks the *face-to-face* dimension of some forms of offline hate speech—recall the street example described above—online hate speakers operate without the normal social-psychological cues of empathy and censure that tend to keep harmful or antisocial behaviour in check. For one thing, online communication often means that the immediate impacts of speech acts are unseen by the perpetrator. If one cannot see the emotional hurt wrought by one's online hate speech, one may be more likely to downplay its significance. 'It's only harmless flaming; people shouldn't take it so seriously.' Moreover, when online, hate speakers cannot see the faces of other people who might disapprove of what they are saying. And, according to Danielle Keats Citron, '[p]eople are quicker to resort to invective when there are no social cues, such as facial expressions, to remind them to keep their behaviour in check' (Citron, 2014: 59).

However, are these aspects salient *by themselves* to understanding what is special about online hate speech? I do not believe so. First, it should not be forgotten that traditional methods of spreading hate speech, such as through newspaper articles, printed leaflets, automated voice messages, letters sent through the mail, graffiti, and so on, also create physical distance between speaker and audience. So non-visual online communication does not have a monopoly on invisibility.

Second, online communication is increasingly combining text, audio and visual data streams (online videotelephony). This means that visual interaction is no longer the preserve of offline communication. Thus, someone speaking to a group of people via a video communications link over the Internet using Webcams or Smartphones can see the faces of the audience.

Third, even without seeing people's faces, there can be explicit verbal social cues that remind online hate speakers (as well as offline hate speakers) to keep their behaviour in check. People can be criticised, condemned, or publicly shamed online for engaging in cyberhate—and potentially this negative feedback can occur almost as instantaneously as in face-to-face communication.

At this stage it might be countered that a person who is committed to engaging in offline, face-to-face hate speech is unlikely to be deterred by seeing the anguish on the face of the victim, and may even take pleasure in it. It could be that hate speakers' own consumption of hate speech—such as reading articles telling them that certain others 'aren't really like us' or 'are animals'—has already undermined or created deficits in natural empathy and sympathy for the suffering of members of the out-groups in question so that, even if they see first-hand the anguish etched across the face of someone, it does not register with them in the normal way (Brown, 2015: 132–137; Taylor, 2006). A hardcore hate speaker might also be indifferent to the risk of physical confrontation, and may even get off on the excitement of that risk; and may delight in disapproving glances by onlookers. If so, online hate speech could potentially lack something that would actually motivate the hard-core hate speaker. Then again, keeping in mind some of the points made above, today the hard core hate speaker can also harness digital technologies to perform visual forms of online hate speech that mimic or replicate many of the qualities and dangers of offline hate speech; producing a similar buzz.

Community

Perhaps, then, the real difference has to do with what the Internet means for situational online hate speakers—people who might or might not otherwise engage in offline hate speech but who are given pathways into cyberhate by the Internet. One pathway already discussed is anonymity. Another is people's innate desire (including people with non-mainstream attitudes) to engage with like-minded others allied to the power of the Internet to put people in touch with each other—people who otherwise might be unable to connect due to geography or people who might be simply ignorant of each other's existence (Posner, 2002: 149–151). Take members of diasporas, for instance, for whom the Internet can be a useful means of communication, a way of keeping in touch with other people from their homeland, but also publicly affirming their ethnic and cultural identities. This, of course, may involve using the Internet to communicate with people still living in their homeland but also with other members of the diaspora in the countries they now reside in. Perhaps unsurprisingly, some hate speakers perform their hate speech as part of commenting on events in their homeland and as part of the wider practice of publicly affirming their identity, not least to other members of their diaspora. In *R. v. Ahmad*,[5] for example, a dual nationality British-Pakistani business information technology graduate using the pseudonym Abu Jahiman engaged in an on-line discussion via the website *IslamicAwakenings.com* concerning a college in India that had allegedly banned its students from wearing

the Burka. Mr. Ahmad wrote, 'Where are the Muslims? They should storm these filthy rabid sub-monkeys and stomp on their jaws until they hear the sweet crack sound and then some.' For this he was found guilty on one count of publishing written material with intent to stir up religious hatred and given a one-year sentence.

To give another example, people who lead hate groups may have a desire to grow their memberships both domestically and internationally, and the Internet may be invaluable for this purpose. On the other hand, new or prospective members of these groups may engage in online hate speech, using the websites set up by these groups, if they feel isolated or if they crave the good opinion of other people who are members of hate groups, as part of a more general human drive for acceptance by others. Thus, hate groups have often tended to use the Internet to attract new members and ensure that their existing members feel connected (Citron, 2014: 61–62; Perry and Olsson, 2009). Here hate speech is addressed, not to victims, but to like-minded people. This is noticeable in the widespread use of negative stereotyping, vilification, group defamation, and Holocaust denial among online hate groups (Cohen-Almagor, 2015; Tsesis, 2001). In the case of far right movements in Europe and the US, for example, researchers have pointed to the fact that groups have been assisted in building large networks based on shared racist ideologies by using the Internet and related forms of electronic or digital communication (Solomos and Schuster, 2002: 45–46). Similarly, Cass Sunstein maintains that what is particularly striking about the activities of online hate groups is that they 'provide links to one another, and expressly attempt to encourage both recruitment and discussion among like-minded people' (Sunstein 2007: 57–58). 'It is,' as Sunstein puts it, 'clear that the internet is playing a crucial role in permitting people who would otherwise feel isolated, or move on to something else, to band together and spread rumors, many of them paranoid and hateful' (58).

In that sense online hate speech is different in one sense simply because it has become the method of choice among hate groups for cementing in-group statuses and fermenting a sense of intra-group community. Of course, this fact itself also relies on some other distinctive features of the Internet. One feature is that the Internet is relatively cheap and easy to use compared to other comparable means of communication (Delgado and Stefancic, 2014: 323)—that is, other means of achieving complex group communication involving targeted delivery or else reaching a mass audience, in short periods of time, and spanning huge distances. The ability of groups to render lack of physical proximity almost irrelevant in the maintenance of intra-group relations cannot be underestimated. Indeed, this virtue of the Internet is precisely what attracted the creators of social networking websites like Facebook and Internet messaging services like Twitter. The Internet helps bring people together, *every* kind of person.

Then there is hate speech that is intended not for an audience of like-minded thinkers but people whose opinion one is trying to change or people who may implicitly feel or think as one does but who are not presently disposed to articulating or venting those thoughts and feelings publicly. Some forms of incitement to

hatred, or stirring up hatred, are addressed to such people. But what makes stirring up hatred online different from stirring up hatred offline? The Internet is not the only way to reach a mass audience after all—consider printed leaflets, automated telephone messages, billboard and graffiti, newspaper and magazine articles, speeches at large public meetings or demonstrations. Then again, the Internet has democratised mass communication. Access to traditional printed newsletters or leaflets requires either the social capital one needs to be on friendly terms with printers or else have enough purchasing power to get them to print leaflets that could have the potential to attract unwanted attention. Then one needs access to distribution networks capable of getting the printed material into the hands of a mass, geographically dispersed audience. Few individuals could have done what Henry Ford did in the 1920s when he used his nationwide Ford dealerships to distribute copies of his *The Dearborn*, containing anti-Semitic stories (Woeste, 2012). Contemporary hate speakers, by contrast, do not need their own newspapers or national distribution networks to reach large audiences through social networking websites, online chat rooms and notice boards, comments pages, Internet messaging services, and video-sharing platforms, for instance.

The above points speak to the reasons why people might choose to utilise the Internet in order to transmit their hate speech. But what if the forms of hate speech that speakers transmit via the Internet are more or less indistinguishable from the hate speech they would find ways to transmit had the Internet never been invented and that they have traditionally transmitted through public appearances, printed material, and older forms of telephone and radio communication? What if the content of hate speech across online and offline platforms is similar and mutually reinforcing in situations where both are used equally heavily?[6] If this is the case, then what the above points reveal perhaps is not the sui generis nature of the content of online (as compared to offline) hate speech, but simply something about the propensity or reasons some hate speakers have, perhaps the majority of hate speakers, for preferring the Internet over other modes of communication.

