Hate speech

Sponholz, Liriam

Erstveröffentlichung / Primary Publication
Sammelwerksbeitrag / collection article

Digital
Communication
Research.de

Recommended citation: Sponholz, L. (2023). Hate speech. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 143–163). Digital Communication Research. https://doi.org/10.48541/dcr.v12.9

Abstract: Hate speech—communication that attacks a person or a group on the basis of identity factors, such as gender, race, or religion—is one of the main digital threats to democracy. Hate speech has manifold, empirically evidenced consequences for targeted individuals and groups experiencing systematic discrimination and for social cohesion as a whole. Yet, while the upheaval of social media has put the concept in the spotlight, such attention has also structurally transformed its meaning, turning hate speech from a concept with clear defining properties into a family resemblance comprising all kinds of online abuse. This process is far from causing only academic issues. It also sidesteps historical oppression as a defining property and as the reason for which one is targeted by hate speech. Thus, the process has been belittling public animosity against historically oppressed groups, reducing hate speech merely to a matter of offensive language on social media. This chapter shows how and why this conceptual change has taken place and the consequences it unleashes. It specifically addresses the problems of concept stretching, concept shrinking, and the inflation of concepts. Finally, it concludes that such conceptual issues jeopardize the potential that digital media research on hate speech has to provide guidance to a broad range of social actors.

DOI 10.48541/dcr.v12.9

Liriam Sponholz

Hate Speech

1 Hate speech: What is the concept actually for?

Berlin, 2021: The artist Prince Ofori goes to a supermarket and is called the N-word. Customers, workers, security guards—none of them defends him. To the contrary, they start to collectively disparage Ofori, a Black man. Eventually, even the supermarket manager accuses him of being a security risk and throws him out of the store (Amjahid, 2021). Vienna, 2019: A woman wearing a headscarf is spat on in a train station. The woman is called a w***re, pig, and dog and told to go back to the place where she is supposed to belong. "The FPÖ [Austrian far right "Freedom Party," L.S.] will take you all," shouts her harasser ("Alltagsrassismus. Angespuckt und beschimpft," 2019).

In the 1980s, law scholars associated with historically oppressed groups sought to tackle this kind of situation by coining the concept of hate speech to describe the communication of animosity against, or disparagement of people of a historically oppressed group on the basis of identity factors (cf. Matsuda, 1989; Stone, 2000; Delgado & Stefancic, 2004, among others). These scholars were involved with critical race theory and had similar experiences on US-American campuses. To tackle the problem, they proposed that severe cases of hate speech (cf. Matsuda, 1989), such as those in Berlin and Vienna, should be outlawed.

At that time, the concept raised concerns about freedom of speech and was highly criticized. Thirty years later, the debate changed (cf. Tontodimamma et al., 2021). People from historically oppressed groups continue to experience the same experiences, but the concept of hate speech, instead of being rejected, has now been stretched and made ambiguous, leading to a downplaying of the problem.

Nowadays, digital communication enables everyone to gain insight into what it means to be publicly disparaged. By doing so, the digital transformation of the public sphere put the concept of hate speech in the spotlight but has also led to concept stretching (Collier & Mahon, 1993), that is, to the application of the term "hate speech" to cases that do not match its defining properties (cf. Sponholz, 2020).

Researchers on digital communication have been particularly guilty of damaging the clarity of the concept, often without realizing that they are doing so. For instance, they mention the original concept of hate speech in the theoretical part of their studies and then apply the term to cases of online harassment against journalists, online incivility, online abuse, and other forms of conflict that do not match the defining properties of the concept they have just mentioned (cf. Tontodimamma et al., 2021).

However, incidents such as those in Berlin and Vienna have been framed as other than hate speech, with a resultant downplaying of their severity. Therefore, the concept was appropriated by the same patterns of power asymmetry that it was intended to counter.

This chapter sheds light on the problem by analyzing how a concept coined by critical race theory came to be equated with online harassment, what role academic research has played in this development, and why this equation is a problem. As will be shown, hate speech is not a catch-all term for all conflict-related issues involving online communication, nor can it be replaced by catch-all terms such as "online hate." While on the one hand there are serious conceptual issues, such as concept stretching and shrinking, on the other hand, there is a broad consensus among different social actors about what hate speech is. Thus, communication and media scholars should play a significant role in overcoming these issues since hate speech is the key to understanding, explaining, empirically assessing, and tackling extreme forms of symbolic discrimination, one of the most severe digital threats to democracy and social cohesion.

2 Why do concepts matter?

Concepts are not merely a matter of abstract discussion among academics. They constitute a symbolical resource. They are deployed not only to determine how a research subject is assessed empirically but also to evaluate a situation, to define a problem (Thielmann 2004, p. 292, p. 310), to examine the way in which that problem should be tackled, which policies should be employed to tackle it (Palonen, 1999), which statistics should be used to underpin those policies, and—in the case of conflict regulation—who and what should be outlawed and how. When so many rides on a concept, the process of defining the concept becomes a struggle over a resource (Cobb & Elder, 1972), with politicians, governments, and digital platform companies fighting for a definition that best suits their political or economic interests.

