The Routledge Handbook of Language in Conflict

Matthew Evans, Lesley Jeffries, Jim O’Driscoll

Hate speech

Publication details

Sharon Millar
Published online on: 23 May 2019

How to cite: Sharon Millar. 23 May 2019, Hate speech from: The Routledge Handbook of Language in Conflict Routledge
Accessed on: 26 Sep 2023
8

Hate speech

Conceptualisations, interpretations and reactions

Sharon Millar

8.1 Introduction

Hate speech has increasingly become a source of societal and political concern across the globe, as witnessed by recent measures and initiatives to tackle it. For instance, the Japanese Parliament introduced its first hate speech law in 2016 to deal with rising ethnic tensions, particularly directed towards residents of Korean descent; the Law Commission in India has proposed a tightening up of hate speech law because of heightened anti-minority political rhetoric; the European Commission and the major global technology companies (Google, Facebook, Twitter and YouTube) agreed on a code of conduct in 2016 to remove illegal hate speech from their platforms within 24 hours; the Norwegian government announced in 2017 a strategy against hateful expressions. At the same time, in the wake of political upheavals, increased migration and terrorist attacks, what might be viewed as trigger events (Burnap and Williams, 2015), there have been reports from human rights organisations and official institutions of a rise in hate incidences against minorities in many countries, including the US, the UK, Germany, France, Poland, Turkey, South Sudan, China and Russia. The internet offers a myriad of spaces where hate speech thrives and, although difficult to monitor, there are indications that incidences of hate speech online are rising (Banks, 2010; Bartlett et al., 2014; Foxman and Wolf, 2013; Gagliardone et al., 2015). It is these online environments that are attracting increasing political attention.

Hate speech is clearly a global phenomenon and one which rests on some form of inter-group antagonism. Referring to online hate speech, Gagliardone et al. (2015, p.7) argue that it:

is situated at the intersection of multiple tensions: it is the expression of conflicts between different groups within and across societies; it is a vivid example of how technologies with a transformative potential such as the Internet bring with them both opportunities and challenges; and it implies complex balancing between fundamental rights and principles, including freedom of expression and the defence of human dignity.

Hate speech then feeds off conflict, but it also can feed into it as illustrated most strikingly by cases of genocide and atrocities in Nazi Germany, Rwanda, Kenya and the Balkans.
(Gordon, 2017; Thompson, 2007). For example, the use of dehumanising metaphors in the media to construct outgroups as the enemy and a threat to be eliminated have been frequently observed in relation to subsequent atrocities, such as Jews being portrayed as vermin in Nazi Germany or the Tutsis as dirt or cockroaches in Rwanda during the 1990s (Gordon, 2017). In some cases, hate speech has been directly linked to ethnic violence, such as during the 2007 elections in Kenya (Halakhe, 2013).

However, the notion of hate speech itself is highly controversial and is argued to be an “essentially contested concept”, characterised by continual dispute and normative disagreement (Boromisza-Habashi, 2010, p.276). There is, for instance, no accepted definition of hate speech in law and indeed, the term itself is generally not used in legislation. Brown (2017, p.565) suggests that the lack of conceptual transparency of hate speech is best tackled through a mapping of “the stunning heterogeneity of the phenomena to which it refers and as much of the vast array of different connotations it carries as possible”. Similarly, Boromisza-Habashi (2013, p.5) argues that hate speech has multiple local meanings that are contextually embedded and the objective must be to “capture the complicated cultural life of hate speech”.

Acknowledging this heterogeneity and multiplicity, the aim of this chapter is to explore understandings of online hate speech and its social control from institutional and corporate perspectives. Such perspectives are important as it is these institutional and corporate actors that play a fundamental role in establishing, disseminating and enforcing norms for acceptable online communication: what may be expressed and what should be suppressed, and why. As noted by Blommaert (2007, p.2):

people continuously need to observe “norms” – orders of indexicality – that are attached to a multitude of centers of authority, local as well as translocal, momentary as well as lasting.

For Blommaert, norms involve indexical meaning, which “anchors language usage firmly into social and cultural patterns” (Blommaert, 2005, pp.11–12). As differing normative authorities operate at various levels of scale, for example, a parent, a school, a state, norms involve vertical dimensions of scale predicated on stratification and different power differentials. Hence, norms can index different social meanings depending on the scalar level; for example, a regional accent may be assigned a more positive value in local than national or international contexts. My concerns are with higher levels of scale – European bodies, global social media corporations and the state – and how these centres of authority construct norms for online public discourse through their textual representations of hate speech.

8.2 Hate speech as practice

It is quite common to find the label “hate speech” written in scare quotes. Boromisza-Habashi (2013, p.2) explains his graphical usage as a “visual reminder that the term is not a simple window on reality”. Hate speech is not a pre-defined object that can be described, but rather its meanings must be found in language use. For example, in Hungary, he found that media debates about hate speech often revolved around the content versus the tone of expression. I will similarly adopt a practice approach, proceeding from “doing to meaning” (ibid., p.5), and avoiding the “infinite regress of definitions” that occurs when trying as an analyst to establish a definition of hate speech (Janicki, 2015, p.77). Janicki suggests that defining hate speech will ultimately prove pointless, but I would argue that institutional
Hate speech attempts to secure definitions are an important part of the normative social actions surrounding hate speech. For example, Aldo Patriciello, an Italian Member of European Parliament, put forward a motion to the European Parliament in March 2017, asking for the European Commission to explore the feasibility of having a single legal definition of hate speech across EU member states.