Now it might be objected at this point that, because of its democratic character, the Internet allows anyone (or almost anyone) to engage in hate speech and to attract an audience online, even if they lack rhetorical, linguistic and artistic skills. This, in turn (so the objection runs), means that online hate speech is likely to be different in quality as compared to offline hate speech; it is more likely to be basic, unsophisticated, unskilled, for instance. However, I find this generalisation to be unfounded. Although virtually anyone can gain access to the Internet, it is not the case that virtually anyone can attract a mass audience of followers, likes and clicks. In order for someone to engage in online hate speech in an effective way, that is, in a way that enables them to gain followers and to harm their victims, it seems to me that they are likely to need similarly high levels of rhetorical, linguistic and artistic skills as are required to engage in effective offline hate speech. Moreover, studies of the content of the hate speech used by online hate groups to communicate on shared issues and to attract and retain members shows a great deal of creativity, sophistication and nuance (Douglas et al, 2005; McNamee et al, 2010). Studies

comparing content in the press and on Internet blogs also reveal that bloggers use detailed analogies and metaphors in denigrating protected groups (Musolff, 2015).

In light of all this, I want next to propose a way in which the Internet might itself encourage differences in the nature or content of the hate speech that people may choose to publish or send, along similar lines to the way that anonymity is thought to promote particularly vicious forms of hate speech.

Instantaneousness

As compared to printed leaflets, automated telephone messages, billboards, graffiti, newspaper and magazine articles, and speeches at large public meetings or demonstrations, the Internet provides people with almost instantaneous publishing. On the Internet, the time delay between having a thought or feeling and expressing it to a particular individual who is located a long distance away, or to a group of like-minded people or to a mass audience can be a matter of seconds. By contrast, if, for example, an ordinary member of the public wants to distribute a group libel about Jews to a mass audience using traditional media, it can take a not-inconsiderable amount of time to design and print leaflets and to hand out those leaflets to people on the street or to send them to people through the post. It also takes time to create an automated telephone message, to set up the necessary phone accounts, to obtain a set of telephone numbers and to complete the automated calls. There is much less of a time-lag in the case of online publication. My hypothesis, then, is that, as compared to offline modes of communication, the Internet encourages forms of hate speech that are spontaneous in the sense of being instant responses, gut reactions, unconsidered judgments, off-the-cuff remarks, unfiltered commentary, and first thoughts. Among the common types of online hate speech that may be spontaneous in this sense are uses of abusive or cruel insults or demeaning language or threatening words directed at or against a person or group of persons identified by their race, ethnicity, nationality, religion, sexual orientation, disability, gender identity, or other protected characteristics. The point is that the Internet not merely facilitates, but also encourages, instant responses that are by their nature more spontaneous in the aforementioned sense. I do not mean to suggest that online hate speech never takes the form of careful, well-thought-out, considered, painstaking, and extensively planned statements. Clearly it does. Instead, what I am suggesting is that some parts of the Internet or Internet services encourage gut reactions, unconsidered judgments, off-the-cuff remarks, unfiltered commentary, and first thoughts, because they encourage instant responses.

By way of illustration, on 17 March 2012 Liam Stacey, a third-year undergraduate at Swansea University and user of the Twitter Internet messaging service, had spent the afternoon watching sport and drinking beer. One of the games he watched was an English Premiership football match between Bolton Wanderers and Tottenham Hotspur in which a black player, Fabrice Muamba, collapsed on the pitch with a life-threatening (and ultimately career-ending) heart attack. Not long after the match Stacey posted this message on Twitter, ‘LOL fuck Muamba

he's dead.' Several people responded angrily to the message and exercised their prerogative to engage in counter-speech on Twitter by calling Stacey 'a greasy little welsh sheep shagger' and a 'silly fat wanker'. Stacey responded with a string of further messages including, 'You are a silly cunt your mother's a wog and your dad is a rapist, bonjour you scruff northern cunt.' He was later arrested and pleaded guilty to the offence of racially aggravated harassment, alarm or distress of intent to users of the Twitter Internet messaging service.[7] What would have happened had Mr. Stacey lacked the option of using Twitter or some other Internet-based method of communication? Maybe after watching the Muamba incident he might have planned to wake up early the next day to design and print a racist leaflet explaining why black football players were inferior to white football players. But by the time he woke up the next day it is also possible that some second thoughts would have occurred to him. He might have remembered the fact that growing up he was a fan of several black players. Moreover, because other people would not have reacted to his Tweet, he would not have also instantly reacted by sending them further Tweets containing racist epithets. That still further responses from other Twitter users came hot on the heels of his additional Tweets created a frenzy of Tweeting back and forth. In other words, the speed of messages encouraged yet more, spontaneous messages and, with it, an escalation of the cyberhate (Coffey and Woolworth, 2004). The moral of the story is Tweet in haste, repent at leisure. It could also be the moral of a high proportion of online hate speech.

As I have said, there are also parts of the Internet that facilitate and encourage more considered, non-spontaneous forms of hate speech, manifested in so-called hate sites that are carefully constructed to appeal to designated audiences, based on shared ideologies, senses of grievance, religious beliefs, and so forth. The content is often uploaded slowly and carefully by people who may have given a great deal of thought to what they are saying, based on a pattern of saying the same thing over time. On the other hand, where hate sites include comments pages or links to online chat rooms and notice boards available for comments by ad hoc contributors, this can encourage users to post messages of support and agreement instantaneously, sometimes spontaneously reacting to each other. But, at any rate, I do not mean to suggest that all parts of the Internet *equally* encourage spontaneous hate speech.

No doubt some people will point out at this stage that online hate speech does not have a monopoly on spontaneity. And they would be right, of course. After all, person A can be travelling on a bus, and upon witnessing an event or situation involving person B, spontaneously shout 'Fucking X!', where X is a hate slur, at person B. This might not be premeditated and might also be an instant response, gut reaction, unconsidered judgment, off-the-cuff remark, unfiltered comment, or first thought. Person A would probably be able to do this a few seconds quicker than it would take him to reach into his pocket pull out a smart phone and Tweet the following, 'Am on bus and a fucking X just did Y—typical!' Face-to-face hate speech also has the advantage, from the hate speaker's perspective, of being able to target someone without knowing that person's name or user name. Of course, the hate speaker could get out his smart phone, take a picture of person B and post it on

Twitter with the words, 'Am on bus and this fucking X (see pic) just did Y—typical!' But unless person A knows person B's email address or Twitter or Facebook user names, he cannot directly message person B, whereas he can say something directly to person B in a face-to-face encounter; and he can do so instantaneously.

Then again, even if person A does not know person B's identity, email address or Twitter or Facebook user names, A can nevertheless post online a different sort of hate speech, which is about an entire group of people, 'Have you noticed how Xs always do … on buses—fuckers!' Moreover, one advantage that online publishing has over merely speaking one's mind in public places is that, with a smart phone, the hate speaker can instantly publish his first thoughts to a large audience of people who do not need to be physically present.

However, the main response I would make to the above point is simply that the nature of online hate speech is likely to reflect a combination of qualities that make online communication special, each of which would not, when considered in isolation, make a significant difference. In other words, the world of online communication is special because it combines anonymity, lack of physical presence, being relatively cheap and easy to use, and the capacity for instantaneous publishing—together these qualities may drive the spontaneity of some online hate speech. Consider the bus example once again. Now the hate speaker on the bus is anonymous, in the sense that other passengers may not know his identity. And let us suppose he can leave the bus before the driver demands to know his identity. Then again, he is physically present in the situation, which means he opens himself up to counter-actions on the part of the person he has verbally abused or other people on the bus. The hate speaker could end up being shoved, kicked, slapped, punched, or assaulted in some other way. There is always the risk that an online hate speaker could be identified and tracked down in the offline world by vigilantes, but that risk is low and the hate speaker can take some precautions. Engaging in face-to-face hate speech carries a more immediate threat of physical backlash. It seems to me that the very fact of physical presence and risk of physical assault may cause people to think twice before engaging in face-to-face hate speech. This, in turn, could mean that they are less prone to engage in spontaneous acts of hate speech. Of course, there will always be hate speakers who are reckless, not in control of their emotions, or actively embrace the risks involved in face-to-face hate speech. However, my hypothesis is not that online hate speech is always spontaneous and offline hate speech, including face-to-face hate speech, is never spontaneous. Rather, my hypothesis is that, as compared to offline modes of communication, the Internet encourages forms of hate speech that are spontaneous in virtue of the combination of qualities that online communication possesses.