"Hate speech" paradigmatically illustrates the struggle over concepts as symbolic resources. Far right actors build their media capital by making disparaging statements against Black people, Indigenous people, Jewish people, LGBTQ people, women, and Muslims and present themselves as victims of hate speech when they face offensive language in response to these statements. This is the case, for instance, when the Austrian far-right leader Heinz Christian Strache complained of hate against his party (Strache sieht in FPÖ-Hass, 2015).

In the sociotechnical realm, digital platform companies, whose economic model is based on interactions, have managed to establish the idea that "the best remedy against bad speech is more speech" (Brändlin, 2016). In line with this principle, Facebook launched the Online Civil Courage Initiative, a project encouraging people to speak up against hate speech. Its CEO, Mark Zuckerberg, also defended the right of Holocaust deniers "to be wrong" (Levin & Solon, 2018). The company only agreed to ban Holocaust denial content under pressure in 2020 (Bickert, 2020). In this context, hate speech has been treated as a matter of uncivil comments (see Coe et al., 2014, and Bormann & Ziegele in this collection for a discussion of the incivility concept), although incivility is not necessarily linked to identity factors, and hate speech, whether online or not, is not restricted to comments or content. However, by turning hate speech into a matter of incivility, digital platform companies a) veil their own role in triggering hate speech (for instance, through scoring or recommendation algorithms); b) enable haters to continue generating interactions and building networks around discriminatory content; c) induce individual

users and collective actors such as high-profile, well-intentioned organizations from civil society to work for them by producing content against hate speech; d) increase interactions not only through hate speech but also through counter speech; and e) polish their images by promoting such initiatives while tolerating hate speech. In taking this course, they fail to tackle the problems that individuals and societies have been suffering as a consequence of group-targeting, offensive, and inflammatory speech on social media, as the genocide in Myanmar, the riots in Chemnitz in Germany, and the online mobilization that led to the storming of the Capitol in the US illustrate.

3 What actually is hate speech?

Defining hate speech pose a particular challenge for research on digital communication, specifically with regard to online content moderation and automated detection of hate speech. On the one hand, researchers complain that there is no "universally accepted" concept of hate speech (MacAvaney et al., 2019, p. 2). On the other hand, they not only fail to make contributions that tackle this problem but even create more ambiguity by associating the term with different classes of objects, such as:

> *Abusive messages*, *hostile* messages, or *flames*. More recently, many authors have shifted to employing the term *cyberbullying* (Xu et al., 2012; Hosseinmardi et al., 2015; Zhong et al., 2016; Van Hee et al., 2015; Dadvar et al., 2013; Dinakar et al., 2012). The actual term *hate speech* is used by Warner and Hirschberg (2012), Burnap and Williams (2015), Silva et al. (2016), Djuric et al. (2015), Gitari et al. (2015), Williams and Burnap (2015), and Kwok and Wang (2013). Further, Sood et al. (2012a) worked on detecting (personal) *insults*, *profanity*, and user posts that are characterized by *malicious intent*, while Razavi et al. (2010) referred to *offensive language*. Xiang et al. (2012) focused on *vulgar language* and *profanity-related offensive content*. (Schmidt & Wiegand, 2017, p. 2-3)

However, the question remains: What is the problem with the concept of hate speech? Answering this question requires an understanding of what a concept is and what it is made up of.

Concepts are basically a matter of word and meaning (intension) and meaning and things (extension) (Sartori, 1984). This is the classical structure of a concept

(Marsteintredet & Malamud, 2020, p. 1025). In the academic context, concepts are applied by researchers to identify, describe, classify, understand, or explain what they observe (Sellars, 2016, p. 4). The intension of a concept consists of defining properties, that is, *criteria* that delimitate the scope of the term. The extension, in turn, determines the *class of objects* to which this meaning applies. Intension and extension are indirectly proportional: the fewer defining properties a concept has, the more abstract it is. The more abstract it is, the greater the number of objects that match it (Sartori 1984, p. 45).

Deficiencies in the intension and extension of a concept create different issues. Problems with intension create ambiguity. This is the case when the meaning of a term is not anchored in defining properties. Problems with the extension of a concept create vagueness. This is the case when a concept is too abstract, which makes the class of objects it applies unclear (Sartori, 1984, p. 27).

Hence, the question of what the problem with the concept of hate speech actually is can be answered. First of all, the problem does not lie in the intension of the concept.

The term "hate speech" is drawn by the following defining properties (DP): attacks (DP1) based on an identity factor (DP2), which are symbolic in nature (DP3) (Matsuda, 1989; Stone, 2000; Delgado & Stefancic, 2004; among others). Hate speech—whether online or not—is also a matter of communication in places of public space (cf. Sellars, 2016; Delgado & Stefancic, 2004). Nonetheless, this is not a classical defining property, as it may also apply to other cases of communication of disparagement, such as online incivility (cf. Sponholz, 2020).

