

HATE SPEECH LAWS: EXPRESSIVE POWER IS NOT THE ANSWER

1. Introduction

The public discourse of contemporary democracies is rife with intensely disrespectful speech, the worst of which is sometimes labelled ‘hate speech’. While definitions of hate speech vary extensively, they commonly revolve around the idea that hate speech communicates or promotes the inferiority of other members of society. More specifically, hate speech emphatically rejects the basic standing of its targets as equals, typically on the basis of their membership in a vulnerable social group.¹

Thus understood, hate speech might involve newspaper articles portraying vulnerable groups as subhuman (e.g., depictions of immigrants as ‘cockroaches’ or

¹ See, e.g., United Nations, “International Convention on the Elimination of All Forms of Racial Discrimination” (1965); Jeremy Waldron, *The Harm in Hate Speech* (Cambridge, MA: Harvard University Press, 2012), 56–57; Corey Brettschneider, *When the State Speaks, What Should It Say?* (Princeton, NJ: Princeton University Press, 2012), 1; Rae Langton, “Beyond Belief: Pragmatics in Hate Speech and Pornography,” in *Speech and Harm*, ed. Ishani Maitra and Mary Kate McGowan, 72–93 (Oxford: Oxford University Press, 2012), 74–77; Robert Simpson, “Dignity, Harm, and Hate Speech,” *Law and Philosophy* 32 (2013): 701n2; Eric Heinze, “Hate Speech and the Normative Foundations of Regulation,” *International Journal of Law in Context* 9 (2013): 16. Some definitions of hate speech focus less on what it expresses, and more on its likely consequences. For instance, the UK’s Public Order Act of 1986 prohibits, among other things, speech that is *likely to stir up hatred*. However, because I will be examining the expressive dimension of legal regulations of hate speech, it is more useful for my purposes to characterize hate speech primarily in terms of what it expresses.

‘snakes’); public statements ascribing essential attributes to certain groups in virtue of which they are unsuitable for democratic life (e.g., ‘Muslims are terrorists’); or banners that directly express the social exclusion of religious or racial minorities (e.g., ‘Whites only’).

Unless it is countered appropriately, such speech risks inflicting serious harms on its targets. As philosophers of language, political philosophers, and legal theorists have forcefully argued, hate speech can, among other things, cause its targets acute psychological distress,² assault their assurance of dignity,³ damage their autonomy,⁴ stir up violence or animosity towards them,⁵ or silence their speech.⁶

How should we go about countering these potential harms? A prominent proposal recommends countering hate speech with more speech.⁷ Many, however,

² Richard Delgado, “Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling,” *Harvard Civil Rights-Civil Liberties Law Review* 17 (1982): 133–81.

³ Waldron, *The Harm in Hate Speech*.

⁴ Susan Brison, “The Autonomy Defense of Free Speech,” *Ethics* 108 (1998): 312–39.

⁵ Lynne Tirrell, “Genocidal Language Games,” in *Speech and Harm*, ed. Ishani Maitra and Mary Kate McGowan, 174–221 (Oxford: Oxford University Press, 2012).

⁶ Rae Langton, “Hate Speech and the Epistemology of Justice,” *Criminal Law and Philosophy* 10 (2014): 865–873.

⁷ See, e.g., Louis Brandeis, “Opinion in *Whitney v California*, 274 US 357,” 1927; Brettschneider, *When the State Speaks, What Should It Say?*; Katharine Gelber, “Reconceptualizing Counterspeech in Hate Speech Policy (with a Focus on Australia),” in *The Content and Context of Hate Speech*, ed. Michael Herz and Peter Molnar (Cambridge: Cambridge University Press, 2012), 198–216; Nadine Strossen,

contend that such ‘counterspeech’ is insufficient. On this view, the seriousness of the harms hate speech may give rise to demands that we ban it. In other words, we should adopt criminal or civil laws that prohibit hate speech, and that threaten to impose sanctions (such as significant fines or incarceration) on offenders.⁸ This legal response to hate speech is popular in practice as well as in theory: with the notable exception of the United States, democracies throughout the world have enacted criminal or civil prohibitions on hate speech.⁹

There are different possible justifications for hate speech laws. One important argument asserts that, by threatening to impose sanctions on hate speakers, such laws *deter* people from engaging in hate speech. Thus, they eliminate hateful utterances from public discourse, together with the harms they would otherwise have occasioned.¹⁰

But, although the deterrence argument remains influential, I wish to examine a different argument for hate speech laws. This argument, which has garnered widespread support from philosophers and lawyers, holds that an essential part of

“Interview with Nadine Strossen,” in *The Content and Context of Hate Speech*, ed. Michael Herz and Peter Molnar (Cambridge: Cambridge University Press, 2012), 378–98.

⁸ See, e.g., Delgado, “Words That Wound”; Mari Matsuda, “Public Response to Racist Speech,” *Michigan Law Review* 87 (1989): 2320–81; Brison, “The Autonomy Defense of Free Speech”; Waldron, *The Harm in Hate Speech*; Alexander Brown, *Hate Speech Law: A Philosophical Examination* (Abingdon: Routledge, 2015).

⁹ For a philosophically sophisticated overview, see Brown, *Hate Speech Law*.

¹⁰ Delgado, “Words That Wound,” 148.

what justifies hate speech laws is their *expressive* dimension—that is, the message they send out. According to this line of thought, the public statement of condemnation that hate speech laws direct at hate speakers and their worldview plays a key role in combatting hate speech and its potential for harm.¹¹

The expressive argument is *prima facie* attractive for two reasons. First, it is congruent with broader theories of law: legal theorists typically recognize that laws and the punitive sanctions they impose have an important expressive dimension. Second, and perhaps more importantly, this argument seems to circumvent the empirical difficulties that continue to plague the deterrence argument. As we will see, evidence that bans are successful at deterring hate speech remains highly elusive. By emphasizing the law’s symbolic message instead, the expressive strategy appears to sidestep these protracted concerns.

