Stanford Encyclopedia of Philosophy Archive
Fall 2023 Edition

Hate Speech

First published Tue Jan 25, 2022

Hate speech is a concept that many people find intuitively easy to grasp, while at the same time many others deny it is even a coherent concept. A majority of developed, democratic nations have enacted hate speech legislation—with the contemporary United States being a notable outlier—and so implicitly maintain that it is coherent, and that its conceptual lines can be drawn distinctly enough. Nonetheless, the concept of hate speech does indeed raise many difficult questions: What does the 'hate' in hate speech refer to? Can hate speech be directed at dominant groups, or is it by definition targeted at oppressed or marginalized communities? Is hate speech always 'speech'? What is the harm or harms of hate speech? And, perhaps most challenging of all, what can or should be done to counteract hate speech?

In part because of these complexities, hate speech has spawned a vast and interdisciplinary literature. Legal scholars, philosophers, sociologists, anthropologists, political theorists, historians, and other academics have each approached the topic with exceeding interest. In this current article, however, we cannot hope to cover how these many disciplines have engaged with the concept of hate speech. Here, we will focus most explicitly on how hate speech has been taken up within philosophy, with particular emphasis on issues such as: how to define hate speech; what are the plausible harms of hate speech; how an account of hate speech might include both overt expressions of hate (e.g., the vitriolic use of slurs) as well as more covert, implicit utterances (e.g., dogwhistles); the relationship between hate speech and silencing; and what might we do to counteract hate speech.

- 1. What is Hate Speech?
  - 1.1 The Harms of Hate Speech
- 2. Religious Hatred and Anti-Semitism
- 3. Slurs, Code Words, and Dogwhistles
  - 3.1 Slurs
  - 3.2 Dogwhistles and Coded Language
- 4. Pornography, Hate Speech, and Silencing
- 5. Counteracting Hate Speech
  - 5.1 The Case for Bans
  - 5.2 Objections to Bans, and Some Responses
  - 5.3 The Supported Counterspeech Alternative
- Bibliography
- Academic Tools
- Other Internet Resources
- Related Entries

1. What is Hate Speech?

The term 'hate speech' is more than a descriptive concept used to identify a specific class of expressions. It also functions as an evaluative term judging its referent negatively and as a candidate for censure. Thus, defining this category carries serious implications. What is it that designates hate speech as a distinctive class of speech? Some claim the term 'hate speech' itself is misleading because it wrongly suggests "virulent dislike of a person for any reason" as a defining feature (Gelber 2017, 619). That is not, however, the way in which the term is understood among most legal theorists and philosophers. Perhaps it would be useful to start with some examples.

Bhikhu Parekh (2012) lists the following instances as examples different countries have either punished or sought to punish as hate speech:

- Shouting "[N-words] go home," making monkey noises, and chanting racist slogans at soccer matches.
- "Islam out of Britain. Protect the British people."
- "Arabs out of France."
- "Serve your country, burn down a mosque."
- "Blacks are inherently inferior, lecherous, predisposed to criminal activities, and should not be allowed to move into respectable areas."
- "Jews are conspiratorial, devious, treacherous, sadistic, child killers, and subversive; want to take over the country; and should be carefully watched."
- Distribution by a political party of leaflets addressed to "white fellow citizens" saying that, if it came to power, it would remove all Surinamese, Turks, and other "undesired aliens" from the Netherlands.
- A poster of a woman in a *burka* with text that reads: "Who knows what they have under their sinister and ugly looking clothes: stolen goods, guns, bombs even?"
- Speech that either denies or trivializes the holocaust or other crimes against humanity.

Robert Post's four bases for defining hate speech might help us organize the features of Parekh's list:

> In law, we have to define hate speech carefully to designate the forms of the speech that will receive distinctive legal treatment. This is no easy task. Roughly speaking, we can define hate speech in terms of the harms it will cause—physical contingent harms like violence or discrimination; or we can define hate speech in terms of its intrinsic properties—the kinds of words it uses; or we can define hate speech in terms of its connection to principles of dignity; or we can define hate speech in terms of the ideas it conveys. Each of these definitions has advantages and disadvantages. Each intersects with the first amendment theory in a different way. In the end, any definition that we adopt must be justified on the ground that it will achieve the results we wish to achieve. (Herz and Molnar 2012, 31)

The four definitional bases are in terms of: (1) harm, (2) content, (3) intrinsic properties, i.e., the type of words used, and (4) dignity. One could also attempt a hybrid definition by combining the ways mentioned. But, as is made clear in Post's remarks, definitions of this sort are relative to the interests of the definer; "We must evaluate the status of 'hate speech' so defined in order to determine whether it achieves what we wish to accomplish and whether the harms of the definition will outweigh its advantages" (Herz and Molnar 2012, 31). The upshot is a rejection of a univocal definition that captures "the essence" of hate speech as a phenomenon.

It is important to note that many definitions of hate speech will not fall squarely within the categories Post outlines. For instance, the UN's *International Convention on the Elimination of All Forms of Racial Discrimination* identifies hate speech both in terms of its content and its harmful consequences. Most definitions tend to characterize hate speech in multiple ways.

Harm-based definitions conceive of hate speech in terms of the harms to which targets are subjected. Things like discrimination or linguistic violence are candidates, though some (Gelber, 2017) argue that hate speech can harm one's ability to participate in democratic deliberation. Susan Brison (1998a) offers a disjunctive definition that centers on a kind of abuse to targets. She defines hate speech as "speech that vilifies individuals or groups on the basis of such characteristics as race, sex, ethnicity, religion, and sexual orientation, which (1) constitutes face-to-face vilification, (2) creates a hostile or intimidating environment, or (3) is a kind of group libel" (313). 'Harm' as used by Brison refers to what Joel Feinberg describes "as a wrongful setback to (or invasion of) someone's interests" (Brison, 1998b, 42).

Perhaps an immediate reaction to disjunctive definitions of the sort Brison offers is skepticism about the definitiveness

of the purported list. When we go to test the definition's application, we invariably find contestable inclusions and exclusions. Recall the examples from Parekh at the start of this section. Something like "Arabs out of France" might be included as an instance of hate speech on Brison's account on the grounds that it creates a hostile or intimidating environment. Should statements that communicate a similar message in a less abrasive manner also be included? Suppose "Only French Nationals should occupy France" is roughly equivalent content-wise to "Arabs out of France." If the former is indeed a less abrasive presentation though communicating the same content as the latter, what are we to make of its status? Many will find the statement odious; many will not. And since it is certainly not a face-to-face vilification or form of group libel, classifying it as hate speech will depend on how likely it is to create an intimidating or hostile environment.

The previous objection might entice one to opt for a content-based view. Content-based views define hate speech as that which "expresses, encourages, stirs up, or incites *hatred* against a group of individuals distinguished by a particular feature or set of features such as race, ethnicity, gender, religion, nationality, and sexual orientation" (Parekh, 2012, 40). This version makes it easier to conceive of semantically equivalent statements that differ in manner of presentation as instances of hate speech.

Content-based accounts face the challenge of determining which contents meet this standard. If the content that distinguishes hate speech from other types of speech must express, encourage, or incite hatred towards groups or individuals based on certain features, then the proponent of this view will need an account of expression. Is the speech in view that which signals the presence of a particular mental state in the speaker (i.e., hate) or that which is likely to prime feelings of animosity in a specific audience?

Another issue facing content-based approaches concerns distinguishing between speech that "respects 'the decencies of controversy'" and that "which is outrageous and therefore hate inducing" (Post, 2009, 128). The ability to express a wide range of views, even contentious ones, is a cherished aspect of democratic societies. Failure to observe this distinction would broaden the scope of what counts as hate speech perhaps too much. In order to make this distinction, one could follow Post in tying it to "ambient social norms" that distinguish outrageous and respectful behavior. One challenge though is in determining the content of those social norms. For instance, a minority group whose opinions have little impact on the makeup of norms are unjustifiably excluded from influencing the shape of their society's civility norms.