There is a broad consensus, from international organizations to digital platform companies, about the linkage of the term hate speech with these defining properties, as follows:

- United Nations: Any kind of communication in speech, writing or behavior [DP3] that attacks or uses pejorative or discriminatory language [DP1] 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, color, descent, gender or other identity factor [DP2]. (United Nations, 2020, p. 8)

- Facebook Company: We define hate speech as a direct attack [*DP1*] against people on the basis of what we call protected characteristics: race, ethnicity,

national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease [*DP2*]. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation [*DP3*]. (Facebook, 2021)

- Twitter: You may not promote violence against or directly attack or threaten other people [*DP1, DP3*] on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease [*DP2*]. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories. (Twitter, 2020)

- Council of Europe: The term "hate speech" shall be understood as covering all forms of expression [*DP3*] which spread, incite, promote or justify [*DP1*] racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin [*DP2*]. (Weber, 2009, p. 3)

These definitions are not identical. Nevertheless, as Sartori (1984, p. 29) asserts, a single concept can yield several conceptualizations. The same concept may, for instance, yield both denotative and operational definitions, but as long as different definitions possess the same defining properties, they still constitute the same concept.

4 What is not hate speech?

The intension of the concept of hate speech not only enables a determination of what hate speech is but also what it is not.

First, hate speech is not negative stereotypes or misrepresentation but a matter of attacks. Defining hate speech as "negative speech that targets individuals or groups" or even as "statements of disagreement, such as indications that the

group is wrong, what they claim is false or what they believe is incorrect" (Bahador & Kerchner, 2019, pp. 4–6) downplays the severity of the problem.

Negative stereotypes may be used by hate speakers, but they alone are not enough to constitute an attack. To be considered hate speech, it should be applied consciously or intentionally (Delgado & Stefancic, 2004). What precisely constitutes an attack is delineated in, for instance, General Recommendation Nr. 35 of the Committee on the Elimination of Racial Discrimination (2013): incitement of hatred, contempt, exclusion, or violence; threats or expressions of insults, ridicule, or slander (for an overview, see Table 1).

In this context, "conscious" means that the speaker is aware of the disparaging potential of the content, as in the case of identity-targeting offensive speech. "Intentional," in turn, means that symbolic disparagement is a way of achieving a goal. This goal may be hurting someone or derogating a group due to the ideological convictions of the speaker (prior intention), or it may be something other, such as gaining media attention or attracting voters in an election (subsidiary intention) (cf. Searle, 1980; Sponholz, 2018).

Second, not all disparagements of groups qualify as hate speech (cf. Sellars, 2016), only those based on identity factors in correlation with historical oppression (Matsuda, 1989) or systematic discrimination (Gelber, 2021). Hate speech represents the communicative ring on a chain of manufacturing human inferiority (Sponholz, 2018, p. 48), in which antinomies (Marková, 2003) on collective features such as race, gender, origin, religion, and sexual orientation are intentionally activated through communication. It works as another layer in the long-standing process of subordination (Matsuda, 1989). This is why "not everyone has known the experience of being victimized by racist, misogynist, or homophobic speech, and we do not share equally the burden of the societal harm it inflicts" (Lawrence III, 1993, p. 56).

A definition that takes a broader view of groups, contending that theoretically any group can become the target of hate speech, is exactly what critical race theorists were fighting against. It means erasing the power asymmetry that Lawrence III (1993) referred to. This does not mean that people can be attacked symbolically only if they possess one of the designated identity factors, but it clearly means that if there is not an identity factor involved, this kind of abuse or harassment should not be classified as hate speech.

Table 1: Defining properties of hate speech

Who	What		Where
Collective feature corresponding to an identity factor (e.g., gender/race) related to an unprivileged position (e.g., Women/Black people)	Dissemination of	Discriminatory ideas	Places of public life (workplace, school, university campus, media)
	Incitement of	• Hatred • Contempt • Exclusion • Violence	
	Incitement through	• Public denial of genocide and crimes against humanity	
	Threat		
	Justification of	• Genocides and crimes against humanity	
	Expression of	• Insults • Ridicule • Slander	

Source: Own illustration, based on Committee on the Elimination of Racial Discrimination (2013), Delgado and Stefancic (2004), and Matsuda (1989), among others

Third, hate speech is not necessarily a matter of (offensive) language but of communication (Stone, 2000; United Nations, 2020). This is particularly important when it comes to social media, where the media logic is not based on content—as with the mass media—but on interactions (van Dijck & Poell, 2013). As a result, digital communication is not only a matter of media objects (such as online comments) but also of other digital objects (Langlois & Elmer, 2013): network objects, such as hashtags (Poole et al., 2021); and phatic objects, that is, the networks generated by interactions on social networking digital platforms (Chaudhry, 2015). For this reason, hate speech cannot be detected solely by content analysis but also requires social network analysis and social media metrics analysis (Sponholz, 2021). As a consequence, countering the problem should not be limited to content moderation, but should include debates on de-platforming (Ali et al., 2021), cross-platform approaches (Johnson et al., 2019), and broader concepts of platform governance.