Yet the expressive argument poses a puzzle. Opponents of hate speech laws, recall, typically advocate ‘more speech’ (or ‘counterspeech’) as the best way of

¹¹ In theoretical discussions, see Lee Bollinger, *The Tolerant Society* (Oxford: Oxford University Press, 1986), 72; David Kretzmer, “Freedom of Speech and Racism,” *Cardozo Law Review* 8 (1987): 513; Matsuda, “Public Response to Racist Speech,” 2322; David Partlett, “From Red Lion Square to Skokie to the Fatal Shore: Racial Defamation and Freedom of Speech,” *Vanderbilt Journal of Law* 22 (1989): 473; Bhikhu Parekh, “The Rushdie Affair: Research Agenda for Political Philosophy” 38 (1990): 705; Anna Elisabetta Galeotti, *Toleration as Recognition* (Cambridge: Cambridge University Press, 2002), 156; Waldron, *The Harm in Hate Speech*, 80–81; Brown, *Hate Speech Law*, 240. In legal discussions, the expressive rationale for hate speech laws is explicitly advanced in *R. v. Ali, Javed, and Ahmed* (cited in Brown, *Hate Speech Law*, 241.).

dealing with hate speech. And if hate speech laws are defended by appeal to their expressive power, it becomes unclear what they offer that counterspeech does not. As H.L.A. Hart observes when discussing the expressive function of law more generally, “it is not clear, if denunciation is really what is required, why a solemn public statement of disapproval would not be the most ‘appropriate’ or ‘emphatic’ means of expressing this.”¹² The puzzle, in other words, is the following. The expressive defense of legal bans construes them, roughly, as a kind of speech, which conveys a message. But that is what counterspeech is centrally designed to do. So, the expressive defense of bans makes it difficult to understand why bans are needed. After all, if the function of hate speech laws can readily be performed without imposing sanctions on speech—sanctions which, it has been argued, impose pro tanto costs on the freedom or autonomy of hate speakers¹³—then it seems we should forego such laws.

¹² H.L.A. Hart, *Law, Liberty, and Morality* (Stanford, CA: Stanford University Press, 1963), 66. In debates about hate speech, see also: James Weinstein, “A Constitutional Roadmap to the Regulation of Regulation of Campus Map,” *Wayne Law Review* 38 (1991): 245–46; Thomas Scanlon, “The Significance of Choice,” in *The Tanner Lectures on Human Values* (Salt Lake City, UT: University of Utah Press, 1988), 214; Robert Post, “Interview with Robert Post,” in *The Content and Context of Hate Speech*, ed. Michael Herz and Peter Molnar (Cambridge: Cambridge University Press, 2012), 33.

¹³ See, e.g., Thomas Scanlon, “A Theory of Freedom of Expression,” *Philosophy & Public Affairs* 1 (1972): 204–26; Edwin Baker, “Harm, Liberty, and Free Speech,” *Southern California Law Review* 70 (1996): 979–1020.

For the expressive argument to succeed, then, hate speech laws must have a *distinctive* expressive force, which cannot be realized by forms of counterspeech that forego hate speech laws (which, henceforth, I will be referring to simply as ‘counterspeech’). This is precisely what exponents of the expressive argument tend to insist. Lee Bollinger, for example, asserts that enacting a law prohibiting hate speech “is usually a much more powerful demonstration of a community’s commitment[s] [...] than is a simple verbal declaration.”¹⁴ Likewise, David Partlett affirms that “legislation is governmental speech of the most potent kind”.¹⁵

In what follows, I aim to challenge this ‘distinctiveness’ claim. In particular, I will demonstrate that arguments for the expressive distinctiveness of hate speech laws encounter the following problem: *either* they fail to show that counterspeech could not perform the expressive function of hate speech laws, *or* they do identify an expressive function that seems distinctive, but its success depends wholly on the success of the deterrence argument.

More specifically, my argument will unfold as follows. After outlining the expressive argument for hate speech laws (Section 2), I will examine three types of considerations that purport to explain their distinctive expressive dimension: considerations of strength (Section 3); considerations of directness (Section 4); and considerations of complicity (Section 5). These considerations, I will demonstrate,

¹⁴ Bollinger, *The Tolerant Society*, 122.

¹⁵ Partlett, “From Red Lion Square to Skokie to the Fatal Shore,” 468. See also: Waldron, *The Harm in Hate Speech*, 87–89; Brown, *Hate Speech Law*, 263.

either fail to establish that bans are expressively distinctive, or presuppose that bans successfully deter hate speech.

The upshot is that the expressive argument offers no independent support for hate speech bans. To the extent that bans do not play a distinctive expressive role, the expressive argument does not give us any reason to supplement counterspeech with bans. And even insofar as bans may have a distinctive expressive role, this gives us a reason to adopt bans only if the elusive deterrent argument can first be vindicated.

Before proceeding, two clarifications are needed. First, my argument does not purport to establish that bans are altogether unjustified. Rather, it establishes that one of the most influential justifications for bans—the expressive argument—is at best parasitic on another justification—the deterrence argument. Thus, despite its critical form, my argument ultimately has a constructive result for advocates of hate speech laws: to justify such laws, they should focus their efforts, first and foremost, on vindicating the empirical claim that bans deter hate speech. Despite appearances to the contrary, appealing to bans' expressive dimension cannot help them circumvent this empirical controversy.¹⁶

¹⁶ Although my investigation focuses on the expressive and deterrence arguments, there may also be other arguments for hate speech laws. In particular, one might argue that hate speech laws are justified on the retributivist ground that they inflict deserved punishment on hate speakers. Importantly, the existence of this alternative justification does not significantly affect my central contention: that the expressive argument either does not work, or is parasitic on the deterrence argument—and, consequently, that those who embrace the expressive argument must first vindicate the elusive empirical

Second, I will not be relying on the claim that bans' ability to send a morally desirable message depends on their deterrent effect. One might worry that, if hate speech laws fail to suppress hate speech, this will inadvertently send out an *undesirable* message. For instance, it might suggest that the government is only pretending to take hate speech seriously.¹⁷ If so, this provides a straightforward route to my conclusion: if bans cannot unambiguously communicate the condemnation of hateful views unless they succeed as deterrents, then it follows that the expressive argument is at best parasitic on the deterrence argument.

