



Together Against Online Hate Speech

TAOH

Together Against Online Hate Speech

Deliverable 2.3 Literature review

From Online Hate Speech to Online Hostility:
Understanding and managing hate speech across
disciplines

About TAOH

The main goal of TAOH - Together Against Online Hate Speech - is to use AI-driven data collection and analysis alongside trainings to strengthen the resilience of civil society, media and authorities in reducing hate speech in the Netherlands, Germany and Spain.

Consortium Members

- SPF - Sharing Perspectives Foundation - Netherlands | Coordinator
- UvA - University of Amsterdam - Netherlands
- Build Up - How to Build Up Germany - Germany
- IBH - ichbinhier e.V. - Germany
- FundiPau - Fundació par la Pau - Spain



**Co-funded by
the European Union**

Funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

From Online Hate Speech to Online Hostility: Understanding and managing hate speech across disciplines

Anna Claire Shepherd, Caitlin Gray & Sander van Haperen¹

Abstract

This document reports the theoretical background informing analysis of online hostility for the project “Together Against Online Hate Speech”. The goal of this report is to provide an overview of how our thinking about “hate speech” has developed over time: how it is defined in different disciplines and how may it be managed? This overview serves to define and conceptualize online hostility for comparison across Germany, the Netherlands and Spain on Instagram, X and TikTok.

Our review highlights the challenges in defining hate speech, while emphasizing a general consensus among various disciplines calling for the protection against hate speech, through a variety of approaches, and a shift towards subjective and contextual understandings of online hostility.

The project “Together Against Online Hate Speech” consists of the following partners: Sharing Perspectives Foundation, Build Up, IchBinHier, FundiPau, and the University of Amsterdam. The project is funded by the European Union (CERV-2024-CHAR-LITI-SPEECH). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Citation:

Shepherd, Anna Claire; Gray, Caitlin; van Haperen, Sander (2026). “From Online Hate Speech to Online Hostility: Understanding and managing hate speech across disciplines”. University of Amsterdam. Preprint. DOI: 10.21942/uva.33375427

¹ Corresponding author (s.p.f.vanhaperen@uva.nl)

Introduction

This literature review describes the development of the concept “hate speech” generally and online hate speech more specifically across varying academic disciplines to provide a theoretical foundation for the project ‘Together Against Online Hate Speech’ to strengthen the capacity of civil society organizations, media outlets, law enforcement, local authorities and the general public to manage hate speech. The following sections will take a closer look at the evolution of hate speech as an academic concept, its use in a legal sense including its tension in this realm with free speech protections, the impact of online spaces on defining hate speech as well as current approaches on how to detect and engage with hate speech.

The following procedure informed selection of relevant literature for review. The aim is to provide a systematic categorization of hate speech across social and political contexts, including legal and policy spaces. Furthermore, while the relevance of the legal status of hate speech cannot be overlooked, our intention is not to define hate speech across jurisdictions. Rather, we aim to develop a working definition that both covers expressions that violate law and/or policy standards and those that disrupt social/cultural norms.

Our search strategy combined broad and subsequently more narrow search terms. Our initial search began with the terms ‘hate speech’, ‘hate speech definition’, ‘cyber-hate’, ‘hate speech detection’, ‘online hate speed detection’, ‘algorithmic HSD’, ‘AI use in HSD’. We then purposively selected relevant legal and policy definitions for hate speech in the Netherlands, Germany and Spain (the contexts relevant to the project partners). Sources were included if they either explicitly defined or conceptualized hate speech, proposed or applied definitional frameworks within the academic, legal, policy or computational science domain, or if operational or analytical data and/or methods were presented.

Hate speech across disciplines

Law

Within legal scholarship and practice hate speech has developed as one of the most complex concepts to define and as such enforce. As compared to other theoretical discussions around hate

speech, the legal field emphasizes precise definitions as a means to determine which kinds of expressions can be protected and which expressions must be reprimanded. Consequently, legal debates surrounding hate speech evolved through ongoing tensions between freedom of expression, equality, and protections.

The emergence of hate speech as a legal concept came into effect around the 1960s as a mark of both local governments and supranational/international institutions' response to the second World War. Although not ratified into law, the United Nations' 1966 International Covenant on Civil and Political Rights, more specifically article 20 (2) addresses hate speech very generally. The provision states "Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law" (UN 1966). Similarly, the United Kingdom was one of the first countries to make explicit in their constitutional framework, the regulation of hate speech. The 1965 'Race Relations Act', more specifically the provision titled Public Order states: "A person shall be guilty of an offense under this section 'Incitement to' if, with intent to stir up hatred against any section of the public in racial hatred", distinguished by colour, race, ethnic or national origins. Many nations followed suit within the following decades. The most fundamental challenge within the legal sphere as it stands now relates to the tension it poses to other protected freedoms, such as freedom of speech in the United States which makes direct legal protections against hate speech more ambiguous, and discussions of censorship which have become more prominent in relation to online spaces and platforms.

Historically, Mari Matsuda's Public Response to Racist Speech (1989) was pivotal as Matsuda aimed to shift public thinking about hate speech as a merely offensive expression but as a systemic form of discrimination and social subordination. Matsuda notes that a legal response to hate speech would solidify the equality that is the responsibility for the state to uphold; "This absence of law is itself another story with a message, perhaps unintended, about the relative value of different human lives... a legal response to racist speech is a statement that victims of racism are valued members of our polity" (Matsuda 2322). One of the most cited works on hate speech more generally is the book "The Content and Context of Hate Speech: Rethinking Regulation and Responses" by many authors and edited by Michael Herz and Peter Molnar

(2012). Within the collection author Bhikhu Parekh in his chapter ‘Is There a Case for Banning Hate Speech?’ presents an argument that attempts to define hate speech and justifies the need for regulation, outside of the boundaries of constitutional doctrine. Parekh asserts most legal definitions are not targeting hate speech, but lump together actual occurrences of hate speech with other expressions albeit harmful, offensive or even demeaning (40). In such, the blanket term hate speech is unsatisfactory to Parekh: “Hate speech is a distinct kind of speech, and much conceptual confusion is created – and the net of prohibited speech unduly widened – by subsuming all forms of uncivil and hurtful speech under it” (ibid). In order to distinguish hate speech from other forms of harmful language Parekh suggests hate speech is composed of three essential components. Firstly, “it is directed against a specified or easily identifiable individual or, more commonly, a group of individuals based on an arbitrary and normatively irrelevant feature (ibid). Secondly, “hate speech stigmatizes the target group by implicitly or explicitly ascribing it qualities widely regarded as highly undesirable” he notes stigmatization is a matter of degree but, if the quality in discussion is inherent to the groups’ identity, it is permanently condemned (41). And finally, “because of its negative qualities, the target group is viewed as an undesirable presence and a legitimate object of hostility... because society would be better off without it, it may legitimately exterminate or expel the target group... thus hate speech encourages and purports to justify discrimination (ibid). To justify these boundaries, Parekh suggests due to varied reactions to hate speech both within the victims and the wider public, the most relevant element within the hate speech debate is the content of the expression of hate speech, and the long term effects on the group and their social standing, transitioning away from the ‘immediate danger/effect’ that is sometimes integrated into legal proceedings around hate speech. Additionally, this work suggests that creating hate speech laws do not only serve as punitive measures but fulfill many roles including educating and norm-setting as these concrete measures unequivocally assert equal standing for all members of society. Finally, Parekh suggests that the discrepancy among states can be reasonably explained, and even justified in the case of hate speech as local and historical contexts should be considered by governments in their determination of what precautions/punishments are necessary in their attempt to rectify the effects of hate speech.