It is worth noting that neither hate (as an emotion) nor illegality are defining properties of the concept of hate speech. Brown (2017a) labels this first misunderstanding "the myth of hate," that is, the idea that emotions, feelings, or attitudes of hate or hatred are part of the essential nature of hate speech (cf. also Tetrault, 2019). The roots of hate speech are ideologies of inequality, such as racism, sexism, homophobia, and Islamophobia, not affective action. These ideologies gather enough empirical evidence that they can be expressed "rationally"; that is, they are underpinned by arguments (Sponholz, 2018; Meddaugh & Kay, 2009). To put it briefly, David Irving's denying the Holocaust is not an emotional response.

With regard to legality, although law scholars coined the concept, they also made it clear from the beginning that only a very strict range of cases could be regulated (Matsuda, 1989). Further, the upsurge of the concept in public and academic debate has not been triggered by legal issues but by the rise of social media (Paz et al., 2020; Sponholz, 2019b). Media and communication researchers have dominated this research agenda since the 2010s and apply the concept first to observe conflict dynamics, and not to matters of conflict regulation.

As seen above, the concept of hate speech has a clear definition (intension), with broad consensus on the application of the term anchored in the same defining properties. However, in spite of a clear intension, the concept is highly abstract and does not provide a clear extension, that is, an explicit scope for the class of objects to which it applies to (Sartori, 1984). Such vagueness lends it flexibility, but it also poses a challenge for empirical research.

However, it should be highlighted that flexibility in this context does not mean that the concept of hate speech can be applied to all kinds of wrongdoing in online communication, as often happens in digital media research. It actually means that the concept can be applied to a broad range of empirical manifestations, such as online firestorms or hashtag activism, as *long as they match the defining properties.*

By applying the term "hate speech" to offensive language in general, research on digital communication not only fails to tackle the concept's vagueness but also generates new conceptual issues. This is concept stretching—that is, the application of a concept to cases that do not fit its defining properties. In other instances, researchers on digital communication, particularly those working in automated detection, are shrinking the concept by applying it only to identity-targeting

group derogatory labels, such as racial slurs. In the third scenario, an inflation of concepts has taken place, failing to solve old concerns and generate new ones.

5 Conceptual issues within academic research

Epistemologically, definitions are pivotal to increasing and even enabling the efficiency of science (Potthof, 2017). They allow social scientists to avoid talking at cross-purposes when addressing a research subject, which means that empirical findings can be compared and knowledge can be accumulated. This, in turn, allows for the development of theories. Theorizing is imperative to understanding and explaining the puzzle of social reality.

However, when scholars expand the comparative perspective among research areas, they also tend to broaden the meaning of the concepts to incorporate a larger realm of observations under expanded rubrics (Sartori, 1984). The result is a conceptual travelling. This happened to hate speech, a concept coined by law scholars in the 1980s, when the rise of social media in the early 2000s turned the term into an interdisciplinary research subject (Paz et al., 2020; Sponholz, 2019b).

Although conceptual traveling may result in a concept being more relevant, it may also feed a conceptual stretching (Collier & Mahon, 1993). By not considering the intension of hate speech, researchers on media and communication have been applying the concept to a class of objects that do not possess the same defining properties, such as online harassment against journalists (Obermaier et al., 2018).

Different concepts might have the same, contingent, or accidental characteristics, but if they do not possess all the defining properties, they are not the same (Sartori, 1984). By applying the concept to classes of objects that do not belong under the same umbrella, conceptual stretching creates ambiguity and hinders the comparability of empirical evidence, which harms the accumulation of knowledge and makes it harder to provide qualified guidance when policies are formulated to tackle the problem.

Concept stretching jeopardizes hate speech research by erasing not only a defining property of the concept but also the concept's very reason for being: catching disparaging communication that is based on a collective feature linked to historical oppression or systematic discrimination.

6 Concept shrinking

When attempting to identify hate speech by automated means, researchers on digital media often try to solve the problem of vagueness by reducing the concept to a matter of racial slurs or symbols or open threats ("kill," "rape") (cf. Schmidt & Wiegand, 2017). By defining a hate speech message as a message that contains "hate words" (Silva et al., 2016), lexicon- and keyword-based approaches cannot identify cases that do not contain any "hateful" words (e.g., cases that use figurative or nuanced language) but that still deliberately discriminate symbolically against a group (MacAvaney et al., 2019).

Reducing hate speech to a matter of insults, for instance, would mean coming to the conclusion that racist groups are not libeling or inciting discrimination against historically oppressed groups in instances where they target people due to race, origin, or religion but refrain from writing the N-word or displaying a swastika.

Furthermore, lexicon-based approaches, including those based on offensive language, such as group derogatory labels, are capable of empirically assessing only a small proportion of cases. In the case of group libels, they are also considered a less severe form of hate speech (United Nations, 2020). Such approaches also fail to make visible those collective actors and public figures who apply hate speech as a kind of strategic communication, because such actors tend to avoid blatantly discriminatory language (Gerstenfeld et al., 2003; Kleinberg et al., 2021). Even in the case of hate speech during genocides, overt messages, such as open threats ("Go out and kill!"), constitute an exception (Benesch, 2004, p. 67; Straus, 2007, p. 612).