Definitions of hate speech based on intrinsic properties generally refer to those that emphasize the type of the speech uttered. What is at issue is the use of speech widely known to instigate offense or insult among a majority of society. Explicitly derogatory expressions like slurs are paradigmatic examples of this type of view. In general, the type of speech identified on this account is inherently derogatory, discriminatory, or vilifying.

Though attractive at first glance, classifying hate speech along these lines might prove to fall short in two ways. First, defining hate speech in this way might be too constricting. Some of the examples in our initial list would seem not to count as hate speech since they arguably lack the intrinsic features. "Arabs out of France," for example, does not contain explicitly slurring terms. And second, this definition might prove too expansive. In cases where slurs are reappropriated by members of the target group or where artists incorporate them into a creative work, it would appear odd to count these as instances of hate speech. The concern is tied specifically to locating the issue in the terms themselves, as opposed to the use to which the terms are put.

Perhaps a final challenge to intrinsic property views can be derived from the work of Judith Butler (1997). On Butler's account, hate speech is a kind of performative that is "always delivered twice-removed, that is, through a *theory of the speech act* that has its own performative power" (96). More specifically, "[w]hat hate speech does … is to constitute the subject in a subordinate position" (19). Butler locates the trouble with hate speech in its *perlocutionary* effects, a concept introduced by J.L. Austin that refers to the effects a speech act can have on its audience. An example of a perlocutionary effect is feeling amused at a joke or frightened from the telling of a ghost story. Unlike with intrinsic property definitions, Butler shifts focus to the nature of the acts performed rather than the terms in use. (For a critical look at Butler's account, see Schwartzman (2002).)

Lastly, dignity-based conceptions focus primarily on the role of harms to the dignity of targets of hate speech. For instance, both Steven Heyman (2008) and Jeremy Waldron (2014) appeal to dignity in their accounts. Broadly speaking, hate speech on this kind of conception amounts to speech that undermines its target's "basic social standing, the basis of [their] recognition as social equals and as bearers of human rights and constitutional entitlements" (Waldron, 2014, 59). This conception of hate speech will also include characterizations in terms of group defamation or group libel. Section 130 of Germany's penal code is an example of legislation that incorporates a dignity-based conception of hate speech, prohibiting "attacks on human dignity by insulting, maliciously maligning, or *defaming*

part of the population" (see Waldron, 2014, 8).

Worries about application follow dignity-based conceptions as well. Firstly, there may be questions about how we, in particular instances, are to distinguish between false statements about a group as a whole and those about a particular member of a group (Brown, 2017a). Presumably, only the former is consistent with an understanding of hate speech as a group-based phenomenon. Secondly, an implication of the view appears to be that it expands the range of things that would count as hate speech. Any speech that calls into question the basic standing of certain groups falls under this notion, which may make it more difficult to distinguish between contentious political speech and hate speech.

Perhaps a lesson to draw from the profusion of disjunctive definitions is a general skepticism about a definitive description of hate speech. We might concur with Alexander Brown that 'hate speech' is an equivocal term denoting a family of meanings (Brown 2017b, 562). According to Brown, 'hate speech' isn't just a term with contested meanings, but rather, it is "systematically ambiguous; which is to say, it carries a multiplicity of different meanings" (2017b, 564). Because the expression is what is typically referred to as an essentially contested term, the hunt for a univocal or universal definition is futile.

1.1 The Harms of Hate Speech

The harms that have been attributed to hate speech comprise a long and varied list, ranging from the immediate psychological harms experienced in the moment by the person(s) targeted by an instance of hate speech, to much more long-term impacts that affect not only those targeted but whole communities, and even the strength of an entire nation.

A distinction between "assaultive hate speech" and "propagandistic hate speech" is helpful when discussing these harms (Langton 2012; 2018a; see also Gelber and McNamara (2016) who discuss "face-to-face encounters" and "incidences of general circulation"). Hate speech yelled at an individual on the street, or from a passing car, is a face-to-face encounter, and an assaultive speech act. This is, moreover, most often inter-group hate speech, where the speaker(s) are, for example, white, and the targets are non-white. On the other hand, *propagandistic* hate speech is often intra-group speech, spoken by members of one group to fellow ingroup members (e.g., a white person to other white people). The newsletter of the KKK, therefore, would fit into this category.

While this distinction is helpful to keep in mind, it should also not be overstated. Summarizing the results of their study which surveyed the experiences of the victims of hate speech, Katharine Gelber and Luke McNamara conclude that "the distinction between face-to-face encounters and general circulation hate speech is not always clear in the everyday experiences of racism endured by targets" (2016, 326). Any one instance of hate speech might fall into both categories. For example, it may occur in its first instance as an assaultive speech act, and then reports of the event may then take on a propagandistic aspect, as it is spread among the community. Similarly, even if an instance of hate speech is intended as a piece of propaganda, it may, when encountered by a member of the community it disparages, be akin to assaultive speech.

Still, this distinction helps reveal the wide range of the *types of speech acts* that are plausibly harmful, and also offers insight into *how* they harm. For example, Waldron (2014) focuses mainly on hate speech in its propagandistic mode, which he argues undermines the public assurance of equal social standing that members of non-dominant communities are entitled to—in his terms, their assurance of dignity. On this view, public hate speech—e.g., flyers that read 'Muslims Out!'—is "an environmental threat to social peace, a sort of slow-acting poison, accumulating here and there, word by word" (2014, 4). Its harm is therefore one that attacks the broader society, and not just individuals targeted by hate speech.

On the other hand, the essays in the classic *Words that Wound* tend to focus more on what its authors term "assaultive speech," that is, "words that are used as weapons to ambush, terrorize, wound, humiliate, and degrade" (Matsuda et al. 1993, 1). This leads them to focus more on hate speech's ability to produce "direct, immediate, and substantial injury" (Lawrence, 1993, 57), such as "immediate mental or emotional distress" (Delgado, 1993, 93–94). On this approach, the most evident harms of hate speech are psychological. These psychological injuries scale up, however, when hate speech is endemic, and so result in the types of community or social harms highlighted by authors like Waldron. For this reason, the distinction between these approaches may be thought of as more a matter of emphasis.

This relationship between individual harms and broader social harms is also evident once we acknowledge the long-term effects of hate speech on victims, in addition to its more immediate impacts (Delgado and Stefancic, 2004, 14). Victims of hate speech may first experience "psychological symptoms and emotional distress" like heightened stress and fear in the immediate aftermath of assaultive hate speech, but they may also experience far-ranging consequences if they "modify their behavior and demeanor" to avoid receiving further hate messages, limiting their ability to

participate fully in society (Matsuda, 1993, 24). Gelber and McNamara's interview subjects confirm this complex web of effects that hate speech may cause, highlighting how "harms are often enduring and not ephemeral" (2016, 336). In this way, hate speech is both an immediate attack on one's health and dignity, along with a threat to their community's position in society. The cumulative effect of hate speech events, therefore, is a collection of harms located both in individuals and communities, which blurs the distinction between assaultive and propagandistic hate speech events.

Constitutive and Consequential Harms

Another distinction which is similarly helpful, but also fraught, is the distinction between *constitutive* and *consequential* harms—that is, harms that occur *in* the saying of some utterance of hate speech, and those that are its downstream results (see Maitra and McGowan, 2012, 6). This distinction draws on the speech act theory of J.L. Austin (1962) and has served an important role in the examination of hate speech from feminist philosophers of language (see, e.g., Langton, 1993; 2012; Maitra and McGowan, 2012; Maitra, 2012; McGowan, 2004; 2009; 2012; 2019; and others). Constitutive harms are those that correspond to what Austin called the *illocutionary act*, the act performed *in saying X*, while consequential harms correspond to *perlocutionary effects*, the results brought about *by saying X*. Most (though not all) of the harms surveyed above comprise consequential harms, as items such as psychological injury, feelings of fear, and societal withdrawal all most naturally fall into the perlocutionary effects category.