Harm

It is now time to discuss whether or not there is anything special about the harmful effects of online (as compared to offline) hate speech. I take it that this is a quantitative, as well as a qualitative, question: that it pertains to how much hate speech

occurs on the Internet as compared to offline domains as well as whether or not online hate speech has distinctive harmful effects, such as creating a unique type of climate of hatred or causing a singular sort of psychological distress. An inquiry focusing on possible differences between the harmful effects of online, as compared to offline, hate speech would also have to isolate other relevant variables, including the relative harmfulness of different forms of hate speech, such as slurs, negative stereotypes, group defamation, incitement to hatred, denial of atrocities, and so on (e.g. Brown, 2015: ch. 3) and the relative harmfulness of hate speech targeted at different groups, such as racial, ethnic, national, religious, sexual orientation, gender, disability groups, and so on (e.g. Brown, 2016, 2017).

Nevertheless, let us focus for a moment on my previous contention that the Internet encourages spontaneous forms of hate speech. If I am right, what implications might this have for the further issue of the distinctive harmfulness of online (as compared to offline) hate speech? The short answer is that, at present, it is impossible to say. As far as I am aware, there is no social scientific evidence yet available that would support any firm generalisations about the comparative psychological effects of being subjected to spontaneous hate speech as compared to premeditated hate speech, whether online or offline. In the absence of social scientific evidence, one is left with anecdotal evidence and commonsense reasoning; and, perhaps unsurprisingly, the latter is likely to prove inconclusive. Maybe it could be more distressing for *some* victims, in one sense, to experience unthinking hate abuse, insofar as it also suggests that one's status is so low that people do not bother to give consideration to how one might feel. Then again, maybe it could be more unsettling for *other* victims, in another sense, to experience premeditated hate abuse, since it implies that one is the object of a determined and planned campaign.

Putting the issue of spontaneity to one side, however, it might be thought that being verbally abused face-to-face because of the colour of one's skin, or because of one's gender identity, sexuality, disability, or religion, say, is more psychologically harmful than being verbally abused online, based on the same targeted characteristics. Perhaps because the former is more *personal* in some sense: because one's real or offline identity is on show as opposed to merely one's online identity. Then again, why should we assume that one's offline identity is any more inherently real than one's online identity? For some people there may be a sense in which they are being more real, or more authentic at least, online because they feel freer or less inhibited in revealing their true selves. At any rate, what if the online hate abuse is more frequent and comes from a larger number of people? The effect might be worse because of the volume of abuse facilitated by online communication. Moreover, it might be that online hate speech has especially harmful effects because it is done in front of larger audiences, thus ramping up the public shame element. Plus, if it turns out that being verbally attacked tends to produce more shame or anguish when done in front of one's friends, family, and work colleagues, say, then it is surely relevant that targeted online hate speech can be, and often is, made public to a very select audience on social networking websites. Then again, there are a great many other ways for hate speakers to launch verbal attacks in front of a

very large audience besides via the Internet. A right-wing or anti-Semitic comedian, for instance, could give a live comedy performance with a couple of thousand people in the audience, and the comedian could point to a person in the audience and say, 'Oh look, here we have an X. Are you sure you paid to get in tonight?', where X is a hate insult. The targeted person could also be sitting with a friend or spouse.

In addition, when harmful effects are being calculated, quantity will matter as much as quality. Quantity can be a function of the duration of time in which the hate speech is publicly available. I can illustrate this with a point made by Citron (Citron, 2014: 4)—it is that there is a potential difference between the sort of street graffiti hate speech case discussed by Waldron in *The Harm in Hate Speech* (2012) and hate messages posted online: namely, the latter can last longer. Whereas graffiti tends to fade, be washed away by the rain, painted over, or lost when buildings are demolished, hate messages posted online remain in pristine digital form. That being said, the position of hate websites and other websites containing hate content in Internet search engine results can be subject to change over time. Even so, a more pressing point is quite simply whether or not the Internet has become the preferred method of performing acts of hate speech for most hate speakers, most of the time.

What does seem clear is that there is an urgent need for existing research on the harmful effects of hate speech (Brown, 2015; Gelber and McNamara, 2016; Jay, 2009; Leets, 2002) to be extended or augmented to include research directly comparing the harmful effects of online and offline hate speech.

The regulation of online hate speech

What, if anything, is different about the regulation of online (as compared to offline) hate speech? Early discussions of the regulation of cyberhate often focused on the question of whether or not laws and legal cases that traditionally concerned the constitutional limits of free speech in offline spaces could be applied to speech in cyberspace (Tsesis, 2001). This continues to be an important issue, of course. But I want to focus on something else. What I want to explore is whether or not the constantly evolving nature and variety of cyberhate (as compared to offline hate speech) poses an especially serious practical (as opposed to constitutional) challenge for the regulation of hate speech.

Because the Internet allows cheap access to mass communication and easy transmission of words, images, music and videos, it has a tendency to support and encourage ingenuity, creativity, playfulness, and innovation in such content (Fuchs, 2014). The same applies to hate speech. Online hate speech is heterogeneous and dynamic: it takes many different forms and those forms can shift and expand over relatively short spaces of time (Citron, 2014; Cohen-Almagor, 2015; Delgado and Stefancic, 2014; Tsesis, 2001). Tomorrow's instances of cyberhate may not always be easily predicted on the basis of today's instances of cyberhate. The Internet is home to forms of hate speech that are banned by existing hate speech laws in England and Wales, including the stirring up of hatred toward people based on certain protected characteristics and certain public order and

harassment offences aggravated by hostility toward people based on certain protected characteristics, for example.[8] But the Internet is also home to hate speech that is not directly banned by existing hate speech laws in England and Wales, including forms of negative stereotyping, vilification, group defamation, and Holocaust denial.[9] In both instances, it is important to ask whether cyberhate can mutate at a faster rate than governmental regulations can keep pace with, and what implications this has for who should be responsible for its regulation.

Take stirring up hatred offences and certain aggravated public order and harassment offences in England and Wales. In some instances governments have spent years considering whether or not to introduce new legislation, months passing draft bills back and forth between the House of Commons and the House of Lords. Following on from this, legal professionals, including police, prosecutors, and judges have grappled with the wording of the legislation applying it to a relatively small number of cases which have slowly emerged over time. Meanwhile the nature of the hate speech being legislated against, including online hate speech, has developed well beyond anything the legislators might have anticipated. I have in mind, not merely the range of groups who are the subject of online hate speech that is covered by existing hate speech bans in England and Wales, but also the specific bits of language that are used to perform acts of cyberhate. In addition to this, there is the problem that criminal prohibitions necessarily involve the punishment of offenders after the fact. If people commit these offences online, the messages they have sent (or the content they have posted) have already been viewed and the damage done before the case ever reaches a court of law.

Nevertheless, it is not clear that this rapidity of change and the challenge of combating online hate speech by means of legislation and criminal prosecutions is significantly different for online as compared to offline hate speech. Hate speakers who prefer to do their hate speaking face-to-face can also exhibit ingenuity, creativity, playfulness, and innovation in content, and this too can pose a problem for legislators and legal professionals. Think of the hate speaker who prefers to perform his hate speech to large audiences in person—where his charisma can shine—but who also knows full well that in order to be convicted of stirring up religious hatred offences in England and Wales, say, public prosecutors must prove both *intent* to stir up hatred and the use of *threatening* words or behaviour.[10] Such a hate speaker has reason to be ingenious in how he or she goes about performing acts of hate speech in order to stay one step ahead of the authorities, whether he or she engages in online or offline hate speech. Much the same problem afflicts the use of campus speech codes to tackle the perennial problem of hate speech on university campuses in the US. Codes and guidance brochures will be amended and adapted by university authorities and discipline tribunals at much slower rates than hate speech itself mutates (Alexander, 1996; Brison, 1998; Craddock, 1995; Delgado, 1991; Downs, 1993, 2005; Lawrence, 1990; Shiell, 2009; Smolla, 1990; Strossen, 1990; Tsesis, 2017).