For this reason, several stakeholders have underlined that hate speech is not a matter of language but of communication (Stone, 2000; United Nations, 2020). While language refers to a system of codes (Lewandowski, 1994), communication involves speakers, messages, means of dissemination, audience, and historical, social, and political context. This is particularly relevant for research into hate speech on social media, as reducing the problem to content does not fit the media logic of such digital platforms, as analyzed earlier.

7 Inflation of concepts

Researchers have also tried to sidestep conceptual issues with hate speech by replacing it with, for instance, the concepts of "online hate" and "extreme speech."

"Online hate" is one of the catch-all terms that scholars have been using in their attempts to capture empirical developments in digital communication. "Online hate" incorporates issues such as "online toxicity," "online abusive language," "cyberbullying," "online harassment," and "online firestorms." That is, it comprises much more than online hate speech (Waqas et al., 2019). Hence, "online hate" is a family resemblance rather than a classical concept:

> One distinctive feature of family resemblance concepts is the fact that everything that falls under the concept shares at least one similar quality, feature, or descriptive property with at least one other thing that falls under the concept, even if there is no single quality, feature, or descriptive property that is common to all things that fall under the concept. (Brown, 2017b, p. 596)

Changing the concept structure from classical to family resemblances has been a successful strategy to capture empirical developments. Nonetheless, at least three problems can be caused by such conceptual change. First, it inhibits the recognition of hate speech. Second, it creates the danger of causal and conceptual confusion. Third, it may have serious political consequences.

Transforming hate speech into a matter of family resemblance means applying the same term based on different criteria depending on the context (Wennerberg, 1998, p. 64), opening the door to all kinds of political instrumentalization. Such adaptions entail many risks, including freedom of speech. The concept may even be turned against historically oppressed groups seeking to speak out about their situations of oppression (cf. Benesch, 2014; Gagliardone et al., 2015).

Applying "hate speech" and "online hate" interchangeably also means erasing discrimination and power asymmetries as the roots of the problem (cf. also Matamoros-Fernández & Farkas, 2021). By doing so, researchers, instead of investigating, assume that insults, threats, or incitement against Black people, women, Jewish people, Muslims, or other groups are the same as any other kind of disparaging communication, such as individual insults and slanders. In assuming that, they fail to look for empirical evidence that would, in the case of hate speech, prove its true nature.

This failure on the part of researchers is particularly problematic, given that it is often hate speech—and not other forms of online abuse—that plays a pivotal role in political developments, such as the rise of the far right and its linkage to digital communication (Art, 2020; Froio & Ganesh, 2019; Sponholz, 2019a). In a nutshell, family resemblance concepts such as "online hate" are successful at capturing empirical developments in digital communication but cannot replace the concept of hate speech.

Relabeling hate speech, in turn, might raise new issues, as is the case with the concept of "extreme speech," as applied within digital communication research (the term was applied before, at the end of the 2000s, by law scholars such as Hare and Weinstein (2009) in another context).

In digital communication research, "extreme speech" aims to provide an alternative, non-regulatory approach to hate speech since:

> The use of hate speech (...) embodies the colonial logic of "yet-to-be modern" societies prone to "emotions," manipulation, and public frenzy, which have to be tested against the high values of calm rationality of Western liberal democracy. (Udupa & Pohjonen, 2019, p. 3055)

To overcome the "Western bias" of the concept of hate speech, the authors propose the concept of "extreme speech," a framework to "capture digital cultures that push and provoke the limits of legitimate speech along the twin axes of truth-falsity and civility-incivility" (Udupa & Pohjonen, 2019, p. 3051).

This is particularly striking because, in the case of political incivility, the concept is deeply ingrained in the US legal and political debate on civil discourse. Moreover, as it is defined in the framework of deliberative theories, political incivility also relies heavily on values such as rationality (Massaro & Stryker, 2012, p. 379, p. 414). Regarding civility in general, it acts even as a further mechanism of discrimination, such as when the language used by members of historically oppressed groups use is classified as offensive language (Sap et al., 2019). This happens because the concept is supposed to be a high value of Western societies and intimately connected to social rank, class status, political hierarchy, and relations of power (cf. Harcourt, 2012). Hate speech, in turn, requires neither incivility nor irrationality, as it constitutes a matter of discrimination.

8 Why working on the concept of hate speech?

This chapter sheds light on the conceptual change the term "hate speech" has experienced, what role academic research has played in this, and what problems the change causes.

Hate speech is a theoretically sound scientific concept with a clear intention: it is anchored in three defining properties (DP), which work as criteria to disambiguate it: attacks (DP1) based on an identity factor (DP2) and that are symbolic in nature (DP3). Further, hate speech is a matter of communication in public life.

The clear intension of the concept provides a first approach to determining what is not hate speech. Yet, the concept is also highly abstract, which makes it difficult to identify to which class of objects it applies. Not having a clear extension lends it flexibility but also poses a challenge to empirical research.