However, I will assume for the sake of argument that bans can successfully communicate a morally desirable message—e.g., the condemnation of hateful worldviews—even if they fail as deterrents. What I will show is that, even if we grant this, the problem resurfaces at a later stage. To give reasons for adopting bans, the expressive argument must establish not only that bans can successfully express condemnation, but that they can do so in a distinctive way, which could not be realized via counterspeech. My point is that, to establish this further claim, advocates of bans must appeal to their success as deterrents.

claim that bans deter hate speech. Now, in light of these difficulties, one might recommend circumventing both the expressive *and* the deterrence arguments, and focusing on the retributivist argument instead. But this strategy too remains broadly congruent with one of my main points: that, unlike what proponents of the expressive argument often suggest, we cannot avoid empirical controversies surrounding bans' causal effectiveness (whether at deterring hate speech, or at punishing hate speakers).

¹⁷ Brown (*Hate Speech Law*, 249.) acknowledges this worry.

2. The Expressive Argument

According to the expressive argument, a crucial component of what justifies hate speech laws, together with the sanctions they impose, is their expressive or symbolic dimension: roughly, the message they convey. This argument—which has notably been advanced by Lee Bollinger, Larry Kretzmer, David Partlett, Mari Matsuda, Bhikhu Parekh, Anna Galeotti, and more recently Jeremy Waldron and Alexander Brown¹⁸—implicitly relies on an influential strand of legal theory, which presents expressive considerations as central to the function of law and legal sanctions.¹⁹

What message do legal prohibitions—and more specifically, legal prohibitions that impose punitive sanctions—generally express? Legal theorists often suggest that, at the very least, legally prohibiting and sanctioning conduct x (say, by threatening to fine or incarcerate offenders) expresses strong moral disapproval

¹⁸ See note 11 above. Although Bollinger generally recommends tolerating bad speech, he nonetheless holds that legal prohibitions on such speech may sometimes be warranted. And, when doing so, he emphasises the expressive significance of such prohibitions (*The Tolerant Society*, 72–73.).

¹⁹ For overviews, see Matthew Adler, “Expressive Theories of Law: A Skeptical Overview,” *University of Pennsylvania Law Review* 148 (2000): 1363–1501; Richard McAdams, *The Expressive Powers of Law* (Cambridge, MA: Harvard University Press, 2015).

towards *x*.²⁰ This disapproval may be communicated to both the offender and the broader public.²¹

In our context, this suggests that legally prohibiting and sanctioning hate speech conveys strong moral disapproval towards hate speech and the degrading perspective it expresses. Partlett and Parekh are both explicit about this: they defend legislation against racial defamation and ethnic libel, respectively, precisely because such legislation expresses “disapproval” of the regulated utterances.²² In a similar spirit, Galeotti affirms that bans constitute “a public stand against racism which symbolically delegitimizes it”.²³

This is not to say that hate speech laws *only* express disapproval of hate speech. On the contrary, it is often said that, in virtue of condemning hate speech and its degrading message, bans *also* express support for its targets, as well as a

²⁰ This claim is especially widespread in debates about the expressive significance of legal sanctions and punishment. See, e.g., Joel Feinberg, “The Expressive Function of Punishment,” *The Monist* 49 (1965): 400; Igor Primoratz, “Punishment as Language,” *Philosophy* 64 (1989): 188; Dan Kahan, “What Do Alternative Sanctions Mean?,” *University of Chicago Law Review* 63 (1996): 593; Antony Duff, *Punishment, Communication, and Community* (Oxford: Oxford University Press, 2001), 29; Joshua Glasgow, “The Expressive Theory of Punishment Defended,” *Law and Philosophy* 34 (2015): 602; Bill Wringe, *An Expressive Theory of Punishment* (Basingstoke: Palgrave, 2016), 60.

²¹ Wringe, *An Expressive Theory of Punishment*, 57.

²² Partlett, “From Red Lion Square to Skokie to the Fatal Shore,” 469; Parekh, “The Rushdie Affair,” 705. In this paragraph, and the next, I am indebted to Brown’s (*Hate Speech Law*, 241–42) excellent overview.

²³ Galeotti, *Tolerance as Recognition*, 156.

commitment to the egalitarian ideals that hate speech rejects. According to Matsuda, for instance, bans on racist speech are “a statement that victims of racism are valued members of our polity”.²⁴ Similarly, Waldron has prominently argued that hate speech laws assure targets of hate speech of their dignity, which he defines as their good and equal standing in society.²⁵ Thus, while the core message of hate speech laws may indeed consist in strong disapproval of hate speech and the worldview it publicizes, this condemnation arguably implicates other, more positive, messages.

Though all advocates of the expressive argument agree that these various messages play an important role in justifying hate speech laws, they disagree over exactly how important this role is. Many suggest that the expressive dimension of bans constitutes their *primary* source of justification.²⁶ In its strongest form, this position asserts that the expressive argument is sufficient to justify bans.²⁷

However, some adopt a weaker variant of the expressive argument. According to the weaker variant, the expressive argument is better understood as a *supplement*

²⁴ Matsuda, “Public Response to Racist Speech,” 2322.