I also believe that there are certain similarities in the moral responsibilities of Internet companies and broadcast and print media companies when it comes to

certain online and offline hate speech. If Internet companies allow online hate speech to occur by providing social networking websites, Internet messaging services, email accounts, online chat rooms and message boards, comments pages, video-sharing platforms, and so on, then arguably they share some moral responsibility for regulating it (Cohen-Almagor, 2015). The general principle is that, if one invents or is responsible for maintaining technology which facilitates harmful and sometimes illegal actions, especially if one profits from this technology, then it is right that one should be responsible for introducing measures designed to prevent or limit these actions, even if one never intended to facilitate them in the first place. Of course, the nature and extent of the moral responsibility may depend on, *inter alia*, the presence of other agents who may be responsible for creating a climate of hatred in which people think it acceptable to engage in online hate speech, the extent to which hate speech would exist in some form even without the technology, and whether the technology in fact encourages, as well as facilitates, hate speech. But arguably, very similar arguments about moral responsibilities apply to broadcast and print media companies.

Notwithstanding all this, it might be countered that there is another relevant difference between the regulation of online and offline hate speech in terms of the special capacity of companies to regulate their own industry. Partly as a result of the technical challenges in tackling online hate speech perhaps there has been a tendency to assume that the regulation of online hate speech is most effectively done by the very Internet companies whose websites and platforms are being used to transmit the hate speech in the first place. The basic idea is that Internet companies are much better placed than governments to develop and implement terms of usage, community standards, codes of conduct, and so forth, for their users. They are also better placed to train specialist teams of employees to quickly moderate or provide swift adjudications on content that may have been in contravention of the relevant standards, meaning that content could be removed or access to content blocked within a very short space of time after being uploaded or sent. Because these moderators or in-house regulators are likely to be themselves knowledgeable about, and heavy users of, the websites and platforms in question they may be in a good position to keep pace with innovations in online hate speech. If the relevant standards or codes are defined in sufficiently generalised language, the moderators can use their discretion in applying them to specific bits of content. Because the Internet companies are acting as self-regulators they do not have to pause to seek guidance or permission from legislators and legal professionals about the content in question or their adjudications of the relevant standards or codes. They can shift their interpretations of the wording—and even the wording itself—almost instantaneously. And because they are not required to publicise or justify their adjudications to those users whose content is removed or to other stakeholders including Internet rights organisations, they are not hamstrung by precedent. If they feel they need to remove a certain form of hate speech that has previously not been removed or *vice versa*, then they can. They do not need to find reasons for doing so that would satisfy a court of law. All of this builds in a

certain degree of flexibility and swiftness into the tools used to combat online hate speech, the quality of which is partly its being heterogeneous and dynamic.

It should perhaps come as no surprise (it might be further argued) that Internet companies have *for some time* been setting out, and to a greater or lesser extent enforcing, their own rules concerning online hate speech. For example, YouTube's 'community guidelines' states this about 'Hateful content': 'we don't support content that promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics.'[11] And its 'Policy Centre' specifies the following policy on 'Hate Speech': 'We encourage free speech and try to defend your right to express unpopular points of view, but we don't permit hate speech.'[12] Facebook's 'community standards' state 'Facebook removes hate speech'.[13] Likewise, Twitter's 'Ad Policy' states 'Twitter prohibits the promotion of hate content'.[14] And Microsoft's 'Code of Conduct' includes the following code: 'Don't engage in activity that is harmful to you, the Services or others (e.g. transmitting viruses, stalking, communicating hate speech or advocating violence against others).'[15]

Nor should we be surprised to see transnational political organisations joining forces with Internet companies to combat online hate speech, not instead of joining forces with governments but in appreciation of the reality that Internet companies might be not merely more inclined to work with them but also in a better position to get fast results. On 31 May 2016, for example, the European Commission and various major Internet companies announced a new 'Code of Conduct On Countering Illegal Hate Speech Online'.[16] Facebook, Microsoft, Twitter and YouTube have agreed to clarify on their terms of usage that they will prohibit illegal incitement to hatred. Moreover, they have agreed to put in place clear and effective processes to review and remove or disable access to such content within 24 hours.[17] Although the agreement amounted to no more than a public commitment to enforce the new Code and the new Code is not legally binding on these Internet companies, the mere fact that they were willing to make this public commitment demonstrates three salient facts. First, it suggests that Internet companies really do have extraordinary technical ability to control content, despite their sometimes claiming otherwise as a means of abdicating their moral responsibility and, more importantly for them, their legal liability. Second, it is evidence of the fact that Internet companies are increasingly accepting that it is an appropriate function of such companies to police, so to speak, online hate speech and that they have a responsibility to the victims of such speech to make the Internet more secure. Third, it highlights the extent to which Internet companies are willing to work with a range of partners, including not only governments but also intergovernmental organisations and non-governmental organisations, in order to ensure that they are in the vanguard, and therefore partly in control, of creating effective tools to regulate online hate speech.[18] The irony here is that Internet companies are answerable for providing hate speakers with ever-more powerful online tools for spreading their messages but are also best placed to police the use of those tools. And so

the Internet, we might say, provides unprecedented opportunities for hate speakers and regulators of hate speech alike.

But does any of this demonstrate that there is a difference in kind between the regulation of online and the regulation of offline hate speech? Arguably not. For it could also be said that broadcast and print media companies are much better placed than governments to rise to the technical challenges of regulating content that is communicated via their radio and television stations, newspapers and magazines. They, too, have knowledgeable staff who can be trained to intervene swiftly to deal with problematic content.

Furthermore, even if both intergovernmental organisations and governments are willing to entertain the notion that it is Internet companies themselves who can and perhaps *should* take the lead in tackling the problem of online hate speech, this does not mean that there is a division of labour in combating online and offline hate speech. The police and the courts still have a role to play in tackling illegal online hate speech. Now it is certainly true that there is no mention of the words 'computer system', 'the Internet', 'website', 'trolling', 'flaming', 'cyberhate' and other terms relating to online hate speech in legislation defining stirring up hatred offences in England and Wales.[19] Similarly, among the public order and harassment offences that can be charged as aggravated offences (aggravated by hostility toward people based on certain protected characteristics), there is no mention of the Internet.[20] Interestingly, the Communications Act 2003 prohibits messages sent 'by means of a public electronic communications network' that are grossly offensive or of an indecent, obscene or menacing character.[21] But when it comes to prosecuting these offences as aggravated offences the matter is dealt with only in Crown Prosecution Service Guidelines and not in the relevant hate crime legislation.[22] Nevertheless, it is worth emphasising that existing hate speech laws *do* apply, as written and as interpreted, equally to online and offline communication, and have already been applied by the police, prosecutors, and the courts to online communication—consider the Stacey case discussed above.

There are also similarities between the moral responsibilities of governments for regulating online and offline hate speech. Suppose one believes that governments cannot, and should not, abrogate entirely their role in ensuring that citizens' rights are respected, even when it comes to the controversial and technically challenging arena of regulating print and broadcast media. If media companies are responsible for facilitating acts of hate speech in virtue of providing access to printed and broadcast communications, then governments are also responsible for the environment in which media companies operate. So, on this view, governments should at least work to ensure that there is some form of appropriate and effective regulatory framework for dealing with hate speech in the media, even when it comes to hate speech that might not fall under the scope of existing hate speech laws in England and Wales. If one holds this view about print and broadcast media regulation, then surely the same logic would apply to the regulation of hate speech on the Internet. There are, of course, different models for this responsibility. At one extreme is for governments to fully devolve regulatory responsibilities to individual

Internet companies, and not seek to provide any legislative or regulatory framework, but instead liaise with the companies and provide guidance and suggestions. Or it could devolve power to independent Internet regulatory bodies that operate within a broadly defined legal framework, and compel Internet companies to comply with its standards. Or it could devolve power to an industry-based self-regulating body which writes and enforces its own standards. But the bottom line is that these questions about the moral responsibility of governments for Internet regulation seem quite similar to perennial questions about the moral responsibility of governments for the regulation of traditional media.

At this point it might be objected, however, that if, as I have suggested, the Internet has become the medium of choice for hate speakers, and that online hate speech may be more dangerous than its offline equivalent—because of the quantity of harm caused by online hate speech if not also the quality of harm caused by online hate speech—then, arguably, there should be a greater role for the state in governing Internet regulation, for example, by imposing a statutory regulatory framework on Internet companies and by specifying strict standards to be employed by the regulator. However, the one does not follow from the other. When governments choose not to impose a statutory regulatory framework on printed media, they do so not because they downgrade the potential for harm but because they place an emphasis on freedom of the media. Perhaps the same goes for the Internet. If governments impose a statutory regulatory framework on Internet companies and specify strict standards to be employed by the regulator this also poses a significant threat to freedom of the Internet, that is, a threat to the free exchange of ideas online. For example, based on values of democracy and legitimacy, the Internet might be viewed as an especially important site of political speech and public discourse more generally. So a higher risk of harmful speech does not necessarily imply that a higher level of government intervention is appropriate, all things considered. I shall return to discuss free speech objections to the regulation of hate speech in the next section.