A completely different approach was adopted by author Michel Rosenfield in his chapter ‘Constitutional Jurisprudence: A Comparative Analysis. Rosenfield embarks on a comparative analysis of the myriad ways Western democracies have up until ‘this point’ [2012] chosen to ratify hate speech laws and codes within their Constitutions, explicitly making note of the obvious departure the United States regulates, or rather, does not regulate hate speech. Although Rosenfield recognizes the merit to understanding hate speech laws through a contextual lens, even noting it might be ‘tempting’ he also states “given the trend toward globalization and the instant transactional reach of the Internet, a purely contextual approach would seem insufficient, if not downright inadequate” (243), as such a comparative analysis that sheds light on the differences in regulating hate speech can likewise set the basis for possible ways to “explore how best to deal with hate speech as a problem for contemporary constitutional jurisprudence” (ibid). In section II and III, Rosenfield compares the United States to other ‘Western Democracies’. The US essentially protects deeply offensive and even hateful expression, only acting on hate speech if it directly incites ‘lawless action’ or constitutes a narrowly defined exception such a ‘true threat’, which reflects a wider culture of distrust of state power and understanding of free speech as essential to democratic liberty. However, as largely explored by other authors, this strategy becomes inadequate when other protections such as intimidation, social exclusion and the reinforcement of social inequality are also paramount to establishing democratic liberty. In line with this viewpoint many Western democracies such as the explicitly named Canada, Germany, and the United Kingdom, have adopted hate speech laws that restrict speech that threatens social cohesion, born out of a history of fascism, there is a general trend that democracy must defend itself against anti-democratic expressions (that lead to actions). The author notes challenges to both approaches; as not all hate speech is equal and its consequences are highly varied (281). The chapter concludes by suggesting “ a connection with regulation of hate speech with an approach of fundamental normative principles, and in light of key contextual variables [...] these principles (openness to pluralism and respect for the most elementary degree of autonomy, equality, dignity and reciprocity) must transcend geographical, cultural and historical differences, and at the same time be open to accommodate relevant historical and cultural variables (history and nature of discrimination, status as a minority or majority group, customs, common linguistic practices and the relative power or powerlessness of speakers and their targets within the society involved (287-288).

Meanwhile, at the level of the United Nations, it seeks to provide “a unified framework for the United Nations to address the issue globally”. UN Strategy and Plan of Action on Hate Speech defines hate speech as “any kind of communication in speech, writing, or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor” (United Nations, 2026). However, the UN definition is immediately followed with “to date there is no universal definition of hate speech under international human rights law. The concept is still under discussion, especially in relation to freedom of opinion and expression, non-discrimination and equality (ibid). As such, the greatest challenge to a legal framework, comes from balancing the protection of freedom with the need to prevent discrimination, exclusion and harm particularly as hate speech has continued to evolve (legally) from explicit incitement to more subtle and covert expressions whose meaning may vary across cultural, political, national contexts.

Further expanding the boundaries of what defines hate speech in a legal sense, Waldron’s “The Harm in Hate Speech” (2012) is one of the leading contemporary works positing hate speech should not primarily be understood as solely offensive, or even, as a harmful expression, but rather as communication that actively undermines the dignity, social standing and inclusion of vulnerable minorities (Waldron 2012). More explicitly, Waldron attempts to distinguish hate speech from offensive speech by establishing the concept of “implicit assurance”, which is the notion that regardless of one’s beliefs/allegiances (essentially the facets of identity that are targeted by expressions of hate speech), they will be treated with equal right to membership in a society” (Waldron in Fish 2012). It is acts of speech that threaten this assurance that should be monitored and regulated as it is in the state’s realm of responsibility to protect the dignity lost by victims of hate speech. Waldron also offers insight into the debate regarding possible infringement on (Constitutional) protections of free speech noting, “hate speech and defamation are actions performed in public, with a public orientation, aimed at undermining public goods.” (Waldron 2012: 102). Waldron’s work moves the legal debate away from defining hate speech on the narrow basis of offense or incitement but towards concerns about the responsibility of the law

to a commitment of equality, democratic inclusion and the social conditions necessary for full citizenship.

To summarize, legal scholarship treats hate speech as a contested but necessary category for determining when expression loses protection because it threatens equality and social inclusion, rather than merely causing offence. Emerging after the Second World War through instruments such as Article 20(2) of the International Covenant on Civil and Political Rights and early national laws including the UK's Race Relations Act, legal approaches have developed amid persistent tension between free expression and protection from discrimination. Matsuda frames racist speech as systemic subordination requiring a legal response that affirms the equal value of targeted groups, while Parekh argues that hate speech should be distinguished from broadly offensive or uncivil speech through three elements: targeting an identifiable group on arbitrary grounds, stigmatising it through negative attributes, and representing it as a legitimate object of hostility, exclusion, or discrimination. Waldron similarly shifts attention away from offence and immediate incitement towards speech that undermines minorities' public dignity and their assurance of equal membership in society. Although the UN currently defines hate speech broadly as identity-based communication that attacks or uses pejorative or discriminatory language, it acknowledges that no universal definition exists in international human-rights law; consequently, legal regulation must balance shared principles of autonomy, equality, dignity, reciprocity, and pluralism with the cultural, historical, and power-related contexts in which both speech and its harms occur.

Social Sciences

Generally, the social sciences have moved our understanding of hate speech beyond strictly legal definitions. We find two strands of literature on hate speech in the social sciences: one focusing on the characteristics of targets, the other considering the context of such language. Generally, hate speech is understood as hostile or disparaging language targeting individuals or groups because of characteristics such as race, religion, gender, disability, or sexual orientation. At the same time, it is widely acknowledged across the social science literatures that the concept

remains difficult to define and operationalise consistently, which led to consideration of the context within which hate speech becomes meaningful.

Initially, consideration of hate speech arose as a concept in relation to free speech. In tracing the evolution of hate speech literature over the three decades, Tontodiamamma et al (2020) note a shift from general themes around the nature of speech and free speech versus hate speech to research focused on the technical aspects of hate speech detection and methodologies, especially within the field of machine learning (Tontodiamamma et al, 2020:168). In their analysis they also expand on the countries that are the most influential in hate speech research, finding that especially in Anglo-Saxon states focus on the nuances between hate speech and freedom of expression, particularly in the United States (ibid, 172). In contrast to other European countries, which prohibit speech and the expression of people who deny crimes of genocide e.g in Germany or Italy where the display of ideas and symbols of facism was banned under the Mancino law (Vergani, 2024:4).

The rich tradition in US scholarship emerged from critical race theorists coining hate speech as a term early on. In the 1980s, Mari Matsuda used the term hate speech to “highlight how the United States failed victims of racist hate speech due to inadequate means of seeking redress, civil or criminal” (Brown, 2017:424). In this same tradition stands one of the most frequently cited articles on hate speech, John Nockleby’s entry in the Encyclopedia of the American Constitution on Hate Speech, defining hate speech as encompassing “communications of animosity or disparagement of an individual or group on account of a group characteristic such as race, colour, national origin, sex, disability, religion or sexual orientation” (Nockleby, 200: 1277). While this definition is useful based on the broad spectrum of identity characteristics that is covered by this definition and why it has been cited so widely in the literature, this broad and abstract definition presents difficulties in implementation.

In this same vein, others have focused on characteristics targeted by hate speech. For instance, Alexandra Siegel notes “hate speech is understood to be bias-motivated, hostile, and malicious language targeted at a person or group because of their actual or perceived innate characteristics” (Siegel, 2020:57). Giron et al (2024) argue that hate speech is a composite concept defined

through a set of recurring conditions that appear across legal standards, platform policies, and normative frameworks (Giron, 2024:5). Instead of adopting a single operational definition of hate speech, they instead advocate for consideration of core conditions to adopt when identifying hate speech. They argue such a checklist should include a Protected Target Identification, whether the text targets an individual or group on the basis of an inherent protected characteristic is foundational, as without a protected target the expression may be abusive or offensive but it does not meet the basic definition of hate speech (ibid, 5). The second condition concerns hate speech as requiring content to convey hostility, contempt or animus toward that group, and operationalises this by checking for a derogatory tone, insults, accusations or negative generalisations tied to the protected identity (ibid). They summarise that hate speech encompasses dehumanisation, scapegoating narratives, calls for exclusion or discrimination, threats and explicit or implicit incitement to violence (ibid). Relatedly, the use of implicit hate speech, defined by El Sherief (2021) as “coded or indirect language that disparages a person or group on the basis of protected characteristics like race, gender and cultural identity” (El Sherief, 2021:1).