However, by trying to overcome conceptual challenges in empirical research, digital media research has been creating new conceptual issues, such as: a) concept stretching—applying the concept to cases that do not match the defining properties of hate speech; b) concept shrinking—reducing the problem to a matter of content, as in the case of lexicon-based approaches; and c) an inflation of concepts—using the term interchangeably with "online hate" or replacing it with new terms, which creates its own conceptual issues.

This is problematic because concepts are not merely a matter of abstract discussions among academics. Concepts constitute a symbolic resource for defining problems and determining how they are going to be tackled. As a consequence, erasing the defining properties of hate speech creates many political issues. Transforming hate speech into a matter of family resemblance means that the concept can be "adapted" to any context, opening the door to all kinds of political instrumentalization. Applying the term "hate speech" and other forms of online abuse interchangeably downplays the problem as merely a matter of bad behavior among users in online conversations.

Replacing the concept of hate speech with that of online hate erases discrimination and power asymmetry from digital media research as the roots of this specific but highly harmful kind of communication.

So, what is the purpose of the concept of hate speech? Why should digital media researchers work on a concept that raises so many conceptual issues? First, of all the concepts applied to analyze threats in digital communication, hate speech is

probably the one with the longest research tradition. Many issues being discussed in the field of platform governance, for instance, have been analyzed for decades in hate speech research. Second, in spite of vagueness with regard to which cases the concept can be applied to, the concept is unambiguous: there is a broad consensus among different social actors about what hate speech means. Third, hate speech is a much more severe digital threat than insulting people on social media.

By abandoning, making ambiguous, or shrinking the concept, digital media research is jeopardizing its potential to tackle one of the most socially relevant problems in its field.

Liriam Sponholz is a postdoctoral researcher at the German Centre for Integration and Migration Research (DeZIM) in Berlin, Germany. https://orcid.org/0000-0001-7875-4273

References

Ali, S., Saeed, M. H., Aldreabi, E., Blackburn, J., De Cristofaro, E., Zannettou, S., & Stringhini, G. (2021). Understanding the effect of deplatforming on social networks. *WebSci '21*, June 21–25, Virtual Event, United Kingdom. https://seclab.bu.edu/people/gianluca/papers/deplatforming-websci2021.pdf

Alltagsrassismus. Angespuckt und beschimpft: Junge Muslima wird auf offener Straße angegriffen [Everyday racism. Spat on and verbally abused: Young Muslim women are attacked on the street]. (2019, April 3). *Stern*. https://www.stern.de/neon/wilde-welt/gesellschaft/alltagsrassismus--junge-muslima-wird-in-wien-auf-der-strasse-angespuckt-8650518.html

Amjahid, M. (2021, April 29). Neuköllner über Alltagsrassismus: "Nur Zufall, dass es bei Aldi war" [Neuköllner on everyday racism: "It was just chance that it was at Aldi"]. *Taz*. https://taz.de/Neukoellner-ueber-Alltagsrassismus/!5762911/

Art, D. (2020). The myth of global populism. *Perspectives on Politics*, 1–13. https://doi.org/10.1017/S1537592720003552

Bahador, B., & Kerchner, D. (2019). *Monitoring hate speech in the US media* (Working Paper). The George Washington University. https://cpb-us-e1.wpmucdn.com/blogs.gwu.edu/dist/8/846/files/2019/03/Monitoring-Hate-Speech-in-the-US-Media-3_22-z0h5kk.pdf

Benesch, S. (2004). Inciting genocide, pleading free speech. *World Policy Journal,* 21(2), 62–69.

Benesch, S. (2014). Defining and diminishing hate speech. In P. Grant (Ed.), *State of the world's minorities and indigenous peoples 2014* (pp. 19–25). http://minorityrights.org/wp-content/uploads/old-site-downloads/mrg-state-of-the-worlds-minorities-2014-chapter02.pdf

Bickert, M. (2020, October 12). *Removing Holocaust denial content. Facebook.* https://about.fb.com/news/2020/10/removing-holocaust-denial-content/

Brändlin, A.-S. (2016, February 26). Facebook's Zuckerberg to tackle hate speech. *Deutsche Welle*. https://www.dw.com/en/facebooks-zuckerberg-to-stamp-out-hate-speech-in-germany/a-19078185

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

Brown, A. (2017b). What is hate speech? Part 2: Family resemblances. *Law and Philosophy*, *36*(5), 561–613.

Chaudhry, I. (2015). #Hashtagging hate: Using Twitter to track racism online. *First Monday*, *20*(2). https://doi.org/10.5210/fm.v20i2.5450

Cobb, R. W., & Elder, C. D. (1972). *Participation in American politics: The dynamics of agenda building*. Allyn and Bacon.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Collier, D., & Mahon, J. E. (1993). Conceptual "stretching" revisited: Adapting categories in comparative analysis. *The American Political Science Review*, *87*(4), 845–855. https://doi.org/10.2307/2938818

Committee on the Elimination of Racial Discrimination. (2013, September 26). *General recommendation No. 35*. http://tbinternet.ohchr.org/_layouts/treatybodyexternal/Download.aspx?symbolno=CERD/C/GC/35&Lang=en

Delgado, R., & Stefancic, J. (2004). *Understanding words that wound*. Westview Press.