Within this practice orientation, I adopt insights from van Leeuwen’s (2008, 2016) approach to discourse as the recontextualisation of social practice. He argues that discourses, defined as “context-specific frameworks of making sense of things”, are based on social practices, but “transform” these practices through, for instance “including and excluding different things, and doing so in the service of different interests” (van Leeuwen, 2016, p.138). Actual social practices are characterised by certain elements, such as actions, actors, times, spaces, resources, and it is these that are transformed in discourses about the social practices. Discourses may not represent certain aspects of the social practice (e.g. actors may be deleted) or may “add reactions and motives to the representation of social practices” (ibid., p.144). For van Leeuwen, discourse is not the same as text, but by means of textual analysis, discourses can be identified or reconstructed on the basis of similarities and repetitions across different texts. To transfer this thinking to hate speech, we can say that hate speech and the means to counter it (such as monitoring, censorship, legal punishment) are performed social practices, discourses of hate speech and combatting hate speech represent these practices, and through examination of a variety of texts, we can find evidence of these discourses. Van Leeuwen (2008, 2016) identifies numerous components for this type of analysis, but I will focus particularly on the following: a) social actions (e.g. recognising/identifying hate speech, flagging/reporting online comments and posts, removing online content), and social actors (participants in the action); b) the transforming of social practices in discourse through deletion (not representing an aspect of a social practice) and the addition of reactions (how actors react) and motives (purposes of, and legitimations for, the action); and c) the role of times and spaces, where I will also take account of different dimensions of scale where relevant (Blommaert, 2007).

The analysis requires different types of texts and I have selected a number produced by, or under the auspices of, institutional and corporate bodies operating at different scalar levels. The global level is represented by two of the global technology corporations, Facebook and Twitter, while the European level will be represented by two European organs, the Council of Europe and the European Commission. The European institutions are major social actors with regard to hate speech, but they are fundamentally different. The Council of Europe is a European human rights organisation with members from 47 countries, including the EU member states. The European Commission is an EU executive body, consisting of representatives from all 28 EU member states, and is responsible for drawing up EU legislation and implementing decisions. The national level will focus on one EU member state, Denmark, and consider how hate speech online has been interpreted as criminal or not in national case law. The choice of Denmark owes much to my participation in an EU action project (CONTACT) on hate speech, which involved collecting data in the Danish context (Assimakopoulos et al., 2017).

The Council of Europe’s understandings of hate speech will primarily be investigated through Bookmarks: a manual for combating hate speech online through human rights education (Keen and Georgescu, 2016) and published in ten languages as part of the Council’s youth campaign, the No Hate Speech Movement. According to Thorbjørn Jagland, the Secretary General of the Council of Europe, this campaign was “launched to reduce the acceptance of hate speech online and put an end to its ‘normalisation’” (Bookmarks, p.3). The European
Commission and the global technology companies will be considered together in light of the common code of conduct on hate speech agreed upon in May 2016, again because of political concerns about the extent of hate speech online. In addition to this code, Facebook and Twitter policies on hate speech will be drawn upon. Danish case law is accessed through the public database of the Office of Public Prosecutions (Anklagemyndigheden), which provides summaries of court decisions (but not transcriptions of court proceedings) and practical guidelines for law enforcement and legal professionals on how to interpret the relevant Danish legislation.

While the texts come from different genres (training manuals, policy guidelines, summaries of court proceedings), they all relate to internet governance to a greater or lesser degree. They can also be seen as representing actions and reactions to hate speech in the sense that they are normative documents dealing with its social control, be this through education, awareness-raising, censorship or criminalisation. Importantly, they are all framed by international conventions on freedom of expression and human rights and hence, intertextuality is in evidence in the form of, for example, direct quoting, using similar formulations, lexical choices or arguments from the conventions and statutes. Citing all the relevant instruments is beyond the scope of this contribution, but I will point to some of the most salient. Both Article 29 of the Universal Declaration of Human Rights (1948) and Article 10(2) of the European Convention on Human Rights (1950) formulate restrictions on freedom of expression in terms of protection of the rights and freedoms of others, public order or safety and morality. From a more legal perspective, Article 20 of the International Covenant on Civil and Political Rights (1966) and Article 4 of the International Convention on the Elimination of Racial Discrimination (1969) sanction activities such as incitement to racial discrimination and violence or dissemination of ideas of racial superiority and hatred. The criminalisation of racist and xenophobic written, visual and other expression online that involve advocacy, promotion or incitement of hatred, discrimination and violence was introduced by the Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems (2006).

As the texts deal with hate speech, they have metalinguistic dimensions since they include reflection about, and evaluation of, language use. As such, they are not simply “reflexive activity” (Lucy, 1993, p.17) about language use, but “ideologically saturated” (Silverstein, 2003, p.196), involving moral, social and political beliefs. As noted by Cameron (2004, p.313) “metalinguistic resources seem very often to be deployed to connect various aspects of linguistic behaviour to a larger moral order”. The textual analyses will include these metalinguistic aspects, particularly in relation to two key pragmatic concepts which are commonly appealed to in hate speech discourses. These are the intention of the communicator and the effect of the expression he/she produces on the receiver, often referred to as perlocution or perlocutionary effect in the academic literature.

8.3 The Council of Europe and hate speech

The Council of Europe, which is the major human rights actor in Europe with regard to combatting hate speech, locates its work at the interface between the right of freedom of expression and the need to protect the rights of vulnerable groups. For example, one of its current webpages on hate speech is embedded under “media” and “freedom of expression”, and emphasises that, in working to counter hate speech, the Council adopts:

- a “freedom of expression perspective” which focuses on co-operation with member states in preparing, assessing, reviewing and bringing in line with the European
Hate speech

Convention on Human Rights any laws and practices that place restrictions on freedom of expression.

(https://www.coe.int/en/web/freedom-expression/hate-speech)

The Council then legitimises its approach to hate speech as one resting on the authority given by the European Convention of Human Rights to protect freedom of expression and guard against restrictive regulation. The need to emphasise freedom of expression derives from the fact that addressing hate speech often involves the control of what may be expressed. In the preface to Bookmarks (p.3), the Secretary General of the Council of Europe adopts a legitimisation strategy of authoritative argument, tinged with morality, to counter such concerns:

The manual Bookmarks you have in your hands is a precious tool to stop hate speech and strengthen human rights. You may ask: ‘Why should we bother? Don’t people have the right to express themselves freely in a democratic society?’ It is true that freedom of expression is a fundamental human right that applies also to ideas that may offend, shock or disturb people. But exercising this right carries clear duties and responsibilities. Hate speech is not ‘protected’ speech; words of hate can lead to real-life crimes of hate, and such crimes have already ruined and taken the lives of too many people.