By contrast to a focus on hate speech as specific language targeting characteristics, other scholars have emphasized the context within such language is used and attempts to determine which commonalities are shared across various definitions. Sellars’s seminal work argues that despite the extensive literature on the causes, harms and responses to hate speech, there is a paucity of articles that systematically define the term (Sellars, 2016:24). While Sellars does not provide a definition of hate speech he suggests common themes for any definition including 1) targeting a group or individual as a member of a group 2) the content in the message that expresses hatred, 3) the speech that causes harm 4) the speaker intending harm or bad activity 5) the speech incites bad action beyond the speech itself 6) the speech is either public or directed at a member of the group 7) the context makes violent response possible 8) the speech has no redeeming purpose (ibid, 14).

Advancing on Sellars (2016) work, Hietanan and Eddebo go further in tracing the definition of hate speech by examining definitional modes ranging from A) teleological definitions where the intentionality and tendency of the speech act towards certain ends is focused on B)

consequentialist definitions which focuses on the effect of the speech act alone, C) formal definitions which builds on the essential character of the speech act and ideas involved and D) consensus definitions where any speech act can become denoted by hate speech by fiat (ibid, 454). They argue for a multi-pronged definition combining elements of each definitional mode, and argue this can create a workable foundation for nuanced utilisation of automated hate speech suppression (ibid, 455), while being aware of the danger of value-laden definitions and situations where sensitive or inflammatory political criticism could be construed as hate speech and become a target for suppression (Hietanen and Eddebo, 2023: 444). In doing so, they highlight where definitions fall short, where they can be misconstrued for political gain, and where they can be improved, and how the different definitional modes create different unintended outcomes.

The consideration of context has expanded to online spheres in recent years. As Kentmen-Cim argues (2025) political science scholarship has suggested that platforms' flexible and shifting definitions of harm, violence and hate speech are shaped by shifting political climates and existing power hierarchies. They argue that this means that greater protection is offered to socially and politically dominant groups thereby leaving marginalized communities exposed to abuse and ingrained groups which are deemed worthy of protection. (Kentmen-Cim, 2025:4). Equally, as they note the rules and definitions which govern hate speech are subject to frequent and often undocumented, changes over time (ibid: 5). As Kentmen-Cim notes a common theme in the literature is now how social media companies are leading and developing their own understandings of hate speech, particularly as already noted by Sellars in 2016, platforms benefit from a vague definition of hate speech as it allows them to respond at will. This was also noted by Hietanen and Eddebo (2022) where a definition or model of hate speech must mesh with their overarching goals relating to profit and expansion (Hietanen and Eddebo, 2023:454).

In summary, social-scientific research broadly understands hate speech as hostile language directed at individuals or groups because of protected or perceived identity characteristics, but scholars differ over whether these target-based features suffice to cover the concept. Emerging from debates over free expression (particularly pronounced in US scholarship, where critical race theorists such as Matsuda foregrounded failures to protect victims, in contrast to more restrictive European legal approaches) Nockleby, Siegel, and Giron et al. emphasize the protection of

targets. Giron et al. develop this approach into a set of recurring conditions, including identity-based targeting, animus, and expressions such as dehumanisation, exclusion, threats, or incitement, while El Sherief extends it to coded and indirect forms of hostility. In contrast, others argue that no stable definition can be derived from target and content alone, identifying additional recurrent contextual features such as harm, speaker intent, publicity, incitement, and possibly violent reception. Hietanen and Eddebo similarly reject a single definitional formula, proposing an approach that combines intent, consequences, formal properties, and social consensus while warning that overly value-laden definitions can enable the suppression of legitimate political criticism. This contextual concern becomes especially important online: Tontodiamamma et al. identify a shift from normative free-speech debates towards automated detection, whereas Kentmen-Cim highlights how platforms' mutable definitions of hate, harm, and violence are shaped by political power, commercial incentives, and changing institutional priorities, potentially offering uneven protection to dominant and marginalised groups.

Philosophy

Sharing similar roots as the social science literature, philosophical debates about hate speech have also traditionally centered on freedom of expression in contrast to harm, but have tended to emphasize more the nature of the individual in relation to others, for instance through an understanding of “justice”.

For instance, Jeremy Waldron's (2012) book 'The Harm in Hate Speech' contributed to the extant literature on the free speech and hate speech debate. Waldron argued that public hate speech should be seen as a group libel, an attack on the dignity of individuals through their group membership. He argues that hate speech legislation is necessary for the assurance of justice and that hate speech pollutes the social environment. At the same time, regulating and monitoring hate speech is challenging. Babak Bahador (2023), explores how the current definitions of hate speech can influence and restrict hate speech monitoring. He defines hate speech as 'communication aimed at groups of people to dehumanize, demonize or incite violence against them. If hate against a particular outgroup (the group targeted for hate speech) is successfully sold to an in-group (the group hate speech attempts to persuade) then all members of the

out-group are viewed as a negative stereotype, losing their individuality and humanity' (Bahador, 2023:292). Bahador questions the utility of current definitions of hate speech noting that most definitions limit targets to groups that hold immutable qualities e.g. race, gender, nationality, however, he noted that there is a growing number of groups which are not traditionally covered by hate speech definitions, who are now being targeted e.g. journalists (ibid). It is thus problematic to restrict hate speech definitions to only the most severe types, dehumanisation, demonization and incitement, if other speech also contributes to hate and disdain towards groups (ibid, 294).

Similarly, Alexander Brown (2017) critically interrogates the concept of hate speech itself, arguing it is a myth that emotions, feelings, or attitudes of hate or hatred are part of the essential nature of hate speech, suggesting instead that hate speech expresses a family-resemblance concept, and therefore we cannot identify one essential feature of hate speech. Brown argues that different people have different sensibilities and conceptions of what is perceived as hateful. By this assumption, regulating speech on this basis is inconsistent with free speech doctrines.

In consideration of context, Dirk Kindermann (2023) argues that the term hate speech should be dispensed of completely and replaced with the term discriminatory speech. Kindermann identifies the paradigmatic definition of hate speech as having two features in common. First, hate is ascribed and located in the motive, content or effect of the relevant speech indirectly, it is thus ascribed to the speaker (expression, motivation) or their intended audience (incitement). Second, hatred is claimed to be directed at (members of) some social group (Kindermann, 2023: 815). Kindermann argues that the definition of hate speech falls short as it rarely elucidates how it is dependent on structural power relations, as nothing in the definition prevents this application as Kindermann gives the example of the term 'Whitey', suggesting this should not be counted as racist hate speech (ibid, 823). He argues that the term discriminatory speech could replace hate speech, defining it as "speech that disadvantages, or contributes to the disadvantaging of people, on account of their positioning in social categories such as gender, race, class [...], where this disadvantaging is understood in terms of discrimination" (ibid, 824). Kindermann's argument therefore interrogates how current definitions of hate speech suffer from conceptual slippage, failing to include reference to structural power dynamics.

In summary, philosophical debates on hate speech share social-scientific concerns with the tension between freedom of expression and harm, but place greater emphasis on dignity, justice, and moral responsibility. Waldron conceptualises public hate speech as group libel undermining the dignity and equal social standing of group members, arguing that legal regulation is needed to preserve a just and inclusive public environment. Bahador similarly stresses hate speech's capacity to dehumanise out-groups and mobilise in-groups against them, but criticises narrow definitions restricting protection to characteristics or only the most extreme forms of hate speech such as demonisation and incitement. Brown by contrast challenges assumptions that hatred is an essential feature of the concept, treating hate speech as a family-resemblance category with boundaries depending on differing social sensibilities, complicating regulation under free-speech principles. Kindermann takes this critique further by rejecting the term hate speech in favour of *discriminatory speech*, arguing that conventional definitions insufficiently account for structural power and can misleadingly treat speech against socially dominant groups as equivalent to racism or other systemic forms of oppression. Despite their differences, these authors converge in questioning narrow, purely linguistic definitions: Waldron foregrounds dignitary harm, Bahador the broader production of group-based disdain, Brown conceptual pluralism, and Kindermann the unequal power relations through which speech produces discrimination.