²⁵ Waldron, *The Harm in Hate Speech*, chap. 4. See also: Kretzmer, “Freedom of Speech and Racism,” 456; Bollinger, *The Tolerant Society*, 122; Parekh, “The Rushdie Affair,” 705.

²⁶ See, e.g., Bollinger, *The Tolerant Society*, 72; Partlett, “From Red Lion Square to Skokie to the Fatal Shore,” 470; Parekh, “The Rushdie Affair,” 156; Galeotti, *Tolerance as Recognition*, 156–57.

²⁷ Partlett and Galeotti come close to this position by asserting, respectively, that the function of bans is “largely” and “mainly” symbolic. See: Partlett, “From Red Lion Square to Skokie to the Fatal Shore,” 473; Galeotti, *Tolerance as Recognition*, 157.

for other justifications, which adds to the overall justification of hate speech laws, and thereby compensates for the limits of other justifications.²⁸ On this second view, then, the expressive function of hate speech laws contributes in a substantial and necessary way to their justification. But it may nonetheless be less normatively important than, say, hate speech laws' deterrence function. Because this second thesis is weaker, it is in principle easier to defend. Nevertheless, the concern I develop in subsequent sections will apply to weak and strong variants alike.

The expressive argument is *prima facie* highly attractive for two reasons. First, it seems to circumvent the longstanding empirical difficulties that plague the deterrence argument. The deterrence argument, recall, holds that bans induce people to refrain from engaging in hate speech. However, reliable evidence supporting this claim remains famously scarce.²⁹ As Brown notably observes, “there is a dearth of useful evidence comparing the extent of hate speech in countries that do possess hate speech law[s] with the extent of hate speech in countries that do not”.³⁰ This scarcity of reliable evidence results partly from methodological obstacles: hate speech is difficult to measure and often goes unreported; different countries and agencies may have different ways of defining and measuring hate speech; and even if we had data reliably comparing the

²⁸ See, e.g., Kretzmer, “Freedom of Speech and Racism,” 489; Brown, *Hate Speech Law*, 241–42.

²⁹ Brown, *Hate Speech Law*, 246; Heinze, “Hate Speech and the Normative Foundations of Regulation,” 607–9.

³⁰ Brown, *Hate Speech Law*, 246.

incidence of hate speech between countries that do and do not ban it, there are so many other cultural, social, and political differences between countries that it would remain extremely difficult to establish a causal connection between bans and reductions in hate speech.

Moreover, the limited evidence that *does* exist is not altogether promising. Eric Heinze, for instance, observes that, despite adopting increasingly punitive and comprehensive bans, some European states have experienced a rise in hate speech.³¹ Likewise, Katharine Gelber and Luke McNamara report that, in Australia, the incidence of hate speech has hardly decreased (and in some contexts has actually increased) since the introduction of hate speech laws.³² Further, although they do find that the language used to express prejudice in newspaper opinion sections has grown more moderate, they concede that this does not necessarily show that hate speech has grown less severe: it may simply be the case that hate speakers have learned to express their degrading views in ‘coded’ ways.³³

In sum, evidence that bans deter hate speech remains scarce, difficult to generate, and contested. *Prima facie*, this empirical problem constitutes a compelling reason for turning to the expressive argument. Insofar as the

³¹ Heinze, “Hate Speech and the Normative Foundations of Regulation,” 609.

³² Katharine Gelber and Luke McNamara, “The Effects of Civil Hate Speech Laws: Lessons from Australia,” *Law & Society Review* 49 (2015): 644–45.

³³ *Ibid.*, 651–654. For discussion of how hate speech can take a coded form that is equally degrading, see Eric Heinze, *Hate Speech in Democratic Citizenship* (Oxford: Oxford University Press, 2016), 145–48.

justification of bans depends on its expressive dimension rather than its ability to deter hate speech, this enables defenders of bans to circumvent the empirical impasse which continues to surround the deterrence argument. Some defenders of the expressive argument make this motivation explicit. As a prelude to his articulation of the expressive argument, for example, Partlett asserts that even draconian bans “will do little to curb th[e] nefarious activities” of “those citizens bent on disseminating racial defamation”.³⁴

However, and secondly, the expressive argument for hate speech laws is not merely attractive because it avoids problems with the deterrence argument. It also seems theoretically and empirically promising in its own right. From a theoretical standpoint, as we have already seen, legal scholars widely recognize that the law has an important expressive dimension, and that laws prohibiting and sanctioning conduct can forcefully express disapproval.³⁵ As for the empirical side, there is evidence that the public statement issued by hate speech laws actually matters to targets of hate speech. Indeed, while Gelber and McNamara struggle to find evidence supporting bans’ deterrent effect, their interview data does yield support for the expressive argument. “The overwhelming view [among groups targeted by hate speech]”, they report, “was that the laws were useful *as a statement in support*

³⁴ Partlett, “From Red Lion Square to Skokie to the Fatal Shore,” 467. See also: Galeotti, *Tolerance as Recognition*, 155–56; Parekh, “The Rushdie Affair,” 705; Kretzmer, “Freedom of Speech and Racism,” 456.

³⁵ See, e.g., Feinberg, “The Expressive Function of Punishment”; Kahan, “What Do Alternative Sanctions Mean?”; Adler, “Expressive Theories of Law”; McAdams, *The Expressive Powers of Law*.

of vulnerable communities.”³⁶ Thus, there are theoretical and empirical reasons to think that hate speech laws do send an important message condemning hate speech and supporting its targets.