The association of rights with duties and responsibilities derives from a human rights discourse that acknowledges moral duty in relation to rights (Boot, 2017) and is intensified through the association of hate speech with criminal activities. Hate speech is morally wrong and, hence, must be countered. It follows then that for Bookmarks taking action is paramount, but action is also constructed as part of the definition of hate speech. The reader is told that:

Hate speech is rarely a black-and-white, yes-or-no concern. Opinions differ both over how it should be classified and over what we should do about it. Part of the reason for the difference of opinion is that these questions are seen by most people as related: if something is classed as hate speech, it seems to warrant some action. If it is not, we assume it is acceptable, or at least, that it should be tolerated. That means that the definition we use also seems to tell us when we should act.

(Bookmarks, p.148)

So, defining hate speech is combined with the social action of taking (non-specified) material, real-world action. What is noteworthy is how acceptability and toleration are aligned with not being hate speech; this implicitly broadens the notion of hate speech to include all that is judged to be unacceptable or not to be tolerated and associates such expression with action. The social actors are represented as “we”, suggesting an inclusive civic responsibility. It should be borne in mind that the No Hate Speech Movement (of which Bookmarks forms a part) promotes empowerment of youth through participation in internet governance, encouraging active user regulation rather than relying on web providers.

The issue of action is not new to the notion of hate speech, but the typical perspective has been that of the producer of such speech and action in the sense of speech acts, that is expressions that may in themselves constitute an action, such as threats; the act of saying/writing “I will kill you” is, in essence, a threatening act. Associated with speech acts is communicator intention (e.g. in saying/writing “I will kill you”, my intention was to threaten rather than
Sharon Millar

to merely express frustration or anger) and perlocutionary effect, which concerns the effect of an expression on the receiver (e.g. the person feels threatened) (see Birner [2013] for an introduction). Academic approaches to hate speech have been inspired by speech act theory (e.g. Butler, 1997; Langton, 2012; McGowan, 2009), as have institutional definitions of hate speech. Coming back to the Council of Europe’s webpage, we are told that:

Hate speech has no particular definition in international human rights; it is a term used to describe broad discourse that is extremely negative and constitutes a threat to social peace. According to the Committee of Ministers, hate speech covers all forms of expressions that spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance.

Importantly, hate speech is not understood solely in terms of propositional content, which must be not simply negative but “extremely negative”. It has to include some form of action, in this case specified as spreading, inciting, promoting or justifying forms of hatred. These actions can be understood as social actions, which may involve a variety of speech acts (e.g. threats, assertions, accusations, commands, complaints, suggestions). That it is these social actions which are specified is no coincidence as they derive from an earlier definition of hate speech given by the Council of Europe in 1997 (Recommendation No. R (97) 20 of the Council of Europe’s Committee of Ministers to Member States on “Hate Speech” (the scare quotes in the original document)). Note too that these actions which constitute hate speech must have a specific perlocutionary effect, that of constituting “a threat to social peace”. Within peace studies, social peace refers to issues of inclusion and identity, rather than security, and widens processes of peacebuilding to include non-security dimensions and social actors other than the state (Lee et al., 2016). So, from this perspective, the perlocutionary effect required for an expression to be considered hate speech must operate beyond the individual addressed by the communicator, or the individual or group targeted, to include society at large. This is similar to hate speech legislation that focuses on threats to public order; as noted by Brown (2015, p.28), such laws aim to “protect not only the people targeted by hate speech but also the wider community”. Insisting on a perlocutionary impact of threat to social peace raises the bar for an expression to be considered hate speech, but more frequently, perlocutionary effects are formulated in terms of the emotional impact on the individuals and groups addressed or targeted (denigration, humiliation, marginalisation, etc.).

Bookmarks reproduces the Council of Europe’s 1997 definition and in so doing sustains an understanding of hate speech across time, 20 years, and also across medium since it now serves to define hate speech online. How the definition is to be interpreted is not directly elaborated upon, but instead, Bookmarks provides a classification of hate speech, the purpose of which is instrumental to aid the recognition of hate speech. Combatting hate speech as a social practice clearly encompasses social actions such as identifying what hate speech is and selecting an appropriate response. In Bookmarks, recognising hate speech is represented in terms of a continuum, which in itself is not unusual, but here the poles of the spectrum are labelled with the evaluative terms “bad” and “worse” (Bookmarks, p.151). The lexical choices may well be due to the focus of the manual on educating youth, but it also anchors hate speech as inherently negative and unacceptable, regardless of its legal status or intensity. What is especially revealing for my purposes is that brief examples of this spectrum are given according to some of the typical definitional components of hate speech: content and tone of expression, intent of the speaker, targets, impact of the expression. I will

150
Hate speech

illustrate with intent (intention) and impact (perlocutionary effect) as these are important if problematic notions in pragmatics (Haugh, 2008) and how they are represented in hate speech discourses could give interesting metapragmatic insights.

Speaker intent is exemplified by the same phrase “wipe out gays” but written in two different contexts: “a private email to a friend – as a ‘joke’” (placed as bad on the continuum) and “someone’s Facebook page, knowing that he’s gay” (placed as worse) (Bookmarks, p.152). The latter is worse because it exemplifies “an intention to hurt” and hence it requires a different response (ibid., p.151). There are two points to be made here. One is the significance given to context in the evaluation of intention. As a metapragmatic object, intention can only be evaluated retrospectively and, without recourse to the person who produced the expression, such evaluation relies on contextual factors and plausible assumptions (Edwards, 2008). Here, pivotal factors are addressee (as member or not of a minority group) and medium (a one-to-one email correspondence versus a one to many post on a personal Facebook page). However, the examples also suggest a hierarchy of intention in that joking about a group is deemed less unacceptable than to directly hurt a member of the group.