Communication Sciences

Communication science is the field with some of the highest volume of papers within the topic as well as, some of the most widely cited papers on the topic of HS (Paz et al., 2020: 4). By contrast to the social sciences generally, communication science research has increasingly come to emphasize hate speech not simply as offensive language but as a communicative process involving speakers, audiences, messages, platforms, interpretations, and social effects that can reproduce fear, exclusion, discrimination, and unequal group hierarchies. Understanding hate speech involves “emitters, receivers, messages, channels and interactions, not to mention its effects and interpretations, which fuel fear, intimidation, harassment, abuse, and discrimination... establishing social hierarchies based on inequality and dominion” (Paz et al,

2020: 2). The field has burgeoned especially given the advent of online hate speech, giving rise to concerns about hate speech in relation to social media. Consequently, much attention in the discipline is given to the interaction between content and medium.

Instrumental in the move of studying hate speech beyond the legal field by way of communication theory is the work of Clay Clavert (1997), providing “courts and legislative bodies with distinct frameworks for analyzing harm caused by hate speech” (1997:12). The respective theories in question are the ‘transmission model of communication’ which makes explicit the direct harms caused through the use of hate speech in which immediate effects/behavioral changes within the targets are noted (1997:6). In contrast, the ritual model of communication determines the long-term and cumulative harms produced through the continued use of hateful speech. This model demonstrates how language propagates and sustains a ‘social reality of racism that promotes disparate treatment of minorities’. The use of racist speech creates both convenient categorization of class and attribute all negative attitudes imputed to the class (Delgado 1995: 163) and slurs stand as conventional ways of treating people as subordinates (Altman 1995: 127). The tandem use of these communication models contributes to a more well-rounded understanding of the impact hate speech has on its victims’. This work emphasized the importance of recognizing how protections against hate speech must go beyond the legal realm as a ritualistic use of hate speech reproduces and sustains social inequalities and hierarchies. Clavert’s paper justifies the need to apply communication sciences when defining the scope and impact of hate speech specifically by examining the weaknesses of its legal application, stressing the complex nature of the topic.

The discursive nature of hate speech is further explored through the book *The Communication of Hate* (2011), wherein Wallman and Haas demonstrate the connection between communication and hate (and as such, hate speech). While hate speech might serve a number of goals, what is clear is that “hate comes alive in our language and our actions” (2011: 3). As such, hate speech is specific discourse founded on four main components (which are not necessarily present simultaneously for speech to be hateful). Firstly, hate speech is “discourse that seeks to politicize social differences (often in a non-neutral and persuasive manner)” the positive construction of in and out group through hateful discourse “are not simply cognitive processes but are labels put

into practice through social discourse (2011:4). Secondly, hate speech is characterized by attempts to “construct the out-group in negative and dehumanized terms, often while praising the in-group and constructing them in positive terms (2011:35). The third component of their definition centers on the context in which hate speech is produced noting “hate speech is most identifiable as such when the objects of hatred belong to a historically oppressed social group [...] hate speech capitalizes on these out-group experiences and attempts to marginalize and strike fear” (2011:36). Lastly the use of hate speech to ‘conquer’ the out-group in such using speech to call for the destruction or glorified killing of outgroups). Together, these features help us to understand how hate speech is constructed discursively.

Others have advanced this discursive understanding of hate speech through distinguishing a variety of different types of related language, and how hate speech has developed linguistically through the forms in which it is communicated. For instance, Rossini (2020) contributes to delineating hate speech as compared to other ‘bad’ speech. She notes that in the past most scholars (communication and others) have treated ‘all’ negative rhetoric online as harmful, toxic, or even outright hateful but warns that by blanketing all negative discourse as hate speech will impact the ability to challenge and/or moderate ‘true’ instances of hate speech. As such she distinguishes between incivility and intolerance, two distinct communicative phenomena and of which, intolerance is the only of the two that poses a threat to democratic pluralism/equality (2020: 400). Drawing on political communications theory, Rossini notes that uncivil discourse “in which people express themselves with foul language or harsh tone, is a rhetorical act that can serve different strategic goals in a political discussion... and is not in itself, incompatible with political discussion” (ibid). Through an explorative content analysis with a contingent focus on platform affordances, Rossini was able to show “uncivil and intolerant behavior occur in substantially different conversational contexts and are associated with distinct features of discourse” (2020: 414). The main differences noted by Rossini are where intolerant discussions take place and their resilience to disagreement (2020: 416). This contribution is important when better understanding how to approach defining and moderating hate speech as it emphasizes decoupling expressions of incivility which refer to tone from intolerance which refers to content, which, within the current context, is paramount as growing concerns regarding content moderation continue to mount.

Barth et al (2023) contribute a major reconceptualization of hate speech as communication and what consequences that has for defining the concept. He contrasts an intentionalist perspective which considers the motivations to generate hate with a consequentialist perspective which refocuses on victims and consequences of hate speech. This suggests a more comprehensive understanding of the complex nature of hate speech as hateful communication, to take up “a systems-theoretical perspective on online hate speech is not interested in singular speech acts, motives, intentions, or experiences of hate speech but in communicative processes. This perspective does not address the question of what online hate speech is but how communicative acts can be understood as hate communication within a social system” (2023: 228). The authors note that as social media has transformed communication making it decentralized and an environment in which antagonism, incivility and polarization are welcomed. Rather than seeking to establish a single authoritative definition of hate speech, they emphasize communicative contexts, reflecting a broader shift in the field away from attempts to analyze hate speech as isolated expressions.

In considering such context, the online sphere has led to marked shifts in our understanding of hate speech. As noted above, research on hate speech in the field of communication science has been marked by the influence of the internet generally and social media more specifically on the proliferation of hate speech. Parvaresh (2023) notes that hate speech, especially as it is engaged with online, can no longer be understood through the traditional approaches that mark hate speech as use of explicit slurs, offensive lexicon or overall clearly hateful words [...] hate speech now is often covert, contextual and discursive marketing, a pragmatic turn within this area of communication sciences. Authors note that now hate speech is expressed in myriad ways such as through humour, insinuation, irony or accompanied with self-deprecation, all of which “provide so-called haters with a unique opportunity to express hate whilst evading detection” (2023:75). This emphasizes the need to move beyond primarily understanding hate speech lexically but also pragmatically and discursively as well.

For instance, Kim et al explore the ‘prism of social media’ by characterizing who comments on social media and factors that contribute to rhetoric being described as uncivil or toxic (937). The article conclude three main findings; frequent commenters on social media are already more polarized (they have self-reported higher rates of political interest and/or are more ideologically more extreme as such, they have more polarized attitudes), online comments (in this case Facebooks’) are substantially more toxic than ‘real life’ (represented by YouGov), and platform affordances lead to the social spread of toxicity, both as comments deemed toxic received more engagement and exposure to toxic comments caused participants to engage in more toxic responses (932). This paper demonstrates that hate speech is amplified by both platforms and toxic and hateful speech have created communicative norm adoption online, and through the algorithmic affordances and lack of moderation, hate speech has become ubiquitous online. Similarly, Clausen et al. (2026) explore the specifics of online infrastructure in the spread of hate speech (with a more explicit focus on hate speech against sexual and gender minorities). These authors offer that social media platforms are not neutral communication spaces and multiple affordances produced by each social media outlet such as anonymity, visibility, collectivity and interactivity facilitate the production and reproduction of hate speech. This emphasises the notion that the source of hate speech does not only exist within individuals/groups but, it is facilitated by communication infrastructures and the technologies built-into these platforms actively influence social conflict, exclusion and democratic participation.