However, philosophers have also drawn attention to how hate speech can injure in a different way by indirectly affecting the positions of the social groups targeted by hate in a social hierarchy. That is, "by fixing facts about the distribution of social power, including facts about who has this power, and who lacks it" hate speech harms in a way not captured in the above account of individual injuries and their cumulative effects (Maitra and McGowan, 2012, 7). This is an *immediate* harm that occurs in the saying of the speech act, which (given appropriate circumstances and uptake) produces a shift in the normative landscape. It is in this way that an instance of hate speech may not only cause the injuries surveyed above but may also, for example, *rank* Indigenous Peoples as inferior, *legitimate* discriminatory behavior towards them (perhaps via incitement), or potentially *silence* them. (We return to the notion of silencing as an illocutionary harm of hate speech in Section 4 below.)

One reason to direct our attention towards the constitutive harms of hate speech is its potential to productively advance the debate over the legitimacy of potential restrictions. Mary Kate McGowan (2009) has made this case most explicitly. "Rather than focus on what a certain category of speech causes," she writes, we ought to be "interested in what such speech actually does, in and of itself" (2009, 389–90). The idea here is that by focusing only on the harms *caused* by hate speech, we are inevitably drawn into a debate about balancing the costs and benefits of permitting or regulating speech, which often leads to an impasse. Alternatively, turning our attention to the acts hate speech constitutes can reveal features that help us avoid question of balancing harms, and opens the door to regulation. On this approach, some instances of hate speech can be seen to constitute acts of (verbal) discrimination, and should be considered analogous to other acts of discrimination—like posting a 'Whites Only' sign up at a hotel—that US law recognizes as illegal. As a speech act, hate speech can enact discriminatory rules in much the same way the physical sign does, and so ought to similarly be restricted (McGowan, 2012). This argument proceeds by a development of Austin's notion of "exercitives," which are speech acts that enact rules in a given domain, and is one example of the fruitful use of speech act theory to the philosophy of hate speech.

At the same time, however, it's worth acknowledging that the distinction that this analysis relies on—between illocutionary acts and perlocutionary effects—is one that some argue is untenable (for one example, see Kukla, 2014). As illocutionary acts are indeterminate or incomplete without some form of audience uptake, it is difficult to articulate precisely how we ought to distinguish a speech act's effects from its inherent qualities. Furthermore, the testimonials of victims of hate speech "suggests that there is a close and complex relationship between constitutive and consequential harms, and the harms are experienced cumulatively" (Gelber and McNamara 2016, 336–37). As such, any attempt to draw too neat of a distinction between these two types of harm risks misrepresenting victims' experiences, and might tie the attempt to restrict hate speech unhelpfully to a philosophically contested distinction.

As a result, some caution must be applied when marking too stark of a contrast between these harms. Much like the distinction between assaultive and propagandistic hate speech, then, we can consider the distinction between consequential and constitutive harms to be analytically helpful in exploring the variety of harms attributable to hate speech, while recognizing that it is at the same time an abstraction from the on-the-ground realities of hate speech.

2. Religious Hatred and Anti-Semitism

Religious belief is sometimes the source of putative cases of hate speech, and sometimes its target. In both cases, assessing the conceptual addition of religion to hate speech is a difficult task. Speech rooted in religious conviction is sometimes subjected to scrutiny to determine whether instances should count as hate speech or not. For instance, the Westboro Baptist Church's demonstrations often make use of slurs and other explicitly defamatory language. This is an extreme case, which can be accommodated by extant hate speech legislation. Other cases, however, involve religious leaders making contentious statements—for instance, questioning the legitimacy or recognition of LGBT+ individuals, while claiming these are statements of love, not hate. Questions about religious speech of this sort concern whether it is simply contentious speech liberal democratic societies must tolerate or speech that runs afoul of deeply held norms that ought to be proscribed.

Some wonder whether religious sensibilities should be afforded special protection from offense. Amnon Reichman (2009), for instance, notes that some Israeli scholars have argued that providing special protection for religious beliefs is a good idea "so as not to push [religious] believers into having to choose between the authority of the state and the authority of their religion (namely, the authority of God)" (338). This relies on an assumption that religion is an institutionalized normative regime in competition with a legal regime where clashes over religious beliefs threaten the social fabric of society. It is in turn prudent to mitigate such clashes in order to avoid situations of unrest like the incidents involving comedic cartoons of Mohammed in the Dutch newspaper *Jyllands-Posten* and the French publication *Charlie Hebdo*.

It is not clear, however, that religious beliefs warrant special protection over other forms of belief that may be just as strongly held. Clashes over deeply held political beliefs can pose a similar threat to the social fabric as religious beliefs. Thus, there is no reason to think the same concern should not apply quite broadly. Providing certain types of speech special protection on these grounds would threaten to introduce quite repressive legislation on speech in general.

Holocaust denial, denial of the Armenian Genocide, and the denial of other crimes against humanity have also been the subject of special legislation, especially in Europe. As Michael Whine (2009) notes, 16 European states, as well as Israel, have criminalized Holocaust denial (543). In these contexts, at least one rationale for banning speech that denies or trivializes the Holocaust concerns its role in inciting hatred (Altman, 2012). One possible justification for such legislation rests on claims about what denial speech is. According to Martin Imbleau (2011), denial speech poses as an historical endeavor but is really propaganda. The denier's aim is to "eradicate the awareness of the truth that prevents the resurgence of past criminal ideologies" (2011, 238). But if this is the rationale, it potentially opens up justifications for much broader application since similar claims might be made of other forms of propaganda. (For a general overview on Holocaust denial, see Robert Wistrich (2012) and Behrens et. al (2017).)

3. Slurs, Code Words, and Dogwhistles

4. Pornography, Hate Speech, and Silencing

That hate speech and pornography are discussed so frequently together in philosophy might, at first glance, seem surprising. But given the overlap made in the arguments made by anti-porn feminist about pornography and anti-racist theorists about racist hate speech, the two are now intimately linked—for better or for worse. (One important fact that led to this development is, of course, the ruling that pornography is protected by the US first amendment as speech [see *Miller v. California* (1973)].) According to anti-porn feminists, much of what is said of racist hate speech and the harms that befall its targets also applies, with the appropriate changes, to pornography and women—including, it's worth emphasizing, women of color.

Many of the important initial moves in this literature were crafted by feminist legal scholar Catherine MacKinnon, along with Andrea Dworkin. One of MacKinnon's most significant claims that has received sustained philosophical attention is the idea that (degrading and misogynist) pornography *silences* women. With some modifications, a similar claim may be applied to hate speech, namely, that hate speech *silences* its targets. However, as the literature has focused on the case of pornography and women, it's worth examining these arguments in detail first.

This silencing argument begins with MacKinnon's observation that there are "words that set conditions" for other speech acts' successes or failures (1993, 63–68; see also Hornsby and Langton, 1998, 27). That is, there are some speech acts that fix the possibility of other speech acts. In other words, they make it possible for some persons to perform some speech acts, and make it impossible for others. This is most evident in formal settings, like a legislature, where the formal rules determine who may speak when, and in what manner. Pornography, the argument continues, does just this. It sets rules of behavior that, in effect, inhibit the speech of women. The result of which is that the speech acts of pornography—performed by those who produce and distribute it—create a climate that undermines women's capacity to perform certain speech acts of their own. The speech of some (pornographers), therefore, curtails the speech of others (women).