However, there is, I think, at least one important difference between the form of self-regulation operated by traditional media companies and the moderation or content regulation imposed by Internet companies. Internet companies involved in providing social networking websites and Internet messages services are, like government organisations and agencies, in a position of reacting to hate speech messages and content once it has been sent through the Internet or posted on the Web. Now the speed of that reaction may be much faster for Internet companies than for government. As mentioned, some Internet companies have committed to removing hate speech content or access to hate speech content within 24 hours. But it remains a form of ex-post self-regulation. By contrast, in the case of traditional media companies the decisions they make about how to apply editorial guidelines occur and take effect prior to the relevant content being broadcast or published. This means that content which is deemed to fall foul of the editorial standards never sees the light of day. It is a form of prior self-restraint (Jacobson and Schlink, 2012). So, for example, the BBC operates under its own ‘Editorial Guidelines’ and these include the following guideline on ‘Portrayal’.

> We aim to reflect fully and fairly all of the United Kingdom's people and cultures in our services. Content may reflect the prejudice and disadvantage which exist in societies worldwide but we should not perpetuate it. In some instances, references to disability, age, sexual orientation, faith, race, etc. may be relevant to portrayal. However, we should avoid careless or offensive stereotypical assumptions and people should only be described in such terms when editorially justified.[23]

When the editors responsible for particular television or radio programmes apply this guideline to proposed content, or when they refer planned content to the managing output controller for the relevant group within the BBC (e.g. News) or to the Chief Adviser on Editorial Policy, the decisions they reach can result in changes to content prior to broadcast.

Of course, broadcast and print media companies also operate under additional regulatory frameworks that pertain to content that has already been broadcast or published. In the case of the BBC, for example, once content has been broadcast people can make a complaint to the Editorial Complaints Unit, including in instances where they perceive that a hate speech has been broadcast in contravention of the standard on portrayal. If the decision does not go in their favour they can then appeal to the Editorial Standards Committee (ESC) which is operated under the auspices of the BBC Trust.[24] In the case of newspapers in the UK, people can potentially bring a complaint against a newspaper for printing hate speech by referring to clause 12 of the Independent Press Standards Organisation (IPSO) Code of Practice for Editors. The first part of clause 12 states: 'The press must avoid prejudicial or pejorative reference to an individual's, race, colour, religion, sex, gender identity, sexual orientation or to any physical or mental illness or disability.'[25] This Code of Practice has its weaknesses, but the key point I want to stress here is that prior self-restraint might be less commonly employed by Internet companies than by traditional media companies.

Free speech objections to the regulation of online hate speech

What, if anything, is different about free speech objections to the regulation of online (as compared to offline) hate speech? There is potentially a great deal for defenders of free speech to be concerned about when it comes to the regulation of hate speech, whether online or offline—concern, for instance, that hate speech laws can be subject to abuse of power, are ineffective, and tend to chill speech in ways that are harmful to autonomy on the part of speakers and damaging to democracy and legitimacy on the part of society as a whole (e.g. Baker, 2009, 2012; Dworkin, 2012; Hare, 2012; Heinze, 2016; Post, 2012; Strossen, 1990, 2012; Weinstein, 2009, 2017). There are things that advocates of (some) hate speech laws can, and do, say in response to these standard objections, of course (e.g. Brison 1998; Brown, 2008, 2015, 2017d; Delgado and Stefancic, 1996, 2009; Gelber and McNamara, 2014). Nevertheless, the point I wish to make here is

that many specific concerns that people may have about the regulation of online hate speech by Internet companies in particular are not *sui generis* but are instead instances of more generic concerns.

Let me try to justify this with three examples. One concern typically raised about hate speech laws enacted by governments is that such laws tend to be vague and because of this chill various forms of valuable speech, including not least political speech or public discourse more broadly construed. Put simply, if a law is expressed in a way that is too unclear for a person of average intelligence to reasonably forecast whether or not his or her speech falls under it, then to avoid the risk of adverse legal consequences he or she may refrain from saying anything remotely controversial, critical, or provocative (Baker, 2009: 157; Strossen, 1990: 521; Weinstein, 2009: 51; Weinstein, 2017). But exactly the same concern might also be applied to community standards adopted by Internet companies because of the potential vagueness in the terminology used to specify what is not permitted, including terms such as 'the promotion of hatred'. If people think that their messages and content could be removed and, more importantly, that their user accounts could be closed by Internet companies, they might think twice about sending or uploading certain content. The thought of being denied access to Facebook and Twitter, say, may be unconscionable for many people. In other words, it is not clear that the relevant free speech objection to the community standards or codes of conduct used by Internet companies is of a different order to the objection to criminal laws banning hate speech including offline hate speech. Then again, perhaps the suggestion is that when national governments introduce laws banning incitement to hatred, say, the legislation is subject to scrutiny by democratic chambers and civil servants and the result is that the language of the legislation can become more precise, to the point that it has less potential to chill speech. However, there is nothing to stop Internet companies from going through a similar process of perfecting the definitions of hate speech used in their community standards or codes of conduct, so as to make them more precise, perhaps in response to feedback from users seeking clarification on what is or is not permitted, and in response to suggestions from their own in-house moderators as they work with the standards over time. Indeed, just as the UK parliament added freedom of expression clauses into the new offences of stirring up hatred on grounds of religion and sexual orientation, so Internet companies have inserted freedom of expression clauses into their community standards and codes of conduct. All of these clauses clarify the limits of the respective hate speech regulations in terms of what would, or would not, count as permissible speech.

Second, government legislation on hate speech is often criticised on the grounds that either it is ineffective in reducing the extent of hate speech (Baker, 2009, 2012; Hare, 2012; Heinze, 2016; Strossen, 1990, 2012) or it is effective and because of this it forces hate speakers to become more secretive or to go underground, which in turn means that society loses an opportunity to monitor hate speakers (Gellman, 1991; Malik, 2005; Smith, 1995; Strossen, 2012) and hate speakers lose an opportunity to blow off steam verbally as a sort of pressure valve before they act out their

beliefs, attitudes and feelings through physical violence (Emerson, 1963; Heins, 1983; Magruder, 1936). I believe that there are some telling replies that can be made to each of the ineffectiveness (Brown, 2015: 239–242; Gelber and McNamara, 2014), pushing underground (Delgado and Yun, 1994a: 1816–1818; Parekh, 2005/2006: 221), and pressure valve (Delgado and Yun, 1994b) objections. Notwithstanding this, the point I wish to make here is that surely the same objections can be levelled against the regulation of online hate speech. Take the ineffectiveness objection. It might be argued that these regulations will not deter users from engaging in hate speech because for every hate message or bit of hate speech content that is removed another, similar hate message or bit of hate speech content will pop up elsewhere on the same Internet messaging service or website or on other parts of the Internet which are controlled by Internet companies who decide not to regulate hate speech. Or take the pushing underground objection. It might be argued that if Internet companies come together and all decide to adopt the same or very similar community standards on cyberhate—perhaps as a result of the work of an international association of Internet companies or an intergovernmental organisation—this might have the effect of forcing hate speakers onto parts of the Internet (including the dark web) which are much less easily regulated by Internet companies because of the use of encryption software. Once again, it might be possible to reply to these objections to online hate speech regulations, but the key point is that these objections are not different in kind as between online and offline hate speech regulations.