The second point is that intention to hurt (assuming “hurt” refers to emotional hurt) is not generally included in institutional definitions of hate speech so Bookmarks is expanding on official understandings. The European Court of Human Rights notes, from the perspective of perlocutionary effects, that information and ideas that “offend, shock or disturb” are protected by freedom of expression. Similarly, although with greater restrictions, the European Commission against Racism and Intolerance (ECRI), a human rights monitoring body under the Council of Europe, stipulates that “satire or objectively based news reporting and analysis that merely offends, hurts or distresses” is not hate speech (ECRI, 2016, p.17). It would seem then that the perlocutionary effect of being “hurt” is not necessarily enough to be hate speech.

Yet when Bookmarks (p.153) emphasises the significance of impact on others in evaluating hate speech, ideas of emotional distress are included:

The actual or potential impact on individuals, groups, or society as a whole is one of the most important considerations in assessing an expression of hate, and in weighing up our response. How a person or group was in fact affected is often more important than how outsiders feel they should have been affected. For example, if a child is severely troubled by comments that others claim to be making in a “friendly” way, the actual pain will probably be more important in simply allowing those others to “express their opinions”.

Actual experiences of targeted individuals and groups would seem to be part of the represented process of recognising hate speech. Yet ultimately these are effects as perceived by the social actors involved in the action of recognising. Bearing in mind that Bookmarks is about combatting hate speech through increased internet literacy, encouraging “critical thinking and information processing” (p.184), it primarily adopts a hearer/receiver-centred perspective on the communicative productions of others.

A noticeable aspect in Bookmarks that has relevance for the recognition of hate speech is the association of hate speech with a range of other concepts. These are modelled as feeding into hate speech in a linear fashion: stereotypes → prejudice → racism → discrimination → hate speech → hate crime (Bookmarks, p.168). This model reflects the plant metaphor for hate speech used to introduce the manual: hate speech as “the leaves of a particularly malicious plant, whose roots lie deep in society”. (Bookmarks, p.1). The argument that...
Stereotypes and hate speech are related is not novel; for example, Haas (2012, p.130) notes in relation to psychological work on prejudiced communication:

Stereotypic talk and hate speech should be understood as existing on a continuum of prejudicial messages. Hate speech constitutes one end of this continuum with subtle, perhaps unconscious, verbal and nonverbal messages falling on the opposite side.

However, this association of hate speech with stereotyping seems to have evolved into a relationship of equivalence: hate speech is negative stereotyping. This is clear, for example, from the definition of hate speech in ECRI’s Recommendation No. 15, On Combatting Hate Speech (2016, p.16) which, in addition to hate speech as “advocacy, promotion or incitement”, includes hate speech as “any harassment, insult, negative stereotyping, stigmatization or threat”. Bookmarks too encompasses what Haas refers to as “stereotypic talk” within hate speech when it addresses the identification of hate speech:

The first task in the battle against hate speech online is being able to identify it when we come across it. This requires knowing what constitutes hate speech and knowing how to assess the possible impact, but it can also demand a deeper awareness of underlying messages and the ability to spot bias and prejudice where these are only implicit.

Consequently, the social action of recognising hate speech is represented as having to ask questions about a range of issues, such as “What is the intention?”, “How ‘bad’ is the expression?”, “Is it hurtful? Damaging? Dangerous?” “Are there any underlying messages?”, “is it true?” (ibid., p.184). The assumption is that these questions are not only answerable but answerable by the social actors Bookmarks is targeting, namely young internet users. Yet, for instance, identifying communicator intention is acknowledged to be notoriously difficult in the scholarly literature (Haugh, 2008). Kecskes (2014, p.32) argues that “what is actually ‘recoverable’ is hearer meaning” and this will be individual since people have “different cognitive predispositions, different commitments, different prior experiences, and different histories of use of the same words and expressions”. Similarly, decisions about the truth of propositional content, needed because “false information […] feeds prejudice against particular groups” (Bookmarks, p.184), are likely to be individual. While Bookmarks gives blueprints and examples to guide normative understandings of hate speech, there will inevitably be a strong subjective component to any endeavour in hate speech recognition.

One issue that remains to be discussed is the space where social actors are expected to be agentive in tackling hate speech. Bookmarks is driven by an ideology of democracy as participatory action and, in this regard, the manual asserts that “online space is public space, and hence, all principles of a democratic society can and should apply online” (Bookmarks, p.8). This statement is noteworthy in that it discursively removes the boundary between private and public in digital contexts and its consequences for the representation of social actions surrounding hate speech need to be considered. In discussing the blurring of the public and private, Livingstone (2005, p.169) argues that the extent of their interpenetration can be assessed along three main dimensions. profit, participation and governance. Under profit, public can be understood as “disinterested” as in the public interest or the public sector whereas private refers to “self-interest” and commercial gain as in the private sector or private enterprise. Under participation, public means “connected or engaged in a shared culture” as in public opinion, public space while private deals with “individualisation” as
in private opinion and personal choice. Under governance, public concerns visibility and accountability as in the public eye, in contrast to private, which is about (rights to) privacy and being “beyond surveillance”. Applying this framework to Bookmarks, there seems to be a conflation of public and private in relation to participation and governance. Constructing online space as public space extends not only the public sphere so that private views function as public opinion, but also the reach of social actors, including the State and internet users, on the governance of these online spaces. Of course, Bookmarks strongly and repeatedly emphasises rights to a private life on the internet, but nonetheless, there remains a dilemma. Discursively erasing the idea of private online spaces for the purpose of combatting hate speech, inevitably compromises the representation of privacy rights in cyberspace. Interestingly, it seems to be the profit dimension that is a motivating factor here since private ownership of online spaces is seen to be detrimental to applying democratic principles to hate speech.