Nevertheless, even if hate speech laws can send an important message, the crucial question articulated in the introduction still needs to be addressed. Why couldn’t we send this important message without bans, via the counterspeech that opponents of bans typically advocate? For expressive considerations to give us reasons to supplement counterspeech with bans, it must be shown that the expressive dimension of bans is distinctively effective. In what follows, I will cast doubt on this distinctiveness claim: the considerations adduced in support of it—namely, considerations of expressive strength, directness, and complicity—are either unconvincing, or parasitic on hate speech laws’ deterrent function.

3. Expressive Strength

The first and most obvious sense in which the expressive dimension of hate speech laws could be distinctive concerns its *strength*. On this view, the statement of disapproval that bans convey is somehow stronger, or “more powerful”, than that conveyed by counterspeech.³⁷

³⁶ Gelber and McNamara, “The Effects of Civil Hate Speech Laws,” 655, emphasis added.

³⁷ Bollinger, *The Tolerant Society*, 122. See also: Partlett, “From Red Lion Square to Skokie to the Fatal Shore,” 468; Brown, *Hate Speech Law*, 263.

Although this line of thought is both intuitive and popular, it is sometimes unclear in what specific sense the disapproval expressed by bans is meant to be stronger or more powerful. Building on the observations of advocates of hate speech laws, I will consider two core dimensions of bans' expressive strength: the authority of the speaker (3.1); and the intensity of the disapproval expressed (3.2). In neither respect, I will suggest, are bans genuinely distinctive.

3.1. Authority

The strength or power of an utterance depends importantly on the speaker's status. Accordingly, one might think that the condemnatory message conveyed by bans is distinctively powerful because it is voiced by an agent who is distinctively *authoritative*. As Partlett expresses this idea, legislation prohibiting racist speech has "the imprimatur of authority".³⁸

To have authority, in the present context, roughly means to have high standing or a high social position, in virtue of which one is taken seriously. Authority, in this sense, endows one's speech with power. For one thing, the speech of authoritative agents tends to have greater persuasive power: insofar as having authority just is being taken seriously, listeners are more likely to believe what an authoritative speaker says. But this is not the only kind of power that authority

³⁸ Partlett ("From Red Lion Square to Skokie to the Fatal Shore," 459; see also 467.). In debates about the law's expressive power more generally, see also McAdams, *The Expressive Powers of Law*, 123–24.

lends to speech. Philosophers of language have widely argued that authority is also needed for one's utterances successfully to constitute numerous speech-acts, such as giving orders, enacting norms, issuing verdicts, and so on.³⁹

Now, hate speech laws are enacted by state officials working collectively within legislative and governmental bodies. In contemporary democracies, these officials have immense social standing, partly because they are typically chosen by majority vote in elections where all citizens are enfranchised. Thus, the disapproval bans convey is voiced by an extremely authoritative agent: the state, via its democratically elected legislative and governmental officials.⁴⁰

By contrast, counterspeech is sometimes criticized for lacking authority. In particular, Ishani Maitra and Mary Kate McGowan worry that, because the targets of hate speech generally come from vulnerable social groups, they may lack the standing needed to successfully respond to hate speech. This is especially likely if, as Maitra and McGowan also suggest, hate speech can erode the standing of its

³⁹ For an influential statement of this point, see: Rae Langton, "Speech Acts and Unspeakable Acts," *Philosophy & Public Affairs* 22 (1993): 304–5; and Rae Langton, "The Authority of Hate Speech," in *Oxford Studies in Philosophy of Law, Vol.3*, ed. John Gardner, Leslie Green, and Brian Leiter (Oxford: Oxford University Press, 2018), 125–26. In Raz's terminology, the notion of authority I use resembles "de facto authority" rather than "legitimate authority". Whereas the latter can actually give people reasons to think and act in certain ways, the former is fundamentally about being viewed as providing such reasons. See Joseph Raz, "Authority and Justification," *Philosophy & Public Affairs* 14 (1985): 5–6.

⁴⁰ Partlett, "From Red Lion Square to Skokie to the Fatal Shore," 467.

targets so much that their public utterances are effectively silenced.⁴¹ If this line of thought is correct, it reveals an important sense in which bans express disapproval more strongly than counterspeech: quite simply, the ‘speaker’ enacting bans is far more authoritative than the speaker who performs counterspeech.

The authority argument fails, however, because it relies on an inadequate conception of who should engage in counterspeech. It assumes that private individuals, including vulnerable targets of hate speech, are the sole agents responsible for condemning hate speech. But proponents of counterspeech have increasingly argued that counterspeech should be spearheaded by the state instead. Indeed, according to Corey Brettschneider and Katharine Gelber, it is first and foremost the state’s responsibility to condemn hate speech and to affirm the countervailing ideal of social equality.⁴²

The state, they suggest, can send these messages via numerous empowered agents—the head of government, legislators, and so on—speaking either individually or collectively. Even when they refrain from enacting bans, these agents have a vast array of expressive tools at their disposal. Besides verbally denouncing degrading utterances, they can also (among other things) adopt non-binding resolutions censuring such utterances, implement federal holidays

⁴¹ Ishani Maitra and Mary Kate McGowan, “Introduction and Overview,” in *Speech and Harm*, ed. Ishani Maitra and Mary Kate McGowan (Oxford: Oxford University Press, 2012), 9–10.

⁴² Brettschneider, *When the State Speaks, What Should It Say?*; Gelber, “Reconceptualizing Counterspeech in Hate Speech Policy (with a Focus on Australia).”

honoring civil rights leaders, erect monuments celebrating those leaders, or fund private groups devoted to supporting targets of hate speech.⁴³

Given this ‘state-driven’ conception of counterspeech, considerations of authority cannot establish that bans are expressively stronger than counterspeech. Although the state-based agents who enact hate speech laws are arguably highly authoritative, these same state-based agents can also condemn hate speech through counterspeech.