Third, it has been argued that the codes of practice operated by traditional media companies which include restrictions on hate speech are less democratic and legitimate than the hate speech laws which are enacted by governments and reviewed by higher courts of law. Consider once again the Code of Conduct On Countering Illegal Hate Speech Online. Here we have an intergovernmental organisation coming together with transnational Internet companies to agree upon a set of codes on cyberhate that restrict the free speech of ordinary people at the national level without each and every national government giving its consent to the agreement. Because the Code is not a piece of European Union law but is instead effectively a form of voluntary self-regulation adopted by Internet companies in consultation with the European Commission, it is not scrutinised nor subject to final approval by the European Parliament and its directly elected Members of the European Parliament (MEPs). This is contrasted with a new piece of hate speech legislation introduced by a national government which would undergo scrutiny and final approval by directly elected representatives. In the words of European Digital Rights (EDRi), an association of civil and human rights organisations from across Europe that lobbies for online rights:

> the 'code of conduct' downgrades the law to a second-class status, behind the 'leading role' of private companies that are being asked to arbitrarily implement their terms of service. This process, established outside an accountable democratic framework, exploits unclear liability rules for companies. It also creates serious risks for freedom

> of expression as legal but controversial content may well be deleted as a result of this voluntary and unaccountable take down mechanism.[26]

In addition to this, the Code does not have to clear the sorts of constitutional barriers which national governments must clear when introducing legislation that limits free speech. There is no supreme court or human rights court to which individuals can appeal on grounds of a violation of their right to freedom of expression by Internet companies. What is more, the day-to-day implementation of the Code including adjudication decisions taken about hate speech content are not subject to any review by higher courts. So users are vulnerable to bad adjudications which they cannot appeal. Allied to this is the fact that, unlike in the case of hate speech laws at the national level which are applied by courts in a public way, the inner workings of Internet moderation operated by companies under the Code remains for all intents and purposes concealed insofar as companies do not make public their adjudications. At least when hate speech laws are passed by democratically elected legislatures, the laws are scrutinised, debated and challenged often over a lengthy period of time (e.g. Bleich, 2011; Brown, 2015: ch. 7, Brown, 2017 c; cf. Heinze, 2016), and, of course, in constitutional democracies whatever legislation is finally settled upon must also demonstrate itself to be compatible with relevant constitutional guarantees of freedom of expression as determined by judges in higher courts of law (domestic), and in some cases by judges in international human rights courts. However, it is hard to see why these sorts of objections relating to democracy and legitimacy are different or special. Arguably the same objections can be put at the door of codes of practice and editorial guidelines employed by traditional media companies in relation to television and radio broadcasts as well as printed newspaper and magazine articles. The codes of practice and editorial guidelines operated by such companies may also cater to a perceived social abhorrence of hate speech but in ways that society is not fully aware of. Traditional media companies may publish their codes of practice but they do not as a matter of course reveal how exactly their editors and content controllers apply or adjudicate these codes of practice in particular cases. Companies do not release this information because they are not required to do so. They can simply say, 'This content was removed for reasons to do with relevant community standards.' And, if push comes to shove, they can stand behind the excuse that their inner operations are commercially sensitive. (These same points may not apply to publicly funded media organisations like the BBC.) According to Arthur Jacobson and Bernhard Schlink, this means that ordinary people are denied the opportunity to know, reflect on and take responsibility for their own abhorrence of hate speech. As such '[w]e lose the power of progressive moral beings' (Jacobson and Schlink, 2012: 231).

Of course, Internet companies can be subject to forms of political online consumerism—meaning, for example, that users can drop Facebook in favour of alternative social networking sites if they do not like the fact that Facebook has adopted a particular community standard relating to hate speech. But without good knowledge of how the community standard is being implemented, users may have

insufficient reasons to switch to a different social networking site especially given the opportunity cost of giving up Facebook. Indeed, even the option of switching to a different company disappears as the number of Internet companies adopting similar community standards on cyberhate increases. So political consumerism may have insufficient power to hold Internet companies to account for adopting the community standards that they do. But once again it is hard to see why these concerns about the accountability of the regulation of online hate speech by Internet companies is any different than concerns about the accountability of the regulation of offline content by traditional media companies including broadcasters and newspapers.

At this stage, some people might try to argue that there is something different about the motives of traditional media companies. In the case of traditional media companies (so the thought goes) restrictions on hate speech are primarily driven by commercial considerations—keeping as many audience members and potential audience members as happy as possible. By contrast, the creators of the large Internet companies are, at heart, technologists, cosmopolitans, and progressives, whose main aim is to connect the world through ever-greater access to forms of online society and to use this online society for good including fighting discrimination and intolerance (it might be argued). However, I think it would be difficult to generalise about differences in the fundamental values of the creators of traditional media companies and Internet media companies. Moreover, it seems naive to ignore the fact that the creators of Internet companies are also entrepreneurs who want to grow small businesses into much large businesses. One also cannot ignore the fact that, once the creators of Internet companies have created large businesses, they often sell them to even larger businesses or float them on stock markets so that they become part-owned by lots of large corporate shareholders. And these large businesses and corporate shareholders may not share the vision of the original creators when it comes to the bottom line. Of course, the relevant community standards and codes of conduct of the Internet companies almost invariably include declarations and clauses that extol the importance of freedom of expression and the need to balance rights. But it is probably the case that these declarations and clauses are also put there for pragmatic purposes: to give the impression (or maintain the conceit?) that these are the sorts of companies that value free speech. Creating this impression might seem especially important for Internet companies given facts about the demographics and personal values of people who spend a lot of time using the Internet for social interaction—the target market of most Internet companies.

Notwithstanding all this, I believe that there is one relevant difference between the content regulation imposed by Internet companies in respect of their community standards and codes of conduct and the self-regulation operated by traditional media companies in respect of their codes of practice or editorial guidelines. As mentioned earlier, traditional media companies are able to censor content before the content is broadcast or published. In that sense codes of practice and editorial guidelines are operated as prior self-restraint, meaning that customers end up never seeing or reading what it is they are assumed to abhor. This is not so with those Internet companies who moderate their users' messages and content *ex post*

(Cohen-Almagor, 2011, 2015). It could be argued that this form of moderation is *less* objectionable on free speech grounds than the prior self-restraint operated by traditional media companies. At least users of Facebook, Twitter and YouTube have a brief window of time in which to see content that is ultimately taken down. This brief window of time enables users to think about and discuss the content and thereafter to reflect on why it was taken down or access removed and whether the decision was the right one. Users who did not get a chance to see the content or did not notice its being removed can be informed of what happened by those who did. Perhaps because of this the argument that, through prior self-restraint, traditional media companies deny their audiences the power of progressive moral beings does not apply with quite the same force to Internet companies who operate *ex post* Internet moderation. In other words, *if* prior self-restraint is the most severe and problematic threat to free speech, and, *if* it is not something standardly operated by social networking websites, Internet messaging services and video-sharing websites, for example, then this is one relevant difference between free speech objections to the regulation of online as compared to offline hate speech. Of course, the same might not be true of Internet Services Providers (ISPs) and web hosting services if they block access to websites that contain hate speech or Internet browsers and Web search engines if they insert default settings for Internet-security controls which block websites based on hate speech content or email packages that insert default settings which place filters on emails containing hate speech in the subject line such as by sending them directly to spam folders.

Of course, arguments against prior self-restraint are not the end of the matter. After all, it might be argued that traditional media companies and Internet companies serve different functions or purposes and that these differences might justify variations in the forms of restraint they each employ. For example, arguably one of the distinctive purposes of newspapers, radio stations and TV channels is to provide editorialised content reflecting particular values and worldviews, and so they will need to censor speech *ex ante* in accordance with their editorial policy. By contrast, Internet companies often maintain that their distinctive purpose is to provide easily accessible platforms for the exchange of content between users and catering to users who adhere to a plurality of values and worldviews. Of course, in reality many Internet companies do engage in user-focused editorialising—consider, for example, the way Facebook employs computer algorithms to control the content that makes it on to users' Newsfeeds based on their previous Facebook habits. But setting aside user-focused editorialising, the aforementioned difference in distinct function or purpose might explain (and to some extent justify) the tighter, stronger form of self-restraint that traditional media companies normally employ. Thus, arguments against prior self-restraint must be viewed in the light of other considerations that might serve to justify it.