8.5 The European Commission, social media companies and hate speech

Turning now from human rights education to codes and policies of regulation, a notable event occurred in May 2016, when the European Commission announced a code of conduct on countering illegal hate speech online which had been agreed with Facebook, Twitter, Microsoft and YouTube. As part of this code, the companies vowed to remove within 24 hours illegal hate speech, defined according to the EU Framework Decision 2008/913/JHA 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law. This definition focuses exclusively on incitement, covering “all conduct 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”. As such, it focuses on communicator intention rather than perlocutionary effect and sets a relatively high threshold for what is to be considered as illegal hate speech. Moreover, as the definition is taken from a decision concerning racism and xenophobia, the protected categories listed belong to that domain. This means that characteristics such as gender and sexual orientation have been deleted from the representation of hate speech in the code.

The legitimation of the code of conduct is, like Bookmarks, framed within the right of freedom of expression and concerns that illegal hate speech is silencing voices and hindering participation in democratic debate. Hence regulation is positioned within an ideology of participatory democracy. This is further emphasised by other sections of the code, which go beyond illegal hate speech and bring in alternative measures and other civic actors to counter broader issues of prejudice. To this end:

The IT Companies and the European Commission, recognising the value of independent counter speech against hateful rhetoric and prejudice, aim to continue their work in identifying and promoting independent counter-narratives, new ideas and initiatives and supporting educational programs that encourage critical thinking.


In this way, the code of conduct shares similar ideals to the human rights initiatives of the Council of Europe. Significantly, it also represents the social media companies as legitimate social actors in the identification and removal of illegal hate speech.
Of course, independently of the European Commission, the global technology companies do represent themselves as social actors in relation to combatting hate speech and they have their own corporate policies, although different terminology may be used; Twitter, for example, does not refer to hate speech but to “hateful conduct”. Both Twitter and Facebook use similar legitimation strategies as the Council of Europe and the European Commission: hate speech regulation is framed in terms of its effects on freedom of expression, participation and even real-world behaviours. Twitter informs its users that:

> Freedom of expression means little if voices are silenced because people are afraid to speak up. We do not tolerate behavior that harasses, intimidates, or uses fear to silence another person’s voice.


In April 2018, in the wake of considerable criticism of its actual and represented social actions surrounding online content (discussed below), Facebook released an updated and more detailed version of their community standards, where hate speech is placed under “objectionable content”, arguing that:

> We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence.

(www.facebook.com/communitystandards/objectionable_content/)

By associating hate speech with a lack of safety (intimidation, fear, even violence), the companies legitimise their policies for removing it. They operate with a much lower threshold for hate speech compared with the EU code of conduct, in terms of the range of social actions represented and the protected characteristics included. Both companies extend protected characteristics to include gender, sexual orientation, disease, disability and, in the case of Twitter, age. The companies differ, however, in how they represent hate speech, bearing in mind that this is done for the user as a potential producer of hate speech. In the Twitter policy, the representation is dynamic using verbal forms, presumably because the focus is on hateful conduct. Users “may not promote violence against or directly attack or threaten others” on the basis of protected characteristics (same link as above). Generalised examples of what will not be tolerated include “violent threats”, “behavior that incites fear” and “repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone”. Compared with Twitter, Facebook shows a preference for what Van Leeuwen (2016, p.149) refers to as “deactivated representations” where actions are represented as “though they are entities or qualities”. In line with their location of hate speech under the link “objectionable content”, Facebook views hate speech as “a direct attack” on the basis of protected characteristics, objectivating hate speech as an entity through nominalisation (an attack versus to attack). Accordingly, what is proscribed is the action of posting such “content”. Users are told “do not post” and there then follows a sub-division of attacks into three tiers of severity: in descending order, “violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation” (same link as above). The hierarchy is decontextualised, focusing on the general forms and content of the different levels of attack. For instance, at tier 1, attack includes “Any violent speech or support for death/disease/harm” and “Dehumanizing speech”, which focuses on dehumanising metaphors relating to “filth, bacteria, disease or faeces” and “animals that are culturally perceived as intellectually or physically inferior”, as well as “reference or comparison to sub-humanity”.

154
Tier 2 attacks include “Statements of inferiority or an image implying a person’s or a group’s physical, mental, or moral deficiency” (e.g. “deformed”, “ugly”, “retarded”, “idiot”, “slutty”, “free riders”), “Expressions of contempt or their visual equivalent” (e.g. “I hate”, “I don’t like”, “X are the worst”) and “Expressions of disgust or their visual equivalent” (e.g. “vile”, “gross” and “cursing at a person or group of people who share protected characteristics”). Tier 3 is not detailed beyond “calls for exclusion or segregation”, but what is permitted is “criticism of immigration policies and arguments for restricting those policies”. Interestingly, slurs are included in this lowest tier of severity and, hence, racial, ethnic and sexual terms of insult are represented as less serious than, for example, swearing at members of a minority group.

A notable feature of Facebook’s approach to hate speech is what is represented as an eligible target to be protected at each tier of severity. All tiers encompass individuals or groups based on the specified protected characteristics (“race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender identity and serious disability or disease”). However, Tier 1 is the only level that explicitly adds “immigration status”, hence giving refugees and asylum seekers recourse to protection, as well as “all subsets except those described as having carried out violent crimes or sexual offences”. The latter has to be understood in the light of the criticism Facebook received after training documents on content moderation were leaked in May 2017 and reported on by the British newspaper *The Guardian* and *ProPublica*, an online investigative newsroom (Angwin and Grassegger, 2017). Here it was revealed that Facebook did not include subsets of protected categories when enforcing its hate speech policy. This meant that, for example, “black children” were not eligible for protection against hate speech as they are a subset of the protected category of race, but “white men” were as both characteristics are protected (race and gender). In their new rules, all subsets will be protected (black children and white men) but this is only specified for Tier 1 attacks.