In an influential account of the phenomena of silencing, Langton (1993) deploys speech act theory to examine the case of sexual refusal. According to the silencing argument, pornography depicts women as not (genuinely) refusing sexual

advances with utterances of 'no.' Indeed, according to the myths perpetuated by pornography (among other social influences), a woman's 'no' is not a refusal, but rather part of an elaborate sexual script. As a result, when a woman says 'no' in a non-pornographic context, intending to refuse a man's sexual advances, she may find herself unable to be heard—that is, her words won't have the force and effect she intends, and her hearer will not take her to be refusing. She may find herself silenced in this particularly horrendous way, unable to use the standard methods of refusing another's sexual advances. The claim is that this occurs as a result of pornography silencing women's refusals in the context of sex. It renders their words powerless.

In making this argument, Langton relies on the distinction between locutionary, illocutionary, and perlocutionary acts, and, correspondingly, locutionary, illocutionary, and perlocutionary silencing. A couple examples will explain these distinctions quickly:

> When *X* says, 'Shoot him!' they are, we can see quite quickly, both saying something: 'shoot him!' and at the same time doing something: *ordering* the hearer to fire. In Austinian terms, we can say that *X* performs the *locutionary* act of making an utterance with a certain meaning, and at the same time is performing an *illocutionary* act of ordering the hearer to fire. In addition to these two things, the speaker is also, with their words, bringing about a number of effects, which Austin termed the '*perlocutionary* act.' In this case, leading to some unfortunate soul to be shot. (Adapted from Langton, 1993, 295, and Austin, 1962, 101)

To be clear, these three acts—locutionary, illocutionary, and perlocutionary—all occur as part of a single utterance and serve to bring out different aspects of any speech act. Austin (and many after him) paid particular attention to the illocutionary act of an utterance, as this, he said, corresponds to the *force* of an utterance. That is, what someone is doing with their words.

With this in mind, we can see that there are, in fact, many ways one could silence someone. You could literally gag or threaten someone to prevent them from speaking at all, which would achieve a type of locutionary silencing. Alternatively, you could let them say what they wish, recognize what act they are performing, but prevent them achieving their goals, and in doing so achieve a type of perlocutionary silencing. Finally, a third alternative occurs when one speaks and is prevented not only from achieving their intended effects, but also is prevented from performing the very action they intend to perform (Langton, 1993, 315). It is this third alternative—illocutionary silencing—that is said to occur when a man fails to even recognize a woman's 'no' as a refusal, owing at least partially to the influence of pornography.

The specific mechanics of silencing—along with the underlying theory that best explains the phenomena—is subject to much dispute in the literature, and numerous accounts with different essential features have been offered (see Langton (1993); Langton and West (1999); Hornsby (1994); Hornsby and Langton (1998); Maitra (2009); McGowan (2004, 2009, 2014); Mikolla (2011; 2019), among others).

Laura Caponetto (2021) distinguishes four different types of silencing, demonstrating the breadth of the concept. First, there is *essential* silencing, which consists in the hearer's failure to recognize the illocutionary point of a speech act. Second, there is *authority* silencing, where a hearer fails to acknowledge a speaker's authority in a relevant domain. Third, there is *sincerity* silencing, when the speaker's utterance is inaccurately taken as insincere. Fourth and finally, there is *seriousness* silencing, which consists in the hearer's failure to acknowledge the speaker's words as appropriately serious. Given these fine-grained ways of understanding silencing, a broad, comprehensive definition of silencing may be put as follows:

> *Illocutionary Silencing*
> A speaker *S* putting forth a speech act *A* addressed to a hearer *H* is illocutionarily silenced iff (i) *H* fails to recognize the obtaining of some conditions for *A*'s success; (ii) *S*'s attempt at *A*-ing meets the conditions that *H* fails to recognize; (iii) normal input and output conditions are met; (iv) the recognition failure on *H*'s part is systematic. (Caponetto, 2021)

In nearly all discussions of silencing, one common piece of contention concerns the notion of 'uptake.' On different understanding of what uptake consists in—ranging from the hearer's recognition of a speaker's intent, or the type of speech act being performed, up to the material consequences of a speech act—we are led to different conclusions about whether a speaker was silenced or not. Disagreement about the conditions of uptake poses difficulty, therefore, for many accounts of silencing. Drawing on these difficulties, Samia Hesni (2018) has argued that the standard account of silencing needs significant retooling, in part because the necessary distinction between illocutionary silencing and perlocutionary silencing cannot hold, as it relies on a problematic—and arguably conceptually untenable

—notion of uptake (Hesni, 2018, 957). In an attempt to avoid these difficulties, we might prefer an account of silencing that uses a Gricean, rather than Austinian or Searlian framework, bypassing the need to fully differentiate the illocutionary from the perlocutionary (see Maitra, 2009).

While much of this literature is explicitly focused on pornography's potential to silence women in the realm of sexual refusal, the notion that racist hate speech may similarly play a silencing function has also been put forward. For example, in a classic paper on the topic, Lawrence wrote that:

> Racist speech … distorts the marketplace of ideas by muting or devaluing the speech of Blacks and other despised minorities. Regardless of intrinsic value, their words and ideas become less saleable in the marketplace of ideas. An idea that would be embraced by large numbers of individuals if it were offered by a white individual will be rejected or given less credence if its author belongs to a group demeaned and stigmatized by racist beliefs. (Lawrence, 1993, 78–79)

Using the above framework, we might therefore say that racist hate speech can itself constitute words that set conditions for the success of other speech acts, and in doing so undermines the speech of its targets—and in some cases, effectively silencing them. That is, racist hate speech may, in cultivating an environment hostile to the voices (and lives) of many, can lead to both locutionary and illocutionary silencing in a way that threatens their freedom of expression. And as is noted above in the section on the harms of hate speech, one long-term consequence of racist hate speech may be the target's partial withdrawal from certain aspects of public life, including public discourse (West, 2012, 237). One further harmful effect of hate speech, then, is its targets' silence.

Another way in which racist hate speech might silence is more immediate. Returning to the distinction between propagandistic hate speech on the one hand, and assaultive hate speech on the other, where the latter consists in hate speech uttered directly to its target, we may note that hate speech often serves as a type of attack. So, despite the common refrain of 'more speech' offered as advice, conceiving hate speech as a personal attack demonstrates how it, in fact, threatens the speech rights of its targets. As Lawrence puts it: "The visceral emotional response to personal attack precludes speech" (1993, 68). He goes on:

> Attack produces an instinctive, defensive psychological reaction. Fear, rage, shock, and flight all interfere with any reasoned response. Words like 'nigger,' 'kike,' and 'faggot' produce physical symptoms that temporarily disable the victim, and perpetrators often use these words with the intention of producing this effect. (ibid.)

So, in both cultivating an environment in which the speech of marginalized groups is systematically devalued, or in serving as an immediate threat, hate speech can be said to silence its targets.

As is the case with pornography and silencing, the details of the mechanisms that sustain this type of silencing, along with what particular type of silencing racist hate speech results in, are subject to dispute. But, just like in the pornography debate, the plausibility of the silencing argument lies partly in how it reframes the overall question surrounding regulation. Rather than simply being a source of harm that merely infringes on the equality rights of its targets, if hate speech silences then it also infringes on the speech rights of its targets (West, 2012). As a result, it is not simply a question of balancing the speech rights of hate speakers against the wellbeing of their targets, but of competing claims to (substantive, and not just formal) freedom of expression. And given the importance that most liberal democracies place on freedom of expression, the challenge presented from hate speech is of central importance. For this reason, the silencing question is one of the most disputed aspects of hate speech and has generated great attention.

5. Counteracting Hate Speech

On the presumption that hate speech is harmful—both particularly harmful for the members of targeted groups, and also generally harmful to democracy—the natural question that follows is: what should we do about it? This question, however, rests on several sub-questions—some empirical, some conceptual—that themselves admit of rich dispute. For example, depending on how one conceives of the value and point of free expression—to better seek the truth, to respect autonomy, to ensure democracy, etc.—different answers to the hate speech question will seem more worthwhile than others. The same consideration applies to empirical matters as well, which are often difficult to properly assess in the absence of uncontroversial data. This means that relatively straightforward empirical questions—does genocidal speech pave the way to actual genocidal violence; do governments abuse hate speech regulation to punish political rivals and disfavored minorities; and others—rarely receive unanimous agreement. Despite these challenges, many theorists have addressed the question of how to counteract hate speech, and what form

that response ought to take.