In this section, I have looked at objections to online hate speech regulations and found many of these objections to be generic, except in respect of objections against prior self-restraint which, at first glance, seem to be more applicable to traditional media companies than to Internet companies. Perhaps there are other free speech

objections to online hate speech regulations that I have not considered and that are *sui generis*. I do not discount that possibility. But I do not know what they are at this point. Nevertheless, I believe that, in the final analysis, whether the regulation of online hate speech (by governments or by Internet companies) can be justified all things considered will depend on finding a principled compromise between principles that support free speech on the Internet but also principles that support the regulation of cyberhate, including not least principles centred on forms of harm prevention.[27]

Conclusion

The main aims of this article have been to try to explain why one might reasonably think that online (as compared to offline) hate speech has a distinctive quality; why the regulation of online (as compared to offline) hate speech has a distinctive quality; and why free speech objections to online (as compared to offline) hate speech have a distinctive quality. In the first respect I argued that the supposed anonymity of the Internet may not be as distinctive as is first assumed but that the instantaneous nature of communication on some parts of the Internet and the spontaneous hate speech that it encourages might be a better, and often overlooked, reason to mark it out as different. The instantaneous nature of the Internet also partly explains why Internet companies are thought to have a special role in regulating online hate speech. But, by the same token, people who object to the regulation of hate speech can point to equally important deficits in democracy, legitimacy and accountability when Internet companies regulate hate speech as opposed to when government organisations and agencies suppress hate speech. However, despite these differences I also observed that the regulation of online hate speech and free speech objections to the regulation of online hate speech by Internet companies are not *sui generis* if one makes a full comparison with the regulation of offline hate speech and free speech objections to the regulation of offline hate speech by traditional media companies—but with one notable exception, the use of prior self-restraint by traditional media companies and objections to that form of restraint.

A more general implication of the observations I have made so far is that *if* parts of the Internet encourage spontaneous hate speech, and *if* online hate speech already outstrips offline hate speech measured as a proportion of total hate speech and will continue to increase in the future, then we might expect to see *more* not *less* spontaneous hate speech in the future. This conditional implication is closely allied, of course, to the ease of use and relatively low cost of being online. And the fact that people are already likely to be connected to the Internet in the course of their professional, private and civic lives—to work, to consume, to maintain friendships, to find romantic partners, to learn, to play, and to participate in the formation of democratic public opinion—means that they do not have to go out of their way to become spontaneous online hate speakers. Of course, this also means that people are not safe from spontaneous hate speech even in their own

homes. What is more, because the regulation of hate speech is especially challenging it may be that Internet companies should play a bigger part in tackling this problem, day-to-day, as compared to government organisations and agencies, irrespective of, and over and above, any commercial reasons that Internet companies might have for doing so.

All of this may also serve to support the idea that, not just intergovernmental organisations, but also all national governments should be working as closely as possible with Internet companies to help find ways of combating online hate speech, with national governments in a sense contracting out the regulation of the Internet to these companies, at least when it comes to day-to-day adjudications on non-illegal hate speech content. No doubt this form of outsourcing of some of the responsibility for tackling online hate speech faces serious free speech objections: one being that decisions about restricting the basic human right to freedom of expression are far too important to be contracted out to Internet companies who on paper lack the democracy, legitimacy and accountability of government organisations and agencies. But I end with one salutary point. If social networking websites, Internet messaging services, and video-sharing websites, for example, do not, unlike traditional media companies, engage in prior self-restraint, then at least the most severe and problematic form of limitation on the right to freedom of expression has not been contracted out. Filtering and blocking functions performed by ISPs, web hosting services, Internet browsers and Web search engines, and email packages are another matter, of course (Cohen-Almagor, 2011, 2015).

Acknowledgements

The author is very grateful to the journal's three anonymous reviewers for their insights and suggestions.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The term 'hate speech' is an opaque idiom with multiple meanings covering a heterogeneous collection of expressive phenomena (Brown, 2017a, 2017b). I shall not seek here to outline the full variety of phenomena captured by the term 'hate speech'. However, in what follows, where I do want to refer to a particular form of hate speech I shall specify that form. I shall also use the term 'cyberhate' as shorthand for hate speech occurring online as opposed to offline.

2. Parts 3 and 3 A of the Public Order Act 1986 (as amended by the Racial and Religious Hatred Act 2006, the Criminal Justice and Immigration Act 2008, and the Marriage (Same Sex Couples) Act 2013).
3. ss 28–33, 50A of the Crime and Disorder Act 1998 (as amended by the Anti-terrorism, Crime and Security Act 2001); and ss 145 and 146 of the Criminal Justice Act 2003 (as amended by the Legal Aid, Sentencing and Punishment of Offenders Act 2012).
4. For a nuanced discussion of anonymity, the Internet and law, see Barendt (2016: ch. 6).
5. Bristol Crown Court, 29 July 2011.
6. For a discussion of the role of the Internet alongside, and in conjunction with, print and broadcast media in the communication of hate messages in Ethiopia, for example, see Gagliardone et al. (2014: 30–35).
7. Swansea Magistrates Court, 19 March 2012.
8. See notes 2 and 3 above.
9. For a classification of different forms of hate speech that are subject to regulations in different parts of the world, see Brown (2015: ch. 2).
10. See note 2 above.
11. www.youtube.com/yt/policyandsafety/en-GB/communityguidelines.html (accessed 8 April 2017).
12. https://support.google.com/youtube/answer/2801939 (accessed 8 April 2017).
13. www.facebook.com/communitystandards# (accessed 8 April 2017).
14. https://support.twitter.com/articles/20170425# (accessed 8 April 2017).
15. www.microsoft.com/en-gb/servicesagreement (accessed 8 April 2017).
16. http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf (accessed 8 April 2017).
17. The new agreement covers illegal incitement to hatred as defined by the Council of the European Union Framework Decision on Combating Racism and Xenophobia by Means of Criminal Law 2008/913/JHA, of 28 November 2008—that is, 'publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin'.
18. Compare the willingness of the big players in the Internet sector to work with the European Commission on formulating and adopting codes of conduct covering online hate speech with the ambivalent response by nearly half of the member states of the Council of Europe to Treaty 189, the Additional Protocol to the Convention on Cybercrime, Concerning the Criminalization of Acts of a Racist and Xenophobic Nature Committed Through Computer Systems, Strasbourg, 28 January 2003, CETS No. 189. Available at: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168008160f (accessed 8 April 2017). To date it has only been ratified by 24 of 47 member states. www.coe.int/en/web/conventions/full-list/-/conventions/treaty/189/signatures?p_auth=BTkvMdrh (accessed 8 April 2017).
19. See note 2 above.
20. See note 3 above.
21. s 127 of Communications Act 2003.
22. www.cps.gov.uk/legal/a_to_c/communications_sent_via_social_media/#a09 (accessed 8 April 2017).
23. www.bbc.co.uk/editorialguidelines/guidelines/harm-and-offence/portrayal (accessed 8 April 2017).

24. The BBC publishes details of upheld complaints on its website, www.bbc.co.uk/complaints/reports (accessed 8 April 2017). The BBC Trust publishes details of appeals on its website, www.bbc.co.uk/bbctrust/our_work/complaints_and_appeals/esc.html (accessed 8 April 2017).
25. www.ipso.co.uk/editors-code-of-practice (accessed 8 April 2017).
26. https://edri.org/edri-access-now-withdraw-eu-commission-forum-discussions (accessed 8 April 2017).
27. For an account of what this compromise might look like, see Brown (2015: ch. 10).

References

Alexander L (1996) Banning hate: Speech and the sticks and stones defense. *Constitutional Commentary* 13: 71–100.

Baker CE (2009) Autonomy and hate speech. In: Hare I, Weinstein J (eds) *Extreme Speech and Democracy*. Oxford: Oxford University Press.

Baker CE (2012) Hate speech. In: Herz M, Molnar P (eds) *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. Cambridge: Cambridge University Press.

Barendt E (2016) *Anonymous Speech*. Oxford: Hart.

Bleich E (2011) *The Freedom to Be Racist? How the United States and Europe Struggle to Preserve Freedom and Combat Racism*. Oxford: Oxford University Press.

Branscomb AW (1995) Anonymity, autonomy, and accountability: Challenges to the first amendment in cyberspace. *Yale Law Journal* 104: 1639–1679.

Brison S (1998) The autonomy defense of free speech. *Ethics* 108: 312–339.

Brown A (2008) The Racial and Religious Hatred Act 2006: A Millian response. *Critical Review of International Social and Political Philosophy* 11: 1–24.

Brown A (2015) *Hate Speech Law: A Philosophical Examination*. New York: Routledge.