To summarize, communication science research increasingly treats hate speech not as isolated offensive language but as a communicative process shaped by speakers, audiences, messages, interpretations, platforms, and social consequences. Calvert’s distinction between transmission and ritual models captures both immediate, target-level harms and the harm through which repeated racist language normalises social subordination, drawing on Delgado’s account of racial categorisation and Altman’s understanding of slurs as conventional acts of subordinating people. Wallman and Haas similarly understand hate speech as a discursive practice that politicises social differences, constructs in-groups against dehumanised out-groups, draws on historically unequal relations, and can lead to violence against the out-group. While such accounts emphasize the social and structural functions of hate speech, Rossini cautions against equating all online

hostility with hate speech, distinguishing *incivility* as a harsh tone that may remain compatible with democratic debate, from *intolerance* which concerns exclusionary content and threatens equality and pluralism. Barth et al. extend this contextual orientation by rejecting singular, intention- or consequence-based definitions in favour of a systems-theoretical account of *hate communication*, focused on the processes through which communication becomes socially recognisable as hateful. Research on online communication further shows why this approach is necessary: Parvaresh demonstrates that hate increasingly operates through indirect and pragmatic forms such as irony, humour, insinuation, and self-deprecation, while Kim et al. and Clausen et al. show that platform affordances including engagement incentives, exposure dynamics, anonymity, visibility, collectivity, and interactivity can amplify toxicity, normalize hostile behavior, and facilitate the production and circulation of hate speech.

Computer Science

Generally, computer science research has come to frame online hate speech moderation as a socio-technical challenge rather than a simple task of identifying offensive words, because automated systems must operate within contested legal, cultural, linguistic, and platform-specific definitions of hate speech. Methodologically, the advent of large language models have moved us beyond simple lexical models and provide new means for understanding hate speech in context.

The legal, philosophical, social and communicative impacts, as explored above, have all provided impetus for governments and platforms to regulate and moderate hate speech. From a technical perspective, computer science scholars have made many contributions to the study of hate speech by demonstrating the challenges and advantages of different approaches in regulating hate speech. Reflective of the trends in other disciplines, computer science has attempted to enable moderation that goes beyond merely flagging common slurs to comprehensive models that account for hate speech and all its complexities in context.

Mirroring the development of legal definitions of hate speech, in the field of computer science, Govers et al (2023) assert the need to define clear hate speech restrictions to protect expression rights and victim groups rights and safety (Govers, 2023:4). Equally, due to human biases and

the risk of mainstreaming intolerance, they note that obtaining viewpoint neutrality to categorise hate speech is challenging (ibid). As a result of this they define hate speech as “the subjective and derogatory speech towards protected characteristics expressed directly or indirectly to such groups in textual form” (2023: 12). This subjective nature of hate speech makes it difficult to detect programmatically.

Warner and Hirschberg (2012) presented one of the earliest approaches to detecting hate speech on the world wide web. They use an approach aimed at detecting key words present in hate speech, however, they note the issues inherent within defining what hate speech is e.g., mentioning or praising an organisation associated with hate crimes does not constitute as hate speech, and disparaging terms and racial epithets may be used by speakers who belong to those groups (Warner and Hirschberg, 2012:20). Additionally, they received data from Yahoo comments which had been flagged as offensive by other readers, finding purged comments which intentionally misspelled words to evade filters, which gave them useful examples of techniques employed to bypass lexical filters e.g. utilising expanded spelling was employed so “jew” became “j@e@w@” (ibid, 21). From this basis, they posit that hate speech employs well-known stereotypes to convey hateful intent, and began to create a language model for each stereotype, in order to build a model for all hate speech. It was one of the first papers to formalise hate speech detection as a natural language processing/classification task, describing a methodology for the early wave of hate speech detection, focused on surface-level indicators, that later scholars such as Fortuna and Nunes expand on.

Grappling with automated content analysis, one of the most influential studies around hate speech in the field of Computer Science is that of Fortuna and Nunes (2018). The authors start by noting that within computer science the study of hate speech is recent and unstable mainly as a result of friction between attempts to operationalize hate speech as a machine-learning classification system rather than through a universally accepted social definition, which as we have explored above does not exist. The article explores the results of a survey in which they map the trends found in research in the field as an attempt to help better detection and prevention practices (2018:1). The paper first maps out different definitions set not only by governmental bodies but also by platforms, highlighting the unique challenge faced by computer science as

developments in this realm now have additional, and highly relevant/powerful stakeholders they must also consider. After this comparison the authors suggest, on the basis of the definitions observed, their own definition of hate speech: “Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used.” (2018:5). The next most substantial finding is a comprehensive overview of what has so far been done in hate speech detection. Their findings suggest that early attempts to text-mine and moderate hate speech focused on surface-level indicators such as slurs, profanity and key-word frequency (2018:85-86) whereas, as time has gone on, now contextual features covering ‘othering’ language, stereotypes, and user metadata have been incorporated (2018:18). This text is foundational as it demonstrates a parallel shift in computer science, similar to the other disciplines discussed above, from a purely lexical detection issue toward a more socio-political-technical understanding of the impact online hate speech can have underscoring the need for moderation to continue to evolve.

A recent shift within the literature on hate speech has focused on the growth of automated moderation, where content can be taken down before it has been flagged. This growth of automated moderation has created a renewed interest in the use of dog-whistles and has Bhat and O’Klein have noted after the introduction of Google’s perspective tool, which was aimed at reducing toxicity and detecting insults, dog-whistles became a popular online strategy (Bhat & O’Klein, 2020:178). They argue that platforms struggle to enforce hate speech, when it is often changing, hidden and covert. Mullah and Zainon (2021) reflect on earlier works by reflecting more narrowly on specific techniques used in moderation, more specifically, a shift from simple text classification toward more sophisticated machine-learning, ensemble and deep-learning models that attempt to capture more semantic, contextual and culturally relevant expressions of hate speech (2021:88369). Although the authors note a need to continue developing these more complex techniques, specifically a lack in use of ensemble and deep-learning models, the underlying issues of annotation (mis)alignment, cultural variation, multi-linguality, imbalance/access to data covert expressions of hate speech have still not been addressed through technological advancements (2021: 88371). This study has been cardinal in providing an

overview of the landscape of techniques in computer science and the development of more advanced models. Recently scholarship has turned to artificial intelligence as a means to grapple with contextual meaning for purposes of detection and moderation of hate speech. Qureshi et al. (2026) firstly note that Explainable AI can act as a bridge between technical and social concerns/challenges previously faced in developing detection models: “In hate speech detection, XAI methods can demystify model predictions for content moderators, platform developers, users, and policymakers” (2026:2). However, argue that hate-speech moderation, with the use of AI, now has evolved from a technical classification task to a wider sociotechnological challenge with concerns around transparency, fairness and accountability. Specifically, XAI tools have created “technical advances, such as transformer-based models and counterfactual reasoning, have improved performance, many of these models remain opaque, lexically biased, and poorly adapted to multilingual, dialectal, or socio-politically embedded content. Likewise, many XAI methods optimize for technical faithfulness but neglect user trust, narrative coherence, or fairness in explanatory communication” (2026:28). The authors note that the use of AI can aid in creating much more effective models but stress “the need to design systems that are not only accurate and scalable but also context-sensitive, culturally aware, and aligned with the expectations of diverse stakeholders, including content moderators and impacted communities” (ibid). Qureshi et al. note technological advancements, especially in the area of hate speech detection, must require an interdisciplinary approach to maintain integrity the complex legal and socio-political context in which ‘hate’ is embedded, as such, these tools are only useful if they address the ‘true’ nature of the problem (embedded in context).