We can divide the most common answers into three broad categories: (1) legally restrict it in some form, as a justified exception to free expression; (2) permit it on the basis of free expression, holding that the harms of censorship outweigh the harms of hate speech; or (3) permit it, but take explicit measures to undo the harm of hate speech.

5.1 The Case for Bans

First, the case for banning hate speech. While this position may be anathema to many (especially in the United States), it is the consensus position of most democratic nations around the globe, as well as the explicit position of the United Nations. In the *International Covenant on Civil and Political Rights*, Article 20 requires a ban on hate speech—or, in their words, "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law" (*Covenant on Civil and Political Rights*; see also, Article 4 of *Convention on Racial Discrimination*). It is worth noting the position of the United Nations and other democracies on hate speech in part because of the contrast they serve for the dominant position in the United States, which recognizes some exceptions to the right to free expression (e.g., obscenity, libel, child sexual abuse material), but not generally on the basis of (racial) hate. Moreover, these exhortations to criminalize hate speech from the United Nations sit alongside commitments that maintain the importance of freedom of expression. For instance, Article 20, quoted above, is immediately preceded by Article 19, which affirms right to freedom of expression (*Covenant on Civil and Political Rights*).

The standard justification offered for restrictions on freedom of expression are based on the necessity of (a) respect of the rights or reputations of others; and (b) reasons of national security or of public order. In other words, a ban on hate speech may be thought to follow from the recognition of the harms it presents, both to the dignity of minority-members of a nation, as well as their physical safety. This position maintains, then, that restrictions on hate speech are a legitimate—and necessary—exception to an otherwise wider understanding of free expression. (For some theorists, it's worth noting, hate speech is best not understood as the type of speech that free speech protections are meant to include—e.g., it serves no purpose in the pursuit of truth—and so is not in fact an *exception* to a free speech principle, but simply not included in a proper understanding of the scope of free speech.) This view naturally follows from the understanding that multiple values and rights must be balanced against each other. This is true both of countries that explicitly prohibit hate speech in order to protect minority rights, as well as in more 'speech-friendly' nations like the United States, where speech that is aimed at and likely to result in "imminent lawless action" may be legitimately restricted (see *Brandenburg v. Ohio*).

However, most advocates for legally restricting hate speech believe that its proper scope is wider than what US law currently allows. Parekh, for example, rejects the position that hate speech may only be restricted when there is "imminent danger" of violence on the grounds that this understanding is too short-sighted. Moreover, he says,

> no action occurs in a historical vacuum, and every action produces consequences not inherently but against a particular background. … Imminent danger occurs against, and is imminent because of, the prevailing social climate, and consistency demands that we concentrate our efforts not only on fighting the immediate source of danger, but also on changing the climate. (2012, 45–46)

On the understanding that the threat of hate speech is not exhausted by cases that concern "imminent danger," we might then ground the prohibition of hate speech on the basis that this may reduce speech that causes harm to its targets, beyond those most immediately affected. Of course, there is also an important role to be played by non-legal means (e.g., moral and social pressure) in erasing or reducing these harms, so legal bans are best understood as part of a broader approach to the ills of hate speech. Furthermore, advocates of bans describe the expressive dimension of these laws as themselves providing a reason in favor of legislation (Waldron, 2014). The law, in this sense, serves as a public statement on a community's values, and has educational and symbolic importance in itself (Parekh, 2012, 46). (For an overview of expressive theories of law, see Anderson and Pildes, 2000.) A ban on hate speech, therefore, is intended both to reduce harms directly, by decreasing instances of hate by the threat of law, as well as indirectly, by shaping the community's moral norms through an expression of value.

5.2 Objections to Bans, and Some Responses

Though many would agree that hate speech can have destructive effects, and that there is a moral imperative on the state to cultivate something like respectful relations between its members, objections to hate speech bans abound. In a wide-ranging response to these concerns, Parekh (2012) considers (and rejects) six common objections to the prohibition of hate speech. These six objections are: (1) that the harm of hate speech, while real, is relatively minor

and a small price to pay given the interest of democratic nations; (2) that bans are not the answer, but rather "better ideas" and "more speech" are; (3) that a prohibition would have a dangerous "chilling effect" and that hate speech bans are a slippery slope to all sorts of unwanted restrictions; (4) that bans give the state too much power to judge the content of speech and decide what can or cannot be said, threatening state-neutrality, skewing political debate, and infringing on individual liberty; (5) that bans are an objectionable form of paternalism or moral authoritarianism, and is incompatible with the assumption that humans are responsible and autonomous individuals and that society is made up of free and equal citizens; and finally, (6) that bans are ineffective at changing attitudes and removing the hate from the hate speaker's heart, with the result that bans have the effect of moving extremists underground, alienating them from wider society, and in doing so rendering us ignorant of their violent potential and impotent to engage in effective de-radicalization.

Each of these concerns merits more space than can be given to them here. Still, considering these objections to bans and the responses available, even briefly, is illustrative of the theoretical concerns bans on hate speech bring forth (see the list of references for fuller development of the relevant theoretical and empirical issues). Again following Parekh (2012, 47–54), we can approach these objections as follows.

In response to (1), the objection that the interests of a vibrant democracy outweigh the harms imposed by hate speech, it may be argued that hate speech does not embody the values of free speech but, in fact, undermines them by promoting irrational fears and hatred over reasoned arguments and public scrutiny. How powerful one takes this response to be depends directly on what one takes the value and justification of a right to free expression to be, which is of course a matter of dispute.

One response to (2), the common 'more speech' objection, is to note that the "marketplace of ideas" is not neutral, and likely requires some regulation (just like a marketplace of other goods). This is what a ban does, and so may be considered to be helping ensure 'fair competition' by countering prevailing prejudices, and encouraging greater participation from the members of communities targeted by hate speech. In other words, bans on hate speech may promote greater freedom of expression, by preventing the type of silencing considered above.

While acknowledging the worries of (3), namely that of a 'chilling effect' or a 'slippery slope,' represent an important objection, we may respond by noting that the problems these signal rest on the vague wording and inconsistent or biased application of hate speech bans. They are not, therefore, direct objections to hate speech bans as such. The remedy, therefore, lies in correcting these aspects of a ban, rather than abandoning it altogether. Moreover, the appeal to a 'slippery slope' may be inapt, as it implies that once one type of speech is prohibited, society cannot help but prohibit even more types. But we have no clear reason to suppose that this is the case, as existing bans on defamation have not led to bans on fair critical comment, for example.

The worries at the core of objection (4) represents a well-founded fear of the state, and so must be taken seriously. But, to defenders of hate speech bans, its understanding of the threat that hate speech bans pose to state-neutrality is nonetheless flawed. It fails to recognize that the state often *already* judges the content of speech (e.g., in banning commercial fraud, criminal solicitation, public displays of obscenity) and often elides neutrality when it speaks in favor of certain positions (e.g., the value of human dignity, equality, liberty). While any defense of hate speech bans must reckon with the possibility of further empowering the state, opponents ought not misrepresent the status quo, exaggerating the reality of state-neutrality.

Objections grounded on the threat of paternalism or moral authoritarianism, like (5), are similarly serious. However, one response on behalf of bans would be to point out how autonomy is always exercised under certain conditions and requires various external circumstances for its development and use. When appealing to personal autonomy, therefore, we should not idealize too greatly so that its real-world exercise is ignored. Rather, the threats that racism and bigotry pose for autonomy must also be acknowledged, alongside praise for our rational faculties.