Brown A (2016) The ''who?'' question in the hate speech debate: Part 1: Consistency, practical, and formal approaches. *Canadian Journal of Law & Jurisprudence* 29: 275–320.

Brown A (2017) The ''who?'' question in the hate speech debate: Part 2: Functional and democratic approaches. *Canadian Journal of Law & Jurisprudence* 30: 23–55.

Brown A (2017a) What is hate speech? Part 1: The myth of hate. *Law and Philosophy* 36: 419–468.

Brown A (2017b) What is hate speech? Part 2: Family resemblances. *Law and Philosophy* 36: 561–613.

Brown A (2017c) The politics behind the introduction of stirring up religious hatred offences in England and Wales. *Politics, Religion, and Ideology* 18: 42–72.

Brown A (2017d) Hate speech laws, legitimacy, and precaution: A reply to James Weinstein. *Constitutional Commentary* 32.

Citron DK (2014) *Hate Crimes in Cyberspace*. Harvard, MA: Harvard University Press.

Citron DK and Norton H (2011) Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review* 91: 1435–1484.

Coffey B and Woolworth S (2004) ''Destroy the scum, and then neuter their families'': The web forum as a vehicle for community discourse? *The Social Science Journal* 41: 1–14.

Cohen-Almagor R (2011) Fighting hate and bigotry on the Internet. *Policy and Internet* 3: 1–26.

Cohen-Almagor R (2015) *Confronting the Internet's Dark Side: Moral and Social Responsibility on the Free Highway*. Cambridge: Cambridge University Press.

Craddock JM (1995) Words that injure, laws that silence: Campus hate speech codes and the threat to American education. *Florida State University Law* 22: 1047–1089.
Delgado R (1991) Campus antiracism rules: Constitutional narratives in collision. *Northwestern University Law Review* 85: 343–387.
Delgado R and Stefancic J (1996) Ten arguments against hate-speech regulation: How valid? *Northern Kentucky Law Review* 23: 475–490.
Delgado R and Stefancic J (2009) Four observations about hate speech. *Wake Forest Law Review* 44: 353–370.
Delgado R and Stefancic J (2014) Hate speech in cyberspace. *Wake Forest Law Review* 49: 319–343.
Delgado R and Yun DH (1994a) The neoconservative case against hate speech regulation: Lively, D'Souzu, Gates, Carter, and the Toughlove Crowd. *Vanderbilt Law Review* 47: 1807–1825.
Delgado R and Yun DH (1994b) Pressure valves and bloodied chickens: An analysis of paternalistic objections to hate speech regulation. *California Law Review* 82: 871–892.
Douglas KM, et al. (2005) Understanding cyberhate: Social competition and group creativity in online white supremacy groups. *Social Science Computer Review* 23: 68–76.
Downs DA (1993) Codes say darnedest things. *The Quill* 81: 19.
Downs DA (2005) *Restoring Free Speech and Liberty on Campus*. New York, NY: Cambridge University Press.
Dworkin R (2012) Reply to Jeremy Waldron. In: Herz M, Molnar P (eds) *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. Cambridge: Cambridge University Press.
Emerson TI (1963) Toward a general theory of the First Amendment. *Yale Law Journal* 72: 877–956.
Fuchs C (2014) *Social Media: A Critical Introduction*. London: SAGE.
Gagliardone I, et al. (2014) *Mapping and Analysing Hate Speech Online: Opportunities and Challenges for Ethiopia*. Oxford: University of Oxford Programme in Comparative Media and Law Policy.
Gelber K and McNamara L (2014) Changes in the expression of prejudice in public discourse in Australia: Assessing the impact of hate speech laws on letters to the editor 1992–2010. *Australian Journal of Human Rights* 20: 99–128.
Gelber K and McNamara L (2016) Evidencing the harms of hate speech. *Social Identities* 22: 324–341.
Gellman S (1991) Sticks and stones can put you in jail, but can words increase your sentence? *UCLA Law Review* 39: 333–396.
Graham G (1999) *The Internet: A Philosophical Examination*. Routledge: London.
Hare I (2012) The harms of hate speech legislation. freespeechdebate.com, March 23. Available at: http://freespeechdebate.com/en/discuss/the-harms-of-hate-speech-legislation.
Heins M (1983) Banning words: A comment on "words that wound". *Harvard Civil Rights-Civil Liberties Law Review* 18: 585–592.
Heinze E (2016) *Hate Speech and Democratic Citizenship*. Oxford: Oxford University Press.
Jacobson A and Schlink B (2012) Hate speech and self-restraint. In: Herz M, Molnar P (eds) *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. Cambridge: Cambridge University Press.
Jay T (2009) Do offensive words harm people? *Psychology, Public, Policy, and Law* 15: 81–101.

Kang J (2000) Cyber-race. *Harvard Law Review* 113: 1130–1208.
Lawrence C (1990) If he hollers let him go: Regulating racist speech on campus. *Duke Law Journal* 1990: 431–483.
Leets L (2002) Experiencing hate speech: Perceptions and responses to anti-Semitism and antigay speech. *Journal of Social Issues* 58: 341–361.
McNamee LG, et al. (2010) A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. *Communication Monographs* 77: 257–280.
Magruder C (1936) Mental and emotional disturbance on the law of torts. *Harvard Law Review* 49: 1033–1067.
Malik K (2005) Hate speech in a plural society. Index On Censorship event: Incitement, hate speech and the right to free expression, Lancaster House, London, December 8–9, 2005. Available at: www.kenanmalik.com/debates/freespeech_index.html
Musolff A (2015) Dehumanizing metaphors in UK immigrant debates in press and online media. *Journal of Language Aggression and Conflict* 3: 41–56.
Parekh B (2005–2006) Hate speech: Is there a case for banning? *Public Policy Research* 12: 213–223.
Perry B and Olsson P (2009) Cyberhate: The globalization of hate. *Information and Communication Technology Law* 18: 185–199.
Poland B (2016) *Haters: Harassment, Abuse, and Violence*. Lincoln, NE: University of Nebraska Press.
Posner R (2002) The speech market and the legacy of Schenck. In: Bollinger L, Stone G (eds) *Eternally Vigilant: Free Speech in the Modern Era*. Chicago, IL: University of Chicago Press.
Post R (2012) Interview. In: Herz M, Molnar P (eds) *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. Cambridge: Cambridge University Press.
Shiell TC (2009) *Campus Hate Speech on Trial*, 2nd ed. Lawrence, KS: University of Kansas Press.
Smith SA (1995) There's such a thing as free speech: And that's a good thing, too. In: Whillock K, Slayden D (eds) *Hate Speech*. Thousand Oaks, CA: Sage Publications.
Smolla R (1990) Academic freedom, hate speech, and the idea of a university. *Law and Contemporary Problems* 53: 195–226.
Solomos J and Schuster L (2002) Hate speech, violence and contemporary racism. In: Evens Foundation (ed) *Europe's New Racism? Causes, Manifestations and Solutions*. Oxford: Berghahn.
Strossen N (1990) Regulating racist speech on campus: A modest proposal. *Duke Law Journal* 1990: 484–573.
Strossen N (2012) Interview. In: Herz M, Molnar P (eds) *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. Cambridge: Cambridge University Press.
Suler J (2004) The online disinhibition effect. *Cyber-Psychology and Behavior* 7: 321–326.
Sunstein C (2007) *Republic.com 2.0*. Princeton, NJ: Princeton University Press.
Taylor J (2006) Humean humanity versus hate. In: Welchman J (ed.) *The Practice of Virtue: Classic and Contemporary Readings in Virtue Ethics*. Indianapolis, IN: Hackett.
Tsesis A (2001) Hate in cyberspace: Regulating hate speech on the Internet. *San Diego Law Review* 38: 817–874.
Tsesis A (2017) Campus speech and harassment. *Minnesota Law Review* 101.
Waldron J (2012) *The Harm in Hate Speech*. Cambridge, MA: Harvard University Press.

Weinstein J (2009) Extreme speech, public order, and democracy: Lessons from the masses. In: Hare I, Weinstein J (eds) *Extreme Speech and Democracy*. Oxford: Oxford University Press.

Weinstein J (2017) Hate speech bans, democracy and political legitimacy. *Constitutional Commentary* 32.

Woeste VS (2012) *Henry Ford's War on Jews and the Legal Battle Against Hate Speech*. Stanford, CA: Stanford University Press.