Facebook. (2021). *Hate speech*. https://www.facebook.com/communitystandards/recentupdates/hate_speech

Froio, C., & Ganesh, B. (2019). The transnationalisation of far right discourse on Twitter. Issues and actors that cross borders in Western European democracies. *European Societies*, *21*(4), 513–539. https://doi.org/10.1080/14616696.2018.1494295

Gagliardone, I., Gal, D., Alves, T., & Martínez, G. (2015). *Countering online hate speech*. UNESCO Series on Internet Freedom. http://unesdoc.unesco.org/images/0023/002332/233231e.pdf

Gelber, K. (2021). Differentiating hate speech: A systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, *24*(4), 393–414. https://doi.org/10.1080/13698230.2019.1576006

Gerstenfeld, P. B., Grant, D. R., & Chiang, C.-P. (2003). Hate online: A content analysis of extremist internet sites. *Analyses of Social Issues and Public Policy*, *3*(1), 29–44. https://doi.org/10.1111/j.1530-2415.2003.00013.x

Harcourt, B. (2012). The politics of incivility. *Arizona Law Review*, *54*(2), 345–374. https://scholarship.law.columbia.edu/faculty_scholarship/638

Hare, I., & Weinstein, J. (Eds.). (2009). *Extreme speech and democracy*. Oxford University Press.

Johnson, N. F., Leahy, R., Restrepo, N. J., Velasquez, N., Zheng, M., Manrique, P., Devkota, P., & Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, *573*(7773), 261–265. https://doi.org/10.1038/s41586-019-1494-7

Kleinberg, B., van der Vegt, I., & Gill, P. (2021). The temporal evolution of a far-right forum. *Journal of Computational Social Science*, *4*, 1-23. https://doi.org/10.1007/S42001-020-00064-X

Langlois, G., & Elmer, G. (2013). The research politics of social media platforms. *Culture Machine*, *14*, 1–17. https://culturemachine.net/wp-content/uploads/2019/05/505-1170-1-PB.pdf

Lawrence III, C. R. (1993). If he hollers let him go: Regulating racist speech on campus. In M. J. Matsuda, C. R. Lawrence III, R. Delgado, & K. W. Crenshaw (Eds.), *Words that wound: Critical race theory, assaultive speech, and the First Amendment* (pp. 53–88). Westview Press.

Levin, S., & Solon, O. (2018, July 18). Zuckerberg defends Facebook users' right to be wrong – even Holocaust deniers. *The Guardian*. https://www.theguardian.com/technology/2018/jul/18/zuckerberg-facebook-holocaust-deniers-censorship

Lewandowski, T. (1994). *Linguistisches Wörterbuch* [Linguistic dictionary]. UTB für Wissenschaft.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE*, 14(8), Article e0221152. https://doi.org/10.1371/journal.pone.0221152

Marková, I. (2003). *Dialogicality and social representations: The dynamics of mind*. Cambridge University Press.

Marsteintredet, L., & Malamud, A. (2020). Coup with adjectives: Conceptual stretching or innovation in comparative research? *Political Studies*, *68*(4), 1014–1035. https://doi.org/10.1177/0032321719888857

Massaro, T. M., & Stryker, R. (2012). Freedom of speech, liberal democracy, and emerging evidence on civility and effective democratic engagement. *Arizona Law Review*, *54*(2), 375–441. https://arizonalawreview.org/pdf/54-2/54arizlrev375.pdf

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, *22*(2), 205–224. https://doi.org/10.1177%2F1527476420982230

Matsuda, M. J. (1989). Public response to racist speech: Considering the victim's story. *Michigan Law Review*, 87(8), 2320–2381. https://doi.org/10.2307/1289306

Meddaugh, P. M., & Kay, J. (2009). Hate speech or "reasonable racism?" The other in Stormfront. *Journal of Mass Media Ethics*, *24*(4), 251–268. https://doi.org/10.1080/08900520903320936

Obermaier, M., Hofbauer, M., & Reinemann, C. (2018). Journalists as targets of hate speech: How German journalists perceive the consequences for themselves and how they cope with it. *SCM Studies in Communication and Media*, *7*(4), 499–524. https://doi.org/10.5771/2192-4007-2018-4-499

Palonen, K. (1999). Rhetorical and temporal perspectives on conceptual change. *Redescriptions: Political Thought, Conceptual History and Feminist Theory*, *3*(1), 41–59. http://doi.org/10.7227/R.3.1.4

Paz, M. A., Montero-Díaz, J., & Moreno-Delgado, A. (2020). Hate speech: A systematized review. *SAGE Open*, *10*(4). https://doi.org/10.1177/2158244020973022

Poole, E., Giraud, E. H., & de Quincey, E. (2021). Tactical interventions in online hate speech: The case of #stopIslam. *New Media & Society*, *23*(6), 1415–1442. https://doi.org/10.1177/1461444820903319

Potthof, M. (2017). Probleme von Begriffsbildung und -verwendung in der Kommunikationswissenschaft [Problems of concept formation and use in communication science]. *Studies in Communication | Media*, *6*(2), 95–127. https://doi.org/10.5771/2192-4007-2017-2-95