One response to (6), that bans are ineffective at changing attitudes, is to admit the law cannot change attitudes (like hatred) directly and maintain that this is no knock against the law, and indeed is no problem for hate speech bans. The aim of these bans, in most cases, is not to prevent hatred but to prevent the harm that *the public expression* of hate can cause. The indirect effects of such a law, however, are an empirical matter, and it is unlikely they admit of a single, general answer, but are highly context-dependent. The subject of the practicability of hate speech bans deserves special attention, however. Opponents to bans may worry that the suppression of hate speech is likely to backfire, not only by failing to reduce hatred, but by *increasing* the sense of oppression and victimization that many bigots thrive on, leading to an escalation of racist violence (Baker, 2012, 77). Again, as an empirical hypothesis, it cannot be settled simply from the armchair. Still, a further response available on behalf of hate speech bans would be to question the legitimacy of this objection. If, by hypothesis, bans generated an increase in violence, it would still be the

responsibility of the state to manage this violence effectively. The role of the state is not exhausted by implementing a ban, but must be seen alongside its enforcement.

This, however, leads to a slightly different objection. The opponent of bans may worry that the enforcement of laws against hate speech would divert the state's energies away from more effective measures against hated, such as "those directed at changing material conditions in which racism festers, material conditions of both the purveyors and targets of hate" (Baker, 2012, 77). That is, the energies and resources that would be directed towards establishing and enforcing hate speech bans may be better spent on alternative policies. The guiding thought rests on two important points. First, that the intended ends of hate speech bans (e.g., reduction in the harms of hate speech that fall on those targeted by it, mitigation of the expansion of racist attitudes, lessening occurrences of violent hate crimes) may be more effectively achieved via different means, such as reducing inequality, improving social safety nets, political empowerment, and more. Second, though the state can do more than one thing at once, it is nonetheless working with limited resources, and efficiency is a value. That these alternative policy options may indeed be more effective is an unresolved empirical matter. And it remains an open question whether indirect approaches like this would fail to achieve the expressive ends of hate speech bans, which more directly communicate to those targeted by hate speech that they are valued members society.

Many of the claims made above, both on behalf of bans and in opposition, raise theoretical and empirical issues whose proper examination spans many articles and books. Suffice to say that the debate over bans is a highly contested one, and each position rests on an understanding of such issues as the value of free expression, the harm of hate speech, the likely effects a ban might have in a particular context, and so on. For instance, one who believes that free expression is valuable in part because of its role in democratic decision-making may maintain that specifically *political* speech deserves increased protections, and that some of what others regard as hate speech might fall into this category, escaping regulation. Alternatively, one may view the immunity for political speech as perhaps a red herring. On the speech act theoretic framework outlined above, some forms of racist hate speech are functionally identical to a 'Whites Only' sign hanging in a public restaurant (McGowan 2012; McGowan and Maitra 2009). The latter expresses a political opinion in the same way as the former expression does, but it is also regarded as unlawful racially discriminatory. The same considerations—legal sanction—might therefore apply to the verbal utterance as the written sign, and the appeal to the political content of the message is irrelevant.

5.3 The Supported Counterspeech Alternative

The preceding summarizes the two main positions in the debate over hate speech: on the one hand, there are those who defend prohibitions, and on the other, those who maintain hate speech as protected under a wide conception of freedom of expression, and so oppose laws that aim at its prohibition. A third position aims to avoid some of the impasses that haunt this debate. On this view, this impasse is the result of a failure by those who oppose hate speech bans (and, as a result, tend to favor 'more speech') to acknowledge the strength of one of the main arguments from those who advocate for bans, namely, that hate speech is a type of assault that often renders one unable to respond. This, along with a failure of those who defend bans from considering non-punitive options for mitigating the harms of hate speech, leads to stalemate. On this understanding, both sides of the debate over bans see the only alternatives as either increased governmental powers to punish, or absent that, 'unsupported' counterspeech on the part of those targeted by hate speech (see Gelber, 2012a; 2012b).

By contrast, the "supported counterspeech" alternative aims to recognize the specific harms inflicted by hate speech and provide state support to empower those who are harmed. Gelber, an advocate for this alternative, places it within the capabilities approach originally developed by Amartya Sen (1992) and Martha Nussbaum (2000; 2003). "If hate-speech acts harm their targets' capacity to develop human capabilities," Gelber says, then "this is what needs to be remedied" (2012a, 54). The impetus for this approach therefore begins from the idea that we must think about remedies to hate speech beyond restrictions and punishment, as neither of these approaches achieve the goal of empowering the target of hate speech. (This is especially true of the latter, punishment, which also carries with it all the negative consequences that anti-carceral advocates have noted.) The supported counterspeech policy is therefore not focused on hate speakers, but rather the targets of hate speech more directly.

The core of this approach lies in an enlarged conception of counterspeech as well as a commitment by the state to provide the material conditions necessary for this speech. In practice, this would mean that the state is committed to responding to an incident of hate speech by empowering its targets to engage in more speech, after the fact. The specific forms this support may take will depend on the conditions of different contexts, along with calibration for the specifics of the incident it is meant as a response to, as well as the needs of the particular communities. Still, to give a sense of what this may entail, examples of the sort of supported counterspeech that this position recommends include things such as: assistance in the production of a community newsletters, op-eds, radio broadcasts, or television

advertisements; the development of antiracism awareness programs, or anti-hate-speech workshops; subsidizing community-led art projects; etc.

In each case, the aim of supported counterspeech is to empower the targets of hate speech, and to increase their capacity for engaging in counterspeech. The goal is thus to undo (as much as one can) the specific harms of hate speech, while avoiding the pitfalls of "private remedies" (as critiqued by Matsuda, 1993). While supported counterspeech could be taken as either an alternative to bans or a supplement to them, it remains an under-explored avenue for considering responses tailored to the particular harms of hate speech.

Bibliography

Anderson, Elizabeth, and Richard Pildes, 2000, "Expressive Theories of Law: A General Restatement," *University of Pennsylvania Law Review*, 148: 1503–1575.

Anderson Luvell and Lepore, Ernest, 2013a, "Slurring Words," *Noûs*, 47(1): 25–48.

——, 2013b, "What Did You Call Me? Slurs as Prohibited Words," *Analytic Philosophy*, 54(3): 350–363.

Appiah, Kwame Anthony, 2012, "What's Wrong with Defamation of Religion?" in Michael Herz, and Peter Molnar (eds.), *The Content and Context of Hate Speech: Rethinking Regulation and Response*, Cambridge; New York: Cambridge University Press, pp. 164–182.

Ashwell, Lauren, 2016, "Gendered Slurs," *Social Theory and Practice*, 42(2): 228–39.

Austin. J. L., 1962, *How to do Things with Words*, Cambridge, MA: Harvard University Press.

Bach, Kent, 2018, "Loaded Words: On the Semantics and Pragmatics of Slurs," in David Sosa (ed.), *Bad Words: Philosophical Perspectives on Slurs*, Oxford: Oxford University Press, pp. 60–76.

Baker, C. Edwin, 2012, "Hate Speech," in Michael Herz, and Peter Molnar (eds.), *The Content and Context of Hate Speech: Rethinking Regulation and Response*, Cambridge, New York: Cambridge University Press, pp. 57–80.

Bauer, Nancy, 2015, *How to Do Things with Pornography*, Cambridge, MA: Harvard University Press.

Beauharnais v. People of the State of Illinois, 343 U.S. 250 (1952).

Behrens,Paul, Terry, N. and Jensen, O (eds.), 2017, *Holocaust and Genocide Denial: A Contextual Perspective*, Abingdon; New York: Routledge, Taylor and Francis Group.

Bolinger, Renée Jorgensen, 2017, "The Pragmatics of Slurs," *Noûs*, 51(3): 439–462.