To summarize, computer science research increasingly frames online hate speech moderation as a socio-technical rather than purely classificatory problem, recognising that automated systems must translate contested legal, social, cultural, linguistic, and platform-specific understandings of hate into operational categories. Fortuna and Nunes describe hate speech as language that attacks, diminishes, or incites hatred or violence against groups defined by protected characteristics, including through subtle or humorous forms, and show how detection has moved from keyword- and slur-based approaches such as Warner and Hirschberg’s towards models that incorporate context, stereotypes, othering, and user-level information. Govers et al. likewise emphasise the need for clear but viewpoint-neutral restrictions that protect both expression and

targeted groups, defining hate speech as subjective and derogatory speech directed, explicitly or implicitly, at protected groups. However, Bhat and O’Klein highlight how moderation systems encourage adaptive practices such as dog whistles, while Mullah and Zainon show that more sophisticated machine-learning, ensemble, and deep-learning approaches still struggle with biased annotations, cultural variation, multilingual content, uneven datasets, and covert expressions. Recent work on explainable AI, particularly by Qureshi et al., argues that transparency can help moderators, developers, users, and policymakers understand automated decisions, but warns that even high-performing models may remain opaque, lexically biased, and insensitive to dialect, culture, political context, and affected communities. Overall, the field increasingly recognises that effective automated moderation requires systems that are accurate and scalable but also explainable, fair, context-sensitive, culturally informed, and developed through interdisciplinary collaboration.

Conclusion

Across disciplines, early attempts to define hate speech were done with specific issues of discrimination in mind, born out of local contexts against the backdrop of the second World War. However, as we have demonstrated through this literature review, our understanding of hate speech has since become increasingly complex. Throughout recent times, the issues that established the foundation for defining, regulating and understanding hate speech, although still relevant, do not comprehensively protect against the harms of hate speech at large.

A review of the legal, philosophical, social science and communication disciplines has demonstrated a shift from understanding hate speech as isolated speech acts that can only be understood through immediate effects, to a recognition of hate speech as subjective and contextual, and as a direct threat to democracy which should be regulated as such. Debates surrounding issues of free speech and censorship on the basis of protecting democratic freedoms become weakened when the impact of hate speech just as severely impacts, if not more so, democratic liberties. If there is a commitment to pluralism, then freedoms of a dominant group cannot be prioritized over protecting the minorities. To reflect that, many scholars, as demonstrated above, have attempted to determine the boundaries of hate speech in order to justly protect citizens, although at present this remains challenging as the nature of the reach of law

changes with the shifts in society more broadly. The study of computer science has demonstrated the need for technological advancements to reflect the conceptual changes highlighted by the other disciplines. Challenges faced regarding the boundaries of hate speech become realized in computer science as models are reliant on definitional clarity, as such, scholars in this field emphasize the need for a joint approach within the field of computer science but also among scholars of other disciplines.

In short, this review highlights the challenges in defining hate speech, while emphasizing a general consensus among various disciplines calling for the protection against hate speech and a shift towards subjective and contextual understandings of online hostility.

Bibliography

1. Alkiviadou, Natalie. 2025. "Platform Liability, Hate Speech and the Fundamental Right to Free Speech." *Information & Communications Technology Law* 34(2): 207–17. <https://doi.org/10.1080/13600834.2024.2411799>.
2. Alkoma, F., and M. Xiaogang. 2022. "A Literature Review of Textual Hate Speech Detection Methods and Datasets." *Information* 13(6): 273. <https://doi.org/10.3390/info13060273>.
3. Altman, A. (2005). Freedom of speech and religion. In *The Oxford Handbook of Practical Ethics* (p. 358). Oxford Handbooks Online.
4. Anagnostopoulos, Dorothea, Shawn Everett, and Carolyn Carey. 2013. "'Of Course, We're Supposed to Move On, but Then You Still Got People Who Are Not over Those Historical Wounds': Cultural Memory and U.S. Youth's Race Talk." *Discourse & Society* 24(2): 163–85. <https://doi.org/10.1177/0957926512469389>.
5. Argyle, L. P., Bail, C. A., Busby, E. C., Gubler, J. R., Howe, T., Rytting, C., ... & Wingate, D. (2023). Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41), e2311627120.
6. Baele, Stéphane J., Lewys Brace, and Travis G. Coan. 2023. "Uncovering the Far-Right Online Ecosystem: An Analytical Framework and Research Agenda." *Studies in Conflict & Terrorism* 46(9): 1599–1623. <https://doi.org/10.1080/1057610X.2020.1862895>.
7. Bahador, B. (2023). Monitoring hate speech and the limits of current definition. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 291-298). Berlin <https://doi.org/10.48541/dcr.v12.17>
8. Baider, Fabienne. 2023. "Accountability Issues, Online Covert Hate Speech, and the Efficacy of Counter-Speech." *Politics and Governance* 11(2). <https://doi.org/10.17645/pag.v11i2.6465>.
9. Bail, Christopher. 2022. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton, NJ: Princeton University Press.
10. Tucker, Joshua A., Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. "Social media, political polarization, and

- political disinformation: A review of the scientific literature." *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* (2018).
11. Barlow, Caitlin, and Imran Awan. 2016. ““You Need to Be Sorted Out with a Knife”: The Attempted Online Silencing of Women and People of Muslim Faith within Academia.” *Social Media + Society* 2(4). <https://doi.org/10.1177/2056305116678896>.
 12. Barth, N., Wagner, E., Raab, P., & Wiegärtner, B. (2023). Contextures of hate: Towards a systems theory of hate communication on social media platforms. *The Communication Review*, 26(3), 209-252.
 13. Bäuml, J., Kaufhold, M. A., Voronin, G., & Reuter, C. (2024). Towards an online hate speech classification scheme for german law enforcement and reporting centers: Insights from research and practice. In *Mensch und Computer 2024-Workshopband*. Gesellschaft für Informatik eV.
 14. Bäuml, J., et al. 2025. “Harnessing Inter-Organizational Collaboration and Automation to Combat Online Hate Speech: A Qualitative Study with German Reporting Centers.” *Proceedings of the ACM on Human-Computer Interaction* 9(2): 1–31. <https://doi.org/10.1145/3710991>.
 15. Bengani, Paresh, and Janet Kao. 2023. “How Verified Accounts on X Thrive While Spreading Misinformation About the Israel-Hamas Conflict.” *ProPublica*, December 13. <https://www.propublica.org/article/x-verified-accounts-misinformation-israel-hamas-conflict>.
 16. Bentivegna, Sara, and Rossella Rega. 2024. “Politicians under Fire: Citizens’ Incivility against Political Leaders on Social Media.” *Social Media + Society* 10(4): 20563051241298415.
 17. Bhat, P., & Klein, O. (2020). Covert hate speech: White nationalists and dog whistle communication on twitter. In *Twitter, the public sphere, and the chaos of online deliberation* (pp. 151-172). Cham: Springer International Publishing.
 18. Bilewicz, Michał, and Wiktor Soral. 2020. “Hate Speech Epidemic: The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization.” *Political Psychology* 41: 3–33.
 19. Brink, David O. 2001. “Millian Principles, Freedom of Expression, and Hate Speech.” *Legal Theory* 7(2): 119–57.
 20. Brown, A., & Sinclair, A. (2019). The politics of hate speech laws. Routledge.