Generally, social media corporations have not been forthcoming about how they implement their hate speech rules/standards. However, both Twitter and Facebook deal with “enforcement” in their policies, representing this as contextual, even though their representations of what hate speech is mainly are decontextualised. Twitter refers to the difference between viewing perceived abusive content “in isolation” and “in the context of a larger conversation”. Facebook’s approach is found in a link from the hate speech page to its “hard questions blog”, where the vice-president of the EMEA region, Richard Allan, asks, “Who Should Decide What is Hate Speech in an Online Global Community?” in a blog written in June 2017. Note that this blog was released approximately one month after the Facebook manuals on online moderation were leaked but predates the new rules on hate speech by ten months. Context is primarily discussed in geographical and cultural terms and reflects Facebook’s acknowledgement of scale when enforcing its global community standards. Allan notes how:

> Often the most difficult edge cases involve language that seems designed to provoke strong feelings, making the discussion even more heated — and a dispassionate look at the context (like country of speaker or audience) more important. Regional and linguistic context is often critical, as is the need to take geopolitical events into account.


Here Allan represents Facebook as an objective, rational social actor in highly-charged contexts, taking considered decisions about the use of language in specific situations.
For instance, he refers to the conflict between Ukraine and Russia, where slang terms traditionally used by each group to designate the other began to be reported as hate speech. Consequently:

We did an internal review and concluded that they were right. We began taking both terms down, a decision that was initially unpopular on both sides because it seemed restrictive, but in the context of the conflict felt important to us.

In evidence here are clear layers of represented reactions: it is first users that take action by reporting perceived hate speech, then Facebook reacts by investigating and subsequently deleting and users react to the decision. The process of enforcement is generalised in that no details are given about what an “internal review” consisted of, or on what basis conclusions were drawn, or how the targeted terms were taken down. Other examples of enforcement given follow a similar pattern of representation, although the scale may differ as in the following example, which takes a global rather than regional perspective. In relation to the migrant crisis in Europe, “feedback” suggesting that posts were “directly threatening refugees or migrants” meant that:

We investigated how this material appeared globally and decided to develop new guidelines to remove calls for violence against migrants or dehumanizing references to them.

Allan also relates enforcement to considerations of the intent of the user, which is assessed from the immediate context. Unlike Twitter, where it is implied that the wider conversational context will be taken into account, Facebook focuses on the specifics of the post:

People’s posts on Facebook exist in the larger context of their social relationships with friends. When a post is flagged for violating our policies on hate speech, we don’t have that context, so we can only judge it based on the specific text or images shared. But the context can indicate a person’s intent, which can come into play when something is reported as hate speech.

Examples of permissible intent given are satirical or humorous use “to make a point about hate speech”, reclaiming offensive terms or repeating hate speech to condemn it. What is not stated is that these examples derive from complaints made against Facebook for removing such content, although Allan does acknowledge that mistakes have been made.

Both Twitter and Facebook represent enforcement as a process beginning with the reporting or flagging of content. In so doing, they include users as social actors in the process of removing hate speech from their platforms. As with Bookmarks, there is emphasis on participation and responsibility. However, as noted by van Leeuwen (2016), social practices require not only actors but resources, and Facebook, in the introduction to its community standards, directs attention to the resources that they have made available for users to take on their responsibility:

We make it easy for people to report potentially breaching content, including Pages, Groups, profiles, individual content and/or comments to us for review. We also give people the option to block, unfollow or hide people and posts, so that they can control their own experience on Facebook.
In addition to reporting others in the hope of getting content taken down, individuals have also the opportunity to remove themselves from the content they do not like. Hence, Facebook is providing alternative resources for users to avoid the action of initiating the enforcement process and possible subsequent deletion of the offending content. With regard to specific resources for reporting hate speech, there appear to be developments underway. Various media sources\(^4\) report that a hate speech button was released accidentally and prematurely by Facebook in the US, meaning that users were asked “does this post contain hate speech?” for even the most innocuous of posts. When and if this particular resource will make a future appearance is not yet known.

The enforcement process, however, also involves those users whose posts have been removed. As noted above, Facebook represents itself as acknowledging mistakes, which are then rectified, so user reactions to actual content removal are part of the represented enforcement process. Indeed, the Verge (a news and media network in the US) reports an interview with Mark Zuckerberg in April 2018, where he talks about the possibility of an appeals process for people who have content removed for violating community standards. Interestingly, in the light of my previous comments about the eligible spaces for hate speech control as represented by Bookmarks, the Verge states that:

The CEO likens the appeal process to Facebook operating more like a government, with the goal of creating a network that “reflects more what people in the community want than what short-term-oriented shareholders might want”.


As in Bookmarks, there is evidence of the blurring between public and private to satisfy the demands for participatory democracy online. If online space is to be public space, it is not just the roles of users that will change, but those who own the spaces may begin to act as public social actors in relation to internet governance.

8.6 Hate speech and the courts: examples of case law in Denmark

One of the criticisms directed against the EU code of conduct discussed above relates to the blurring of public and private as some view the agreement as transferring responsibility for dealing with illegal hate speech to private corporations. The EU, however, insists that it is the courts in member states that arbitrate on matters of criminal hate speech. To illustrate how the judiciary represents illegal hate speech through its interpretations of the legal provisions, I will consider some examples from Denmark.\(^5\) These have been selected to further elaborate on points raised in relation to the Council of Europe and the social media platforms, i.e. publicness, intention and content.

Denmark has ratified all the relevant international and European instruments regarding freedom of expression and restrictions on that expression. The legal provision most generally used to deal with allegations of criminal hate speech is paragraph 266b of the Danish criminal code, popularly known as the “racism paragraph” (racismeparagraph). The first section reads as follows (official translation):

(1) Any person who, publicly or with the intention of wider dissemination, makes a statement or imparts other information by which a group of people are threatened, insulted or degraded on account of their race, colour, national or ethnic origin, religion,
or sexual inclination shall be liable to a fine or to imprisonment for any term not exceeding two years.