Brandenburg v. Ohio, 395 U.S. 444 (1969).

Brandom, Robert, 1994, *Making it Explicit*, Cambridge, MA: Harvard University Press.

Brown, Alexander, 2017a, "What Is Hate Speech? Part 1: The Myth of Hate," *Law and Philosophy*, 36(4): 419–468.

——, 2017b, "What Is Hate Speech? Part 2: Family Resemblances," *Law and Philosophy*, 36(5): 561–613.

Brison, Susan, 1998a, "The Autonomy Defense of Free Speech," *Ethics*, 108(2): 312–339.

——, 1998b, "Speech, Harm, and the Mind-Body Problem in First Amendment Jurisprudence," *Legal Theory*, 4: 39–61.

Butler, Judith, 1997, *Excitable Speech: A Politics of the Performative*, New York: Routledge.

Camp, Elisabeth, 2013, "Slurring Perspectives," *Analytic Philosophy*, 54(3): 330–349.

——, 2018, "Slurs as Dual-Act Expressions," in David Sosa (ed.), *Bad Words*, Oxford: Oxford University Press, pp. 29–59.

Caponetto, Laura, 2021, "A Comprehensive Definition of Illocutionary Silencing," *Topoi*, 40: 191–202.

Cepollaro, Bianca, 2015, "In Defense of a Presuppositional Account of Slurs," *Language Sciences*, 52: 36–45.

Croom, Adam, 2011, "Slurs," *Language Sciences*, 33: 343–358.

Delgado, Richard, 1993, "Words that Wound: A Tort Action for Racial Insults, Epithets, and Name Calling," in Mari Matsuda, Charles Lawrence III, Richard Delgado, and Kimberlé Crenshaw (eds.), *Words that Wound: Critical Race Theory, Assaultive Speech and the First Amendment*, Boulder, CO: Westview Press, pp. 89–110.

Delgado, Richard, and Jean Stefancic, 2004, *Understanding Words that Wound*, Boulder, CO: Westview Press.

Dummett, Michael, 1993, *Frege: Philosophy of Language*, Cambridge, MA: Harvard University Press.

Gallie, W. B., 1955, "Essentially Contested Concepts," *Proceedings of the Aristotelian Society*, 56: 167–198.

Gelber, Katharine, 2012a, "Reconceptualizing Counterspeech in Hate-Speech Policy (with a Focus on Australia)," in Michael Herz, and Peter Molnar (eds.), *The Content and Context of Hate Speech: Rethinking Regulation and Response*, Cambridge, New York: Cambridge University Press, pp. 198–216.

——, 2012b, "'Speaking Back': The Likely Fate of Hate Speech Policy in the United States and Australia," in Ishani Maitra and Mary Kate McGowan (eds.), *Speech & Harm: Controversies Over Free Speech*, New York: Oxford University Press, pp. 50–71.

——, 2017, "Hate Speech—Definitions & Empirical Evidence," *Constitutional Commentary*, 32(3): 619–629.

Gelber, Katharine, and Luke McNamara, 2016, "Evidencing the Harms of Hate Speech," *Social Identities*, 22 (3): 324–341.

Herz, Michael, and Peter Molnar (eds.), 2012, *The Content and Context of Hate Speech: Rethinking Regulation and*

Responses, Cambridge, New York: Cambridge University Press.
Hesni, Samia, 2018, "Illocutionary Frustration," *Mind*, 127(508): 947–976.
Heyman, Steven, 2008, *Free Speech and Human Dignity*, New Haven: Yale University Press.
Hom, Christopher, 2008, "The Semantics of Racial Epithets," *Journal of Philosophy*, 105: 416–440.
Hornsby, Jennifer, 1994, "Illocution and its Significance," in Savas L. Tsohatzidis (ed.), *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives*, London: Routledge, pp. 187–207.
——, 2001, "Meaning and Uselessness: How to Think About Derogatory Words," *Midwest Studies in Philosophy*, 25(1): 128 – 141.
Hornsby, Jennifer, and Rae Langton, 1998, "Free Speech and Illocution," *Legal Theory*, 4: 21–37.
Imbleau, Martin, 2011, "Denial of the Holocaust, Genocide, and Crimes Against Humanity," in Ludovic Hennebel and Thomas Hochmann (eds.) *Genocide Denials and the Law*, New York: Oxford University Press, pp. 235–277.
Jeshion, Robin, 2013a, "Expressivism and the Offensiveness of Slurs," *Philosophical Perspectives*, 27: 231–259.
——, 2013b, "Slurs and Stereotypes," *Analytic Philosophy*, 54 (3): 314–329.
Khoo, Justin, 2017, "Code Words in Political Discourse," *Philosophical Topics*, 45 (2): 33–64.
Kirk-Giannini, Cameron Domenico, 2019, "Slurs are Directives," *Philosopher's Imprint*, 19(48): 1–28.
Kukla, Quill [writing as Rebecca], 2014, "Performative Force, Convention, and Discursive Injustice," *Hypatia*, 29 (2): 440–457.
——, 2018, "Slurs, Interpellation, and Ideology," *The Southern Journal of Philosophy*, 56(S1): 7–32.
Langton, Rae, 1993, "Speech Acts and Unspeakable Acts," *Philosophy & Public Affairs*, 22(4): 293–330.
——, 2012, "Beyond Belief: Pragmatics in Hate Speech and Pornography," in Ishani Maitra and Mary Kate McGowan (eds.), *Speech & Harm: Controversies Over Free Speech*, New York: Oxford University Press, pp. 72–93.
——, 2018a, "The Authority of Hate Speech," in John Gardner, Leslie Green, and Brian Leiter (eds.), *Oxford Studies in Philosophy of Law* (Volume 3), New York: Oxford University Press, pp. 123–152.
——, 2018b, "Blocking as Counter Speech," in Daniel W. Harris Daniel Fogal, and Matt Moss (eds.), *New Works on Speech Acts*, Oxford: Oxford University Press, pp. 144–164.
Langton Rae, and Caroline West, 1999, "Scorekeeping in a Pornographic Language Game," *Australasian Journal of Philosophy*, 77(3): 303–319.
Lawrence, Charles III, 1993, "If He Hollers Let Him Go: Regulating Racist Speech on Campus," in Mari Matsuda, Charles Lawrence III, Richard Delgado, and Kimberlé Crenshaw (eds.), *Words that Wound: Critical Race Theory, Assaultive Speech and the First Amendment*, Boulder, CO: Westview Press, pp. 53–88.
Haney Lopez, Ian, 2014, *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class*, Oxford: Oxford University Press.
MacKinnon, Catharine, 1993, *Only Words*, Cambridge, MA: Harvard University Press.
Maitra, Ishani, 2009, "Silencing Speech," *Canadian Journal of Philosophy*, 39(2): 309–338.
——, 2012, "Subordinating Speech," in Ishani Maitra and Mary Kate McGowan (eds.), *Speech & Harm: Controversies Over Free Speech*, New York: Oxford University Press, pp. 94–120.
Maitra, Ishani, and Mary Kate McGowan (eds.), 2012, *Speech & Harm: Controversies Over Free Speech*, New York: Oxford University Press.
Matsuda, Mari, 1993, "Public Response to Racist Speech: Considering the Victim's Story," in Mari Matsuda, Charles Lawrence III, Richard Delgado, and Kimberlé Crenshaw (eds.), *Words that Wound: Critical Race Theory, Assaultive Speech and the First Amendment*, Boulder, CO: Westview Press, pp. 17–51.
Matsuda, Mari, Charles Lawrence III, Richard Delgado, and Kimberlé Crenshaw (eds.), 1993, *Words that Wound: Critical Race Theory, Assaultive Speech and the First Amendment*, Boulder, CO: Westview Press.
McGowan, Mary Kate, 2004, "Conversational Exercitives: Something Else We Do With Our Words," *Linguistics and Philosophy*, 27: 93–111.
——, 2009, "Oppressive Speech," *Australasian Journal of Philosophy*, 87(3): 389–407.
——, 2012, "On 'Whites Only' Signs and Racist Hate Speech: Verbal Acts of Racial Discrimination," in Ishani Maitra and Mary Kate McGowan (eds.), *Speech & Harm: Controversies Over Free Speech*, New York: Oxford University Press, pp. 121–47.
——, 2014, "Sincerity Silencing," *Hypatia*, 29(2): 458–473.
——, 2018, "On Covert Exercitives: Speech and the Social World," in Daniel W. Harris Daniel Fogal, and Matt Moss (eds.), *New Works on Speech Acts*, Oxford: Oxford University Press, pp. 185–201.
——, 2019, *Just Words: On Speech and Hidden Harm*, Oxford: Oxford University Press.
McGowan, Mary Kate, and Ishani Maitra, 2009, "On Racist Hate Speech and the Scope of a Free Speech Principle," *Canadian Journal of Law and Jurisprudence*, 23(2): 343–372.
Mendelberg, Tali, 2001, *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*, Princeton: Princeton University Press.
Mikkola, Mari, 2011, "Illocution, Silencing and the Act of Refusal," *Pacific Philosophy Quarterly*, 92: 415–437.
——, 2019, *Pornography: a Philosophical Introduction*, New York: Oxford University Press.