Sartori, G. (Ed.). (1984). *Social science concepts: A systematic analysis*. Sage.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019, July). The risk of racial bias in hate speech detection, in *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1668–1678), https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Spain, 1–10, https://www.aclweb.org/anthology/W17-1101.pdf

Searle, J. R. (1980). The intentionality of intention and action. *Cognitive Science*, *4*(1), 47–70. https://doi.org/10.1080/00201747908601876

Sellars, A. (2016). *Defining Hate Speech* (December 1, 2016). Berkman Klein Center Research Publication No. 2016-20, Boston Univ. School of Law. http://dx.doi.org/10.2139/ssrn.2882244

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). *Analyzing the targets of hate in online social media*. https://arxiv.org/pdf/1603.07709.pdf

Sponholz, L. (2018). *Hate Speech in den Massenmedien: Theoretische Grundlagen und empirische Umsetzung* [Hate speech in the mass media: Theoretical foundations and empirical implementation]. Springer Verlag. https://doi.org/10.1007/978-3-658-15077-8

Sponholz, L. (2019a). Hate Speech in Sozialen Medien: Motor der Eskalation? [Hate speech: Social media as trigger?]. In H. Friese, M. Nolden, & M. Schreiter (Eds.), *Rassismus im Alltag: Theoretische und empirische Perspektiven nach Chemnitz* [Racism in everyday life: Theoretical and empirical perspectives after Chemnitz] (pp. 157–178). transcript Verlag. https://doi.org/10.14361/9783839448212-009

Sponholz, L. (2019b). Hate Speech: Viel mehr als böse Wörter [Hate speech: Much more than bad words]. In E. Greif & S. Ulrich (Eds.), *Hass im Netz: Grenzen digitaler Freiheit. Linzer Schriften zu Gender und Recht*, 63 [Online hate: Limits to digital freedom. Linz writings on gender and law, 63] (pp. 1–30). Trauner Verlag.

Sponholz, L. (2020). Der Begriff „Hate Speech" in der deutschsprachigen Forschung. Eine empirische Begriffsanalyse [The term "hate speech" in German-language research: An empirical concept analysis]. *SWS-Rundschau*, *60*(1), 43–65.

Sponholz, L. (2021). Hass mit Likes: Hate Speech als Kommunikationsform in den Social Media [Hate with likes: Hate speech as communication in social media]. In S. Wachs, B. Koch-Priewe, & A. Zick (Eds.), *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen* [Hate speech – Multidisciplinary analysis and options for action] (pp. 15–37). Springer VS. https://doi.org/10.1007/978-3-658-31793-5_2

Stone, G. R. (2000). First amendment. In L. W. Levy, K. L. Karst, & A. Winkler (Eds.), *Encyclopedia of the American Constitution* (pp. 1055–1057). Macmillan.

Strache sieht in "FPÖ-Hass" Beleg für eigene Relevanz [Strache sees hate from "FPÖ" as evidence of own relevance]. (2015, July 14). *Standard.at.* https://www.derstandard.at/story/2000019117104/strache-sieht-in-fpoe-hass-beleg-fuer-eigene-relevanz

Straus, S. (2007). What is the relationship between hate radio and violence? Rethinking Rwanda's "Radio Machete". *Politics & Society*, *35*(4), 609–637. https://doi.org/10.1177/0032329207308181

Tetrault, J. E. C. (2019). What's hate got to do with it? Right-wing movements and the hate stereotype. *Current Sociology*, *69*(1), 3–23. https://doi.org/10.1177%2F0011392119842257

Thielmann, W. (2004) Begriffe als Handlungspotentiale [Concepts as potential for action]. In: *Linguistische Berichte*, *199*, 287–312.

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, *126*(1), 157–179. https://doi.org/10.1007/s11192-020-03737-6

Twitter. (2020, December 2). *Updating our rules against hateful conduct*. https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html

Udupa, S., & Pohjonen, M. (2019). Extreme speech and global digital cultures: Introduction. *International Journal of Communication*, *13*, 3049–3067. https://ijoc.org/index.php/ijoc/article/view/9102/2710

United Nations. (2020). *United Nations strategy and plan of action on hate speech: Detailed Guidance on Implementation for United Nations Field Presences*. https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf

van Dijck, J., & Poell, T. (2013). Understanding social media logic. *Media and Communication*, *1*(1), 2–14. http://dx.doi.org/10.17645/mac.v1i1.70

Waqas, A., Salminen, J., Jung, S.-G., Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLoS One*, *14*(9), Article e0222194. https://doi.org/10.1371/journal.pone.0222194

Weber, A. (2009). *Manual on hate speech*. Council of Europe Publ.

Wennerberg, H. (1998). Der Begriff der Familienähnlichkeit in Wittgensteins Spätphilosophie [The concept of family resemblance in Wittgenstein's late philosophy]. In E. v. Savigny (Ed.), *Ludwig Wittgenstein* (pp. 41–69). De Gruyter.