The roots of this paragraph come from legislation introduced in 1939 to protect the Jewish community in Denmark and which dealt with the dissemination of false rumours and accusations that victimised or incited to hatred on the grounds of religion, descent or citizenship. The requirement of dissemination remains, but the criterion of falseness has been dropped so the truth of the propositional content of an expression has no bearing on the interpretation of 266b. Unlike Bookmarks, the question “is it true?” is not asked.

I begin with court interpretations of publicness and intention to disseminate in online settings. The Office of Public Prosecutions advises that online versions of media (newspapers, radio and tv), online debate fora, blogs and social media platforms are covered by the provision. In the case of closed, non-public Facebook profiles, an intent to disseminate must be shown through documented evidence of numbers with access to the profile and profile settings. In contrast to Bookmarks, online spaces are not by default viewed as public. A case from the District Court serves to exemplify judicial practice (Judgement 1.4.9 2014). The accused had previously been found guilty by the town court for writing the following comment on a friend’s Facebook profile: “one ought to gas Muslim children” (man burde gasse de muslimske børn). He appealed to the District court who acquitted him on the grounds of lack of evidence of intention to disseminate. The Court argued that writing on a Facebook profile is not equivalent to writing on publicly accessible webpages or blogs and that no evidence had been presented to the town court about the number of friends attached to the profile or the settings of the profiles of the friend and the accused. In addition, the Court remarked that only two people had commented on the article that had given rise to the comment. Moreover, the accused had stated that he had limited knowledge of Facebook and its affordances. In evidence are issues of horizontal scale (spread, distribution), defined by number of friends. This reliance on numbers for the evaluation of intention to disseminate is evident in other judicial judgements. For instance, in a court decision from 2014 (judgement 1.1.43), the town court emphasised that racist comments made by the accused in association with a video he/she uploaded onto a Facebook profile could potentially be shared by his/her 92 to 100 friends. In this instance, the accused was found guilty. In another case from 2009 (judgement 1.2.9), the accused was found guilty on the grounds of the nature of the comments written on her blog, but the court added that, even though it was not her intention that her blog should be read by more than her five to six cycling friends, she “must have realized it was highly probable” (måtte have anset det for overvejende sandsynligt) that others would read it. This is an understanding of intention, which in Danish law is “probability intention” (sandsynlighedsfortsæt). It has similarities to the understanding of intention proposed by Culpeper et al. (2017) as the foreseeableability of the consequences (perlocution) of an utterance.

Turning from issues of intention to content, the question of generalisation arises in several cases. The Office of Public Prosecutions points to dehumanising metaphors and generalisations about criminal behaviours when exemplifying how the wording “threatened, insulted or degraded” is to be interpreted. Importantly, there is no requirement of actual perlocutionary impact on the targeted individual or group; what is decisive is that the “expressions in the concrete situation can be seen as suited to cause fear or appear insulting/degrading”. So, it is the potential of the expression to have a perlocutionary impact in a situated context that is paramount. Note too that speaker’s intention is not represented as a factor; the issue is not whether the speaker intended or did not intend to threaten, insult or degrade. Interpretation
of an expression rests on the evaluations of its potential to have an effect by those eligible to do so outside the communicative event: law enforcement and then, if the police decide to prosecute, the courts. Of course, in actual practice, it is likely that the police talk to the producers of the expressions reported and possibly those who reported them and get an insight into producer and recipient meaning. A court case from 2016 (judgement 1.1.54) serves to illustrate the approach to generalisations about criminal behaviours. The accused had written on a Facebook group profile “No thanks to mosques” (Nej tak til moskéer) the comment “Terror container!!”. The Court found that the post “must be understood as a generalising statement that people who come to a mosque are terrorists” (måtte forstås som en generaliserende udtalelse om, at de personer, der kommer i en moské, er terrorist). For the court, the implied meaning was clear. Since it is Muslims who attend mosques, the statement was a “generalising claim about serious criminality among a group of people defined by their religion” (en generaliserende påstand om grov kriminalitet hos en gruppe af personer, der er defineret af deres tro). The accused was found guilty.

The scope of a generalisation about the target group can be a deciding factor in some court decisions. A case that attracted some media attention in Denmark provides an example. In 2013, the accused, after disagreements on the board of a housing organisation about Christmas arrangements, had written on the Facebook wall of someone else the following comment:

ideologien Islam er fuldt ud lige så afskyvækkende, modhydelig, undertrykkende og menneskefjendsk som Nazismen. Den massive indvandring af Islamister her til Danmark, er det mest ødelæggende det danske samfund har været udsat for i nyere historisk tid.

(The ideology Islam is completely just as odious, disgusting, oppressive and misanthropic as Nazism. The massive migration of Islamists to Denmark is the most destructive thing that Danish society has been subjected to in modern times.)

In 2015 (case number 1-848/2015), the town court found him guilty under paragraph 266b for posting “generalising claims which are insulting and degrading to followers of Islam” (generaliserende påstande, der er forhånende og nedværdigende overfor tilhængere af islam). In interpreting the comment, the court took account of the temporal and societal context and judged that the phrase “the ideology Islam” referred to Islam in general and not to the extreme factions. The accused argued that he was referring to the political side of Islam and that “Islamists” was a normal label for extremist groups. So here we see contestations about situated textual meaning. The case was taken to the District Court through a special appeals process and the judgment of the town court was overturned (judgement 1.4.13). This higher court considered the comments to be directed towards the ideology of Islam and Islamists and argued that the legal provision did not cover groups based on their ideology. This conclusion was reported in the media in terms of being able to criticise Islam without being considered to be racist.