Miller v. California, 413 U.S. 15 (1973).
Neufeld, Eleonore, 2019, "An Essentialist Theory of the Meaning of Slurs," *Philosopher's Imprint*, 19(35): 1–29.
Nunberg, Geoff, 2018, "The Social Life of Slurs," in Daniel W. Harris, Daniel Fogal, and Matt Moss (eds.), *New Works on Speech Acts*, Oxford: Oxford University Press, pp. 237–295.
Nussbaum, Martha, 2000, *Women and Human Development: The Capabilities Approach*, Cambridge: Cambridge University Press.
——, 2003, "Capabilities as Fundamental Entitlements: Sen and Social Justice," *Feminist Economics*, 9(2/3): 33–59.
O'Donnell, Patrick, 2017, "Generics, Race, and Social Perspectives," *Inquiry: An Interdisciplinary Journal of Philosophy*, published online 26 January 2017. doi:10.1080/00207543.2016.1266801
Parekh, Bhikhu, 2012, "Is There a Case for Banning Hate Speech?" in Michael Herz, and Peter Molnar (eds.), *The Content and Context of Hate Speech: Rethinking Regulation and Response*, Cambridge, New York: Cambridge University Press, pp. 37–56.
Popa-Wyatt, Mihaela, and Jeremy L. Wyatt, 2017, "Slurs, Roles and Power," *Philosophical Studies*, 175(11): 2879–2906.
Post, Robert, 2009, "Hate Speech," in Ivan Hare and James Weinstein (eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.
Reichman, Ammon, 2009, "Criminalizing Religiously Offensive Satire: Free Speech, Human Dignity, and Comparative Law," in Ivan Hare and James Weinstein (eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press, pp. 331–355.
Richard, Mark, 2010, *When Truth Gives Out*, Oxford: Oxford University Press.
Saul, Jennifer M., 2018, "Dogwhistles, Political Manipulation, and Philosophy of Language," in Daniel W. Harris Daniel Fogal, and Matt Moss (eds.), *New Works on Speech Acts*, Oxford: Oxford University Press, pp. 360–383.
Sbisà, Marina, 2009, "Illocution and Silencing," in Bruce Fraser, and Ken Turner (eds.), *Language in Life, and a Life in Language: Jacob Mey—a Festschrift*, Bradford: Emerald, pp. 351–357.
Schwartzman, Lisa H., 2002, "Hate Speech, Illocution, and Social Context: A Critique of Judith Butler," *Journal of Social Philosophy*, 33(3): 421–441.
Sen, Amartya, 1992, *Inequality Reexamined*, Cambridge, MA: Harvard University Press.
Stanley, Jason, 2015, *How Propaganda Works*, Princeton: Princeton University Press.
Steinberger, Florian, and Julien Murzi, 2017, "Inferentialism," in Bob Hale, Crispin Wright, Alexander Miller (eds.), *A Companion to the Philosophy of Language*, second edition, Oxford: Wiley Blackwell, pp. 197–224.
Tesler, M., and D. O. Sears, 2010, *Obama's Race: The 2008 Election and the Dream of a Post-Racial America*, Chicago: University of Chicago Press.
Tirrell, Lynne, 1999, "Derogatory Terms: Racism, Sexism, and the Inferential Role Theory of Meaning," in Christina Hendricks, and Kelly Oliver (eds.), *Language and Liberation: Feminism, Philosophy, and Language*, Albany: State University of New York Press, pp. 41–79.
Tonhauser, Judith, 2012, "Diagnosing (not-)at-issue content," *Proceedings of Semantics of Underrepresented Languages in the Americas*, 6: 239–254.
UN General Assembly, *International Covenant on Civil and Political Rights*, 16 December 1966.
UN General Assembly, *International Convention on the Elimination of All Forms of Racial Discrimination*, 21 December 1965, [available online].
Waldron, Jeremy,2014, *The Harm in Hate Speech*, Cambridge, MA: Harvard University Press.
West, Caroline, 2012, "Words that Silence? Freedom of Expression and Racist Hate Speech," in Ishani Maitra, and Mary Kate McGowan (eds.), *Speech and Harm: Controversies over Free Speech*, Oxford: Oxford University Press, pp. 222–248.
Whine, Michael, 2009, "Expanding Holocaust Denial and Legislation Against It," in Ivan Hare and James Weinstein (eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press, pp. 538–556.
White, Ismail, 2007, "When Race Matters and When It Doesn't: Racial Group Differences in Response to Racial Cues," *American Political Science Review*, 101(2): 339–54.
Whiting, Daniel, 2008, "Conservatives and Racists: Inferential Role Semantics and Pejoratives," *Philosophia*, 36(3): 375–388.
Williamson, Timothy, 2009, "Reference, Inference, and the Semantics of Pejoratives," in J. Almog, and P. Leonardi (eds.), *The Philosophy of David Kaplan*, Oxford: Oxford University Press, pp. 137–158.
Wistrich, Robert, 2012, *Holocaust Denial: The Politics of Perfidy*, Berlin: De Gruyter.

Academic Tools

How to cite this entry.
Preview the PDF version of this entry at the Friends of the SEP Society.

Look up topics and thinkers related to this entry at the Internet Philosophy Ontology Project (InPhO).

Enhanced bibliography for this entry at PhilPapers, with links to its database.

Other Internet Resources

- Swanson, Eric, 2015, "Slurs and Ideology," unpublished manuscript
- American Library Association, "Hate Speech and Hate Crime"
- European Commission against Hate Speech and Violence
- UN Strategy and Plan of Action on Hate Speech
- UNESCO, "Countering Online Hate Speech"
- Rights for Peace, "Hate Speech"
- Anti-Defamation Leage, "Resources for Responding to Hate Speech in the Community"
- Library of Parliament Research Publication, "Hate Speech and Freedom of Expression: Legal Boundaries in Canada"

Related Entries

cognitivism vs. non-cognitivism, moral | freedom: of speech | pornography: and censorship | respect | semantics: proof-theoretic | speech acts

Copyright © 2022 by
Luvell Anderson <*lander04@syr.edu*>
Michael Barnes <*Michael.barnes@anu.edu.au*>

Open access to the SEP is made possible by a world-wide funding initiative.
Please Read How You Can Help Support the Growth and Development of the Encyclopedia

The Stanford Encyclopedia of Philosophy is copyright © 2023 by The Metaphysics Research Lab, Department of Philosophy, Stanford University

Library of Congress Catalog Data: ISSN 1095-5054