21. Brown, Alexander. 2018. "What Is So Special about Online (as Compared to Offline) Hate Speech?" *Ethnicities* 18(3): 297–326. <https://doi.org/10.1177/1468796817709846>.
22. Brown, A. (2020). Models of governance of online hate speech. *online*, Council of Europe, <https://rm.coe.int/modelsof-governance-of-online-hate-speech/16809e671d> (Retrieved: November 26, 2021).
23. Build Up. 2017. *The Commons: A Pilot Methodology for Addressing Polarization Online*. San Francisco, CA: Build Up. <https://howtobuildup.org/programs/digital-conflict/>.
24. Butler, Judith. 2019. "What Threat the Campaign against 'Gender Ideology'." *Glocalism: Journal of Culture, Politics and Innovation*, no. 3. <https://doi.org/10.12893/gjcpi.2019.3.1>.
25. Calvert, Clay. 1997. "Hate Speech and Its Harms: A Communication Theory Perspective." *Journal of Communication* 47(1): 4–19. <https://doi.org/10.1111/j.1460-2466.1997.tb02690.x>.
26. Cammaerts, Bart. 2009. "Radical Pluralism and Free Speech in Online Public Spaces." *International Journal of Cultural Studies* 12: 555–75.
27. Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and violent behavior*, 58, 101608.
28. Center for Countering Digital Hate. 2023 *X Content Moderation Failure*. https://counterhate.com/wp-content/uploads/2023/09/230907-X-Content-Moderation-Report_final_CCDH.pdf
29. Cisneros, David J., and Thomas K. Nakayama. 2015. "New Media, Old Racisms: Twitter, Miss America, and Cultural Logics of Race." *Journal of International and Intercultural Communication* 8(2): 108–27. <https://doi.org/10.1080/17513057.2015.1025328>.
30. Clausen, A., Frischlich, L., & Mayer, P. (2026). Social media platforms and the spreading of hate speech targeting sexual and gender minorities: a scoping review. *Information, Communication & Society*, 1-21.
31. Cortiz, Diogo, and Arkaitz Zubiaga. 2021. "Ethical and Technical Challenges of AI in Tackling Hate Speech." *The International Review of Information Ethics* 29. <https://doi.org/10.29173/irie416>.
32. Council of the European Union. 2008. "Council Framework Decision 2008/913/JHA of 28 November 2008 on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law."

33. Coutu, Cameron, Martin Blais, Mariia Samoilenko, and Isabel Côté. 2024. "Cyberviolence Exposure among LGBTQ2+ People in Quebec (Canada): An Intersectional Exploration." SSRN, March. <https://doi.org/10.2139/ssrn.4737719>.
34. Delgado, Richard. 1995. "Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling." *Harvard Civil Rights-Civil Liberties Law Review* 17: 133–81.
35. Domingo, David, Thorsten Quandt, Ari Heinonen, Steve Paulussen, Jane B. Singer, and Marina Vujnovic. 2008. "Participatory Journalism Practices in the Media and Beyond." *Journalism Practice* 2: 326–42.
36. ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021, November). Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 345-363).
37. Erjavec, Karmen, and Melita P. Kovačič. 2012. "'You Don't Understand, This Is a New War!' Analysis of Hate Speech in News Web Sites' Comments." *Mass Communication and Society* 15(6): 899–920. <https://doi.org/10.1080/15205436.2011.619679>.
38. European Observatory of Online Hate. 2025. "Definitions: Hate Speech." Accessed September 19, 2025. <https://eoooh.eu/faq>.
39. Evolvi, Giulia. 2019. "#Islamexit: Inter-Group Antagonism on Twitter." *Information, Communication & Society* 22(3): 340-386. <https://doi.org/10.1080/1369118X.2017.1388427>.
40. Fillies, J., Hoffmann, M. P., Reichel, R., Salzwedel, R., Bodemer, S., & Paschke, A. (2025, November). Mapping Toxic Comments Across Demographics: A Dataset from German Public Broadcasting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 18774-18790).
41. Fish, S. (2012). The harm in free speech. *The New York Times Opinionator Blog*.
42. Fish, Stanley. 2012. "The Harm in Free Speech." *The New York Times*, June 4. <https://archive.nytimes.com/opinionator.blogs.nytimes.com/2012/06/04/the-harm-in-free-speech/>.
43. Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4), 1-30.

44. Govers, J., Feldman, P., Dant, A., & Patros, P. (2023). Down the rabbit hole: Detecting online extremism, radicalisation, and politicised hate speech. *ACM Computing Surveys*, 55(14s), 1-35.
45. Harlow, Summer. 2015. "Story-Chatterers Stirring Up Hate: Racist Discourse in Reader Comments on U.S. Newspaper Websites." *Howard Journal of Communications* 26(1): 21–42. <https://doi.org/10.1080/10646175.2014.984795>.
46. Harris, C., Halevy, M., Howard, A., Bruckman, A., & Yang, D. (2022, June). Exploring the role of grammar and word choice in bias toward African American English (AAE) in hate speech classification. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 789-798).
47. Hawke, J., and Helena Puig Larrauri. 2025. "Towards a Third Side on Social Media." *Peacebuilding*: 1–16. <https://doi.org/10.1080/21647259.2025.2583866>.
48. Hawke, J., Mira El Mawla, Caleb Gichuhi, and Helena Puig Larrauri. 2025. "Archetypes of Polarisation on Social Media." OSF Preprint, November 6. https://doi.org/10.31235/osf.io/mfs5r_v1.
49. Hietanen, Mika, and Johan Eddebo. 2023. "Towards a Definition of Hate Speech—With a Focus on Online Contexts." *Journal of Communication Inquiry* 47(4): 440–58. <https://doi.org/10.1177/01968599221124309>.
50. Howard, Jeffrey W. 2019. "Free Speech and Hate Speech." *Annual Review of Political Science* 22(1): 93–109. <https://doi.org/10.1146/annurev-polisci-051517-012343>.
51. International Network Against Cyber Hate (INACH). 2026. "Cyber Hate Definitions." <https://www.inach.net/cyber-hate-definitions/>.
52. Kansok-Dusche, J., Ballaschk, C., Krause, N., ZeiBig, A., Seemann-Herz, L., Wachs, S., & Bilz, L. (2023). A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *Trauma, violence, & abuse*, 24(4), 2598-2615.
53. Kasimov, Andrey. 2023. "Decentralized Hate: Sustained Connective Action in Online Far-Right Community." *Social Movement Studies*: 1–19. <https://doi.org/10.1080/14742837.2023.2204427>.
54. Kentmen-Cin, C. (2025). Hate speech on social media: A systemic narrative review of political science contributions. *Social Sciences*, 14(10), 610.

55. Khurana, U., Vermeulen, I., Nalisnick, E., Van Noorloos, M., & Fokkens, A. (2022, July). Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* (pp. 176-191).
56. Kim, Jin Woo, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. "The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity." *Journal of Communication* 71(6): 922–46. <https://doi.org/10.1093/joc/jqab034>.
57. Kindermann, Dirk. 2023. "Against 'Hate Speech.'" *Journal of Applied Philosophy* 40(5): 813–35. <https://doi.org/10.1111/japp.12648>.
58. Leets, Laura. 2001. "Explaining Perceptions of Racist Speech." *Communication Research* 28(5): 676–706. <https://doi.org/10.1177/009365001028005005>.
59. Lim, Merlyna. 2017. "Freedom to Hate: Social Media, Algorithmic Enclaves, and the Rise of Tribal Nationalism in Indonesia." *Critical Asian Studies* 49(3): 411–27.
60. Luo, C., Bhambhoria, R., Dahan, S., & Zhu, X. (2023, December). Legally enforceable hate speech detection for public forums. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 10948-10963).
61. Malecki, W. P., Kowal, M., Dobrowolska, M., & Sorokowski, P. (2021). Defining online hating and online haters. *Frontiers in psychology*, 12, 744614.
62. Marcks, Holger, and Janina Pawelz. 2022. "From Myths of Victimhood to Fantasies of Violence: How Far-Right Narratives of Imperilment Work." *Terrorism and Political Violence* 34(7): 1415–32. <https://doi.org/10.1080/09546553.2020.1788544>.
63. McDevitt, Jack, Jack Levin, and Susan Bennett. 2002. "Hate Crime Offenders: An Expanded Typology." *Journal of Social Issues* 58(2): 303–17.
64. Mchangama, Jacob, and Natalie Alkiviadou. 2021. "Hate Speech and the European Court of Human Rights: Whatever Happened to the Right to Offend, Shock or Disturb?" *Human Rights Law Review* 21(4): 1008–42. <https://doi.org/10.1093/hrlr/ngab015>.
65. Meddaugh, Priscilla, and Jack Kay. 2009. "Hate Speech or 'Reasonable Racism?' The Other in Stormfront." *Journal of Mass Media Ethics* 24(4): 251–68. <https://doi.org/10.1080/08900520903320936>.
66. Merrill, Samuel, and Mathias Åkerlund. 2018. "Standing Up for Sweden? The Racist Discourses, Architectures and Affordances of an Anti-Immigration Facebook Group."