8.7 Discussion: agency, accountability, authority and AI

In considering how hate speech as social practice is represented in the differing texts, certain commonalities emerge and these can be seen as characterising hate speech discourse. It is very clear that the European bodies and social media platforms emphasise participation as part of the online democratic process and, in representing hate speech as a threat to that process and warranting action, they simultaneously emphasise participation as a means to
deal with the threat. Part of the rationale is to facilitate agency and distribute responsibility in a top-down manner so that hate speech can be tackled at grassroots level. As argued by Mitchell and Haugh (2015, p.207) in the context of interactional evaluations of impoliteness:

Agency involves the socially mediated capacity to act that is afforded through 1. knowing one has the ability to act, 2. knowing that these actions may affect others (and self), and 3. knowing that one will thus be held accountable for those actions.

With regard to hate speech, the institutional position is that people have to be equipped with resources to allow them to recognise the relevant behaviours and then act accordingly, be this through counterspeech reporting, or in the case of social media moderators, removing. These resources can be educational, as in greater knowledge and awareness, and technological, as in accessible means to flag and report content. There is an obvious bias towards recipients as social actors and their interpretations and evaluations of what the communicator is doing. Receivers cover a range of what Goffman (1981) refers to as participant roles, that is the roles people can have in an interaction. Reception roles may be those of addressee or bystanders or eavesdroppers in the interaction. The target talked about in an interaction has the role of figure, bearing in mind that a great deal of hate speech is about individuals and groups, but not necessarily addressed directly to them. The importance given to reception roles is unsurprising since hate speech only becomes hate speech through evaluation. In a sense, what is underway is a capacity-building enterprise, involving education in critical digital literacy and metapragmatic awareness, where internet users, depending on one’s viewpoint, are either being empowered to participate in internet governance and effect change or being trained to be normative, digital surveillance police for purposes of control. Whatever one’s stance, part of the process involves assigning users a moral authority and obligation to act. In so doing, a moral order for public discourse on the internet is being constructed, in terms of not only what may be expressed, how and about whom, but also how perceived unacceptable pragmatic behaviours can be challenged and people held to account for those behaviours. Of particular interest is how the notions of public and private are blurring in hate speech discourse as is evidenced in the texts. This means that the spaces for action have expanded as have the obligations and responsibilities of social actors.

This tendency towards broadening is also seen in the representations of hate speech where definitions and exemplifications have become more encompassing in line with countermeasures that have moved beyond the criminal. There are worries that hate speech is being diluted down to become an issue of giving offence with the subsequent dangers to freedom of expression. However, the Danish legal texts still tend to construct hate speech in narrower terms.

More fraught with problems is the process of regulating or enforcing standards online. As noted already, this is very dependent on the meaning-making of recipients, who typically take the first steps by reporting content. Then the authorities, such as platform moderators, the police and the courts, take over. They are faced with issues that are familiar to academic linguists. Where is the meaning to be found: in the text itself, in the assumed intentions of the communicator, in the interpretations of recipients or in a combination of all three? Recipient interpretation would seem to have less of a role in the later stages of enforcement. Facebook and the Danish courts, for example, put more emphasis on context, textual content, communicator intention and potential effects on targets and society at large than recipient meanings.
8.8 Conclusion: conflicts of legitimacy of expression

It is unlikely that hate speech as concept and practice will cease to be discursively contested given its integral links to freedom of expression. However, there are signs of a discourse about hate speech that cuts across genres and social actors to represent common ideas of legitimisation, social (re)actions, spaces and resources for those (re)actions. The question is whether, in the real world, measures to counter hate speech practices can help relieve inter-group tensions and polarised identity politics. This is probably a big ask since underlying prejudices are not easily changed. For example, a study using bots to counter racist slurs on Twitter used by white men found that the racial identity of the bot influenced whether the human subjects altered their behaviour or not (Munger, 2017). An important function of countermeasures is symbolic to indicate the thresholds of acceptability that a society or community wishes to maintain and the types of behaviour that will not be legitimised. These thresholds will be variable, not always consistent and will be sources of conflict in themselves. Constraining hate speech is an exercise in meaning making and metapragmatic evaluation, but what and whose meanings are prioritised in this process will differ according to interests despite institutional attempts to standardise. Whether artificial intelligence can solve the problems and controversies remains to be seen.

Finally, in the spirit of Critical Discourse Analysis, I need to ask if the discourse about hate speech gives any cause for concern. In many ways, such discourse is to be welcomed in that it addresses problems of inequality, prejudice and marginalisation. Nonetheless, from a pragmatics perspective, complex phenomena are perhaps represented as being easier to evaluate than what they are, and the broadening of hate speech, while motivated by interests to encourage wider and fairer participation in internet governance, needs to be applied thoughtfully to prevent over-policing of online spaces that may result in different forms of dominance.

Notes

1 See www.echr.coe.int/Documents/FS_Hate_speech_ENG.pdf.
5 The examples are taken from two documents available on the “Vidensbasen” of the Office of Public Prosecutions. Guidance on how the law is to be interpreted is given in “Hadforbrydelser” (March 2016), available from https://vidensbasen.anklagemyndigheden.dk/api/portals(6e302527-f0b3-4a5e-889a-668aa67e5491)/Print/h/6dfa19d8-18cc-47d6-b4c4-3bd07bce15ec0/VB/a7a856e5-1a4f-
An overview of practice in the courts is given in “Hadforbrydelser”, En praksis oversigt (October 2016), available from https://vidensbasen.anklagemyndigheden.dk/h6dfa19d8-18cc-47d6-b4c4-3bd07bc15ec0/ VB/0adc60f2-504d-4270-bd8c-e41e4f3ae0e0#ID0E OXAE. One example from 2016 can be found at https://vidensbasen.anklagemyndigheden.dk/h6dfa19d8-18cc-47d6-b4c4-3bd07bc15ec0/ VB/d55ead39-87a2-40dd-97e5-287d995d14.

References


Hate speech