3.2. Intensity

There is nevertheless a different and more promising sense in which hate speech laws might communicate a stronger or more powerful message than counterspeech. The strength of a message of disapproval depends not only on the authority of the speaker, but also on its *intensity*. Roughly, the intensity of a message of (dis)approval corresponds to the degree of (dis)approval that it expresses.

Authority and intensity can come apart. An authoritative agent can express mild rather than intense disapproval. For instance, a legislator might assert that marijuana use is undesirable, but not so undesirable that it should be coercively suppressed. Conversely, non-authoritative agents often express extremely intense disapproval. Students sometimes engage in hunger strikes, monks sometimes self-immolate, and vice versa, to condemn an oppressive government. What makes hunger strikes and self-immolation such powerful statements is not the high

⁴³ Ibid.

standing of the agents engaging in them. They may, like students and monks, have ordinary or even low status. Rather, their power stems from the extreme degree of moral concern expressed.

Hence, even if the message expressed by state-driven counterspeech is as authoritative as that expressed by bans, one might nevertheless argue that it is less intense: it expresses a lower degree of disapproval than bans. Why might this be? The reason, for some, is that the process of enacting hate speech legislation consumes significant amounts of time, effort, and resources. As Brown explains, “legislative time is in short supply, and [...] drafting law is fraught with difficulty for legislative authorities”.⁴⁴ Consequently, he suggests, the decision to enact legal bans despite these significant costs signals that legislators are very serious in their disapproval of hate speech.

However, this argument fails to establish that hate speech laws are more intense than state-driven counterspeech. The argument would be compelling if counterspeech only involved relatively costless verbal denunciations. But, as mentioned in 3.1, counterspeech can take numerous forms, many of which are comparable in cost to bans. Passing a non-binding legislative resolution censoring hate speech also consumes valuable legislative time and effort. So too does enacting legislation mandating the construction of historical monuments or the creation of new national holidays. In fact, these last forms of counterspeech are in some respects *more* costly than hate speech legislation. Politicians, Strossen

⁴⁴ Brown, *Hate Speech Law*, 263.

observes, “don’t have to raise taxes to censor speech.”⁴⁵ In contrast, creating national holidays and monuments can be extraordinarily expensive, and requires spending valuable tax funds that could otherwise be spent elsewhere. So, if we focus simply on bans’ cost, it is unclear why the condemnation expressed by bans is necessarily more intense than that expressed by state-driven counterspeech.

Still, there is a more promising reason for thinking that bans express a distinctively high degree of disapproval: unlike counterspeech, hate speech laws impose sanctions. Typically, civil hate speech laws can impose significant fines on offenders, while criminal laws can impose both fines and prison sentences. One might argue that, by imposing such punitive sanctions, the state communicates that it disapproves of hate speech in the highest degree.

While defenders of hate speech bans sometimes gesture at this point,⁴⁶ it has been developed most extensively within broader debates about punishment. Legal theorists who argue that punishment plays a crucial expressive role commonly hold that punitive sanctions are unique in their ability to express a high degree of disapproval. As Igor Primoratz notably claims, when expressing severe

⁴⁵ Strossen, “Interview with Nadine Strossen,” 381. See also: Heinze, *Hate Speech in Democratic Citizenship*, 164.

⁴⁶ Partlett, “From Red Lion Square to Skokie to the Fatal Shore,” 468; Bollinger, *The Tolerant Society*, 72.

condemnation, “the necessary seriousness and weight can be secured only by punishment”.⁴⁷

Why are punitive sanctions singularly capable of expressing intense disapproval? Though some theorists assert that there is a natural or essential connection between sanctions and disapproval,⁴⁸ they have struggled to convincingly explain what this natural connection consists in.⁴⁹ Therefore, most legal theorists instead argue that the connection between sanctions and intense disapproval is largely *conventional*.⁵⁰ On this view, imposing fines or incarceration is a uniquely apt way of expressing intense disapproval for the same reason that

⁴⁷ Primoratz, “Punishment as Language,” 200. See also: Feinberg, “The Expressive Function of Punishment,” 400, 420–21; Kahan, “What Do Alternative Sanctions Mean?,” 600–601; Duff, *Punishment, Communication, and Community*, 29; McAdams, *The Expressive Powers of Law*, 124; Glasgow, “The Expressive Theory of Punishment Defended,” 616.

⁴⁸ See, e.g., A.J. Skillen, “How to Say Things With Walls,” *Philosophy* 55 (1980): 517; Primoratz, “Punishment as Language,” 199.

⁴⁹ Hanna, for instance, argues that even if there is *some* natural connection between punitive sanctions and disapproval, this does not entail that punitive sanctions are the *only* way of expressing intense disapproval. Note, however, that nothing will hinge on my rejection of the naturalist explanation. This is because I will later concede for the sake of argument that, whether this is for natural or conventional reasons, punishment may be uniquely capable of expressing intense disapproval. See Nathan Hanna, “Say What? A Critique of Expressive Retributivism,” *Law and Philosophy* 27 (2008): 131–33.

⁵⁰ See, e.g., Feinberg, “The Expressive Function of Punishment,” 402; Kahan, “What Do Alternative Sanctions Mean?,” 600; Glasgow, “The Expressive Theory of Punishment Defended,” 617.

drinking champagne is an apt symbol of celebration: like the meaning of champagne, the meaning of sanctions is simply a product of our social norms.

One might take issue with this convention-based argument for the distinctive expressive intensity of punitive sanctions—and by extension, for the distinctive expressive intensity of hate speech laws. According to a commonly voiced concern, if the meaning of sanctions results from conventions, then their distinctive intensity is entirely *contingent*.⁵¹ After all, conventions emerge and change over time. So, even if people currently take the punitive sanctions involved in hate speech laws to be distinctively intense, our conventions could in principle change so that state-driven counterspeech could come to be seen as equally intense.