- Journal of Computer-Mediated Communication* 23(6): 332–53.
<https://doi.org/10.1093/jcmc/zmy018>.
67. Meske, Christian, and Elisabeth Bunde. 2023. “Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection.” *Information Systems Frontiers* 25(3): 743–73. <https://doi.org/10.1007/s10796-021-10234-5>.
 68. Monnier, Angélique, Amandine Boursier, and Antonio Seoane, eds. 2022. *Cyberhate in the Context of Migrations*. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-030-92103-3>.
 69. Mullah, N. S., and W. M. N. W. Zainon. 2021. “Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review.” *IEEE Access* 9: 88364–76. <https://doi.org/10.1109/ACCESS.2021.3089515>.
 70. Mun, J., C. Buerger, J. T. Liang, J. Garland, and M. Sap. 2024. “Counterspeakers’ Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate.” In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–22.
 71. Munger, Kevin. 2017. “Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment.” *Political Behavior* 39(3): 629–49.
 72. Parvaresh, V. (2023). Covertly communicated hate speech: A corpus-assisted pragmatic study. *Journal of Pragmatics*, 205, 63–77.
 73. Paz, María A., Jesús Montero-Díaz, and Alicia Moreno-Delgado. 2020. “Hate Speech: A Systematized Review.” *SAGE Open* 10(4): 2158244020973022. <https://doi.org/10.1177/2158244020973022>.
 74. Peršak, Nina. 2022. “Criminalising Hate Crime and Hate Speech at EU Level: Extending the List of Eurocrimes under Article 83(1) TFEU.” *Criminal Law Forum* 33(2): 85–119. <https://doi.org/10.1007/s10609-022-09440-w>.
 75. Picazo-Sanchez, Pablo. 2025. “Breaking the Silence: The Rise of Hate Speech in Spanish Politics.” *Journal of Computational Social Science* 8(3): 81. <https://doi.org/10.1007/s42001-025-00413-8>.
 76. Qureshi, Muhammad, Muhammad Atif Qureshi, and Wael Rashwan. 2026. “Explainable AI for Hate Speech Moderation: A Stakeholder-Centered and Sociotechnical Review.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 16. <https://doi.org/10.1002/widm.70076>.

77. Rathje, Steve, Jay J. Van Bavel, and Sander Van der Linden. 2021. "Out-Group Animosity Drives Engagement on Social Media." *Proceedings of the National Academy of Sciences* 118(26): e2024292118.
78. Rawat, A., Kumar, S., & Samant, S. S. (2024). Hate speech detection in social media: Techniques, recent trends, and future challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(2), e1648.
79. Rosenfeld, M. (2002). Hate speech in constitutional jurisprudence: a comparative analysis. *Cardozo L. Rev.*, 24, 1523.
80. Rossini, Patrícia, and Rousiley Maia. 2021. "Characterizing Disagreement in Online Political Talk: Examining Incivility and Opinion Expression on News Websites and Facebook in Brazil." *Journal of Deliberative Democracy* 17(1): 90–104. <https://doi.org/10.16997/jdd.967>.
81. Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*.
82. Schäfer, S., Rebasso, I., Boyer, M. M., & Planitzer, A. M. (2024). Can we counteract hate? Effects of online hate speech and counter speech on the perception of social groups. *Communication Research*, 51(5), 553-579.
83. Schirch, Lisa, Katarina Radivojevic, and Christian Buerger. 2025. "A Taxonomy of Response Strategies to Toxic Online Content: Evaluating the Evidence." arXiv. <https://arxiv.org/abs/2509.09921>.
84. Scrivens, Ryan, and Amarnath Amarasingam. 2020. "Haters Gonna 'Like': Exploring Canadian Far-Right Extremism on Facebook." In *Digital Extremisms: Readings in Violence, Radicalisation and Extremism in the Online Space*, eds. Mark Littler and Benjamin Lee, 63–89. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-30138-5_4.
85. Sellars, Andrew (2016). Defining hate speech. *Berkman Klein Center Research Publication*, (2016-20), 16-48.
86. Shepherd, Tamara, Alison Harvey, Tim Jordan, Sam Srauy, and Kate Miltner. 2015. "Histories of Hating." *Social Media + Society* 1(2). <https://doi.org/10.1177/2056305115603997>.
87. Siegel, A. A. (2020). Online hate speech. *Social media and democracy: The state of the field, prospects for reform*, 56-88.

88. Simchon, Almog, William J. Brady, and Jay J. Van Bavel. 2022. "Troll and Divide: The Language of Online Polarization." *PNAS Nexus* 1(1): pgac019.
89. Sobkowicz, Pawel, and Anna Sobkowicz. 2010. "Dynamics of Hate-Based Internet User Networks." *The European Physical Journal B* 73(4): 633–43.
90. Stop Hate Speech. 2025. "Was ist Hate Speech?" Accessed September 18, 2025. <https://stophatespeech.ch/pages/definition-von-hate-speech>.
91. Stray, Jonathan, Ravi Iyer, and Helena Puig Larrauri. 2023. "The Algorithmic Management of Polarization and Violence on Social Media." [INCOMPLETE—publication venue, publisher, page range, and URL/DOI not supplied]
92. Theocharis, Yannis, Pablo Barberá, Zoltán Fazekas, and Sebastian Adrian Popa. 2020. "The Dynamics of Political Incivility on Twitter." *SAGE Open* 10(2). <https://doi.org/10.1177/2158244020919447>.
93. Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: topics of interest and their evolution: A. Tontodimamma et al. *Scientometrics*, 126(1), 157-179.
94. Trifiro, B. M., Paik, S., Fang, Z., & Zhang, L. (2021). Politics and politeness: Analysis of incivility on Twitter during the 2020 democratic presidential primary. *Social Media+ Society*, 7(3), 20563051211036939.
95. Tsesis, Alexander. 2002. *Destructive Messages: How Hate Speech Paves the Way for Harmful Social Movements*. New York: New York University Press.
96. UN General Assembly (21 December 1965). "International Convention on the Elimination of All Forms of Racial Discrimination". United Nations Treaty Series. 660: 195. <https://www.refworld.org/legal/agreements/unga/1965/13974>
97. Van Swol, Lyn M., Sangwon Lee, and Rachel Hutchins. 2022. "The Banality of Extremism: The Role of Group Dynamics and Communication of Norms in Polarization on January 6." *Group Dynamics: Theory, Research, and Practice* 26(3): 239.
98. Vergani, Matteo, Barbara Perry, Joshua Freilich, Steven Chermak, Ryan Scrivens, Rouven Link, Daniel Kleinsman, John Betts, and Muhammad Iqbal. "Mapping the scientific knowledge and approaches to defining and measuring hate crime, hate speech, and hate incidents: A systematic review." *Campbell Systematic Reviews* 20, no. 2 (2024): cl2-1397.

99. Wachs, Sebastian, Michelle F. Wright, and Manuel Gámez-Guadix. 2024. "From Hate Speech to HateLess: The Effectiveness of a Prevention Program on Adolescents' Online Hate Speech Involvement." *Computers in Human Behavior* 157: 108250. <https://doi.org/10.1016/j.chb.2024.108250>.
100. Waldron, Jeremy. 2012. *The Harm in Hate Speech*. Cambridge, MA: Harvard University Press.
101. Waltman, Michael, and John Haas. 2011. *The Communication of Hate*. New York: Peter Lang.
102. Warner, William, and Julia Hirschberg. 2012. "Detecting Hate Speech on the World Wide Web." In *Proceedings of the Second Workshop on Language in Social Media*, 19–26.
103. Whillock, Rita Kirk, and David Slayden, eds. 1995. *Hate Speech*. Thousand Oaks, CA: Sage.