But advocates of hate speech laws might respond that this observation is not exceedingly troubling for them. First, even if our conventions are contingent, they may nonetheless be extremely difficult to change deliberately in a particular direction.⁵² Moreover, and more fundamentally, even insofar as we can deliberately change conventions over time, this is consistent with thinking that, *here and now*, punitive sanctions are uniquely capable of expressing intense condemnation.⁵³ This second point matters significantly in the context of debates about hate speech bans. Normative debates about bans are centrally concerned with what we should do in

⁵¹ Heather Gert, Linda Radzik, and Michael Hand, “Hampton on the Expressive Power of Punishment,” *Journal of Social Philosophy* 35 (2004): 86–87; Hanna, “Say What? A Critique of Expressive Retributivism,” 135–48.

⁵² Kahan, “What Do Alternative Sanctions Mean?,” 630.

⁵³ *Ibid.*, 624; Glasgow, “The Expressive Theory of Punishment Defended,” 618.

actual non-ideal conditions, which are marked by substantial bigotry and imperfect agents. And given the conventions that real-world agents actually embrace, punitive sanctions may be the best way to express intense disapproval towards hateful utterances.

The main problem lies elsewhere. The problem is that, even if the disapproval expressed by punitive sanctions *is* distinctively intense, the present argument for thinking that bans are expressively distinctive once more relies on an unduly narrow conception of counterspeech: it assumes that state-driven counterspeech, unlike hate speech laws, cannot involve punitive sanctions.

In fact, however, bans are not the only sanctions-backed tool that the state has for speaking out against the abhorrent perspectives expressed by hate speech. As Post and Strossen note, enacting civil rights legislation more generally—such as anti-discrimination or hate crime legislation—expresses disapproval of racism, xenophobia, and other degrading attitudes.⁵⁴ Relatedly, civil rights legislation is commonly cited as a paradigmatic example of how the law can express a message of social equality and solidarity.⁵⁵ And, crucially, these laws are enforced by punitive sanctions, such as severe fines or incarceration. So, insofar as the intensity

⁵⁴ Post, “Interview with Robert Post,” 26; Strossen, “Interview with Nadine Strossen,” 391.

⁵⁵ Post, “Interview with Robert Post,” 26; Elizabeth Anderson and Richard Pildes, “Expressive Theories of Law: A General Restatement,” *University of Pennsylvania Law Review* 148 (2000): 1533–45.

of the disapproval expressed by bans derives from their deployment of sanctions, state-driven counterspeech too can presumably achieve this intensity.

One might worry that, although civil rights legislation does express intense disapproval of some kind, it does not express disapproval of the degrading perspectives expressed by hate speech. As discussed in Section 2, legal theorists typically suggest that legal prohibitions express disapproval *of the prohibited conduct*. If so, then what hate crime legislation asserts, in the first instance, is that prejudice-motivated crimes are deeply wrong. Likewise, anti-discrimination law seems to say, first and foremost, that treating people disadvantageously on the basis of their social group membership is deeply wrong. Strictly speaking, this is not the same as asserting that degrading or equality-denying perspectives are wrong or incorrect.

Nonetheless, this worry is ultimately unproblematic. It suggests that civil rights legislation does not assert the wrongness of the degrading perspectives expressed by hate speech. But what an utterance communicates does not reduce to what it asserts. In particular, our utterances can also *implicate* contents. Very roughly, the implicated content of an utterance is content that is not asserted, but that the speaker must nevertheless be committed to for their utterance to make sense.⁵⁶ If I assert *that Tom and Jane are getting a divorce*, for instance, I thereby implicate *that they*

⁵⁶ Andrei Marmor, *The Language of Law* (Oxford: Oxford University Press, 2014), chap. 2.

have previously been married. Importantly, Andrei Marmor argues, the observation that speech can implicate contents also applies to legal speech.⁵⁷

Thus, we can appreciate an important sense in which civil rights legislation does express disapproval of the perspectives expressed by hate speech: even if it does not assert that degrading or equality-denying perspectives are wrong, it nevertheless strongly implicates their wrongness. Consider again hate crime legislation. In inflicting particularly severe sanctions on crimes that are motivated by prejudicial attitudes (including, say, racial or xenophobic animosity) hate crime legislation does not merely assert that such crimes are deeply wrong. It also strongly *implicates* that such prejudicial attitudes are incorrect. Indeed, unless they were incorrect, it would be difficult to make sense of the fact that, other things being equal, prejudice-motivated crimes incur greater sanctions than other crimes. Similarly, anti-discrimination law does not merely assert that it is wrong to treat people disadvantageously simply because of their social group membership. In saying that such conduct is wrong, it also implicates the inadequacy of hateful perspectives that legitimize discriminatory conduct. After all, if perspectives that recommend or require discrimination were correct, it would be hard to argue that discrimination is wrong.

In sum, considerations of intensity fail to establish that bans condemn the degrading perspectives expressed by hate speech in a way that is distinctively intense. The most compelling basis for thinking so is that bans impose punitive

⁵⁷ Ibid.

sanctions. Yet this overlooks the existence of civil rights laws that use punitive sanctions to condemn degrading perspectives, but do so without banning hate speech. Like the authority argument, then, the intensity argument overlooks the full potential of state-driven counterspeech.

To establish this conclusion, however, I suggested that the most intense forms of state-driven counterspeech condemn hateful perspectives by implication. This suggests a different argument for thinking that bans are expressively preferable: even if the disapproval expressed by state-driven counterspeech can in principle be as intense and *strong* as that expressed by bans, it is nevertheless importantly less *direct*. Therefore, in the following section, I will investigate whether considerations of directness can establish the expressive superiority of bans over counterspeech.

4. Expressive Directness

Considerations of directness might motivate two contrasting arguments for the expressive distinctiveness of hate speech bans. The first, which I advertised above, holds that the strongest forms of counterspeech are *insufficiently* direct compared to bans (4.1). The second, by contrast, maintains that counterspeech is in another respect *too* direct compared to bans (4.2). Neither, I will suggest, is ultimately compelling.

4.1. The content of the disapproval

In response to the argument that the disapproval expressed by bans has a distinctive intensity—and therefore, a distinctive strength—I argued that other forms of

legislation, such as civil rights legislation, can express disapproval with a similar intensity. Even so, one might worry that their condemnation is insufficiently direct: *what these laws disapprove of*, their content, is not directly the degrading perspectives expressed in hate speech.

To reiterate, what hate crime and anti-discrimination legislation directly express or assert is disapproval of crimes motivated by group-based prejudice and discriminatory conduct, respectively. *By implication*, I suggested in 3.2, they convey disapproval of degrading or hateful perspectives. In this light, the form of state-driven counterspeech that is arguably greatest in intensity may seem too indirect. It would be preferable, one might think, for the state to straightforwardly say or assert that it intensely disapproves of degrading perspectives than to simply implicate its intense disapproval.

In reply, one might question whether this kind of directness would really be preferable. In 4.2, we will encounter a reason for thinking that highly visible or overt condemnations of hateful utterances may be undesirable. Insofar as the asserted content of an utterance may be more visible than its implicated content, this may constitute a reason for rejecting the present argument. For now, however, let us assume that it really would be better, holding the intensity and authority of disapproval fixed, for disapproval of hateful perspectives to be asserted rather than merely implicated.

The more fundamental problem is that it is also true of hate speech laws that they indirectly condemn the degrading perspectives expressed by hate speech. Legal prohibitions, we have been assuming, assert or directly express the

wrongness of the conduct that they prohibit. Now, hate speech laws do not prohibit hateful or degrading perspectives themselves. Rather, they prohibit the *utterance or expression of* a degrading perspective. So, in the first instance, they assert the wrongness of expressing degrading perspectives. It is only by implication that this statement condemns degrading perspectives themselves. Indeed, it would be difficult (though not impossible)⁵⁸ to explain why it is wrong to express a degrading or equality-denying perspective, if not for the fact that this perspective is profoundly misguided.

In this respect, hate speech and hate crime legislation seem closely analogous. Both prohibit a kind of conduct that is based on a degrading perspective. Saying something (in the case of hate speech laws) or committing a crime (in the case of hate crime laws). And both, in consequence, strongly implicate that the perspective in question is wrong.

Still, advocates of the expressive argument for bans might reply that this misses the obvious point. Even if hate speech laws and civil rights legislation are equally direct in expressing disapproval of degrading or hateful views, there is at least one

⁵⁸ Below, I describe a case where the state disapproves of hateful utterances *while* embracing the perspective they express. Note that this does not undermine my claim that condemning hate speech implicates rejecting the perspective it expresses. It is possible to implicate something without entailing it. This is why, as Davis explains, some kinds of implicature are said to be ‘cancellable’. See: Wayne Davis, “Implicature,” *Stanford Encyclopedia of Philosophy*, 2014, <https://plato.stanford.edu/entries/implicature/>.

thing that bans obviously disapprove of more directly: the expression of degrading perspectives.

This is undeniably true, but does not seriously threaten my argument. At bottom, what needs to be condemned is not so much *the expression* of degrading perspectives, as the degrading perspectives themselves. To see this, imagine a scenario where the state explicitly asserts that it disapproves of the former but not the latter. Perhaps, say, the state disapproves of expressions of degrading perspectives because, though it considers these perspectives to be accurate, it worries that publicly expressing them would produce social unrest. Morally speaking, such a denunciation of hate speech seems entirely inadequate: it intuitively fails to counter what is actually bad about hate speech. Hence, condemnation of hateful utterances hardly seems desirable when it is divorced from condemnation of the content of those utterances.

This, in turn, should not be surprising. The harmfulness of hate speech depends fundamentally on the degrading contents it expresses. In other words, the vilifying views and degrading attitudes that hate speech expresses are crucial to explaining why and how hate speech tends to produce harms. For example, the fact that hate speech expresses the exclusion or inferiority of its targets is crucial to explaining why it undermines those targets' assurance of their good social standing. Likewise, the fact that some hate speech depicts its targets as, say, vicious or worthless is central to explaining why it risks inciting animosity towards them. This relation of dependence lends support to my above suggestion: if hateful utterances are harmful essentially in virtue of the abhorrent perspectives they express, what seems

fundamentally important when voicing opposition to hateful utterances is that we challenge those abhorrent perspectives.

Notice that this last point forestalls a potential worry with the thought-experiment outlined above. The thought-experiment suggests that the condemnation of hate speech is not desirable when it is divorced from condemnation of the degrading perspectives hate speech expresses. But this is consistent with thinking that, when we *are* condemning these degrading perspectives, condemning their expression as well adds value to our condemnation. Why might this be? Perhaps, one might think, because expressing degrading perspectives gives rise to harms that would not arise if people simply held but did not voice degrading perspectives.

But even if expressing degrading perspectives generates harms that would not otherwise arise, I have suggested that these harms arise in virtue of, and are explained by, the content of the degrading perspectives expressed. Because of this fundamental dependence, we can adequately condemn these harms by condemning the hateful views that underpin and explain them. To put this slightly differently: if the abhorrent perspectives that hate speech expresses ground and explain the wrongness of hate speech—that is, if these perspectives are at the heart of that wrongness—then condemning these perspectives gets to the heart of the matter.

To summarize, condemning hate speech seems desirable to the extent that it contributes to condemning the degrading perspectives expressed by hate speech. Now, I have argued that hate speech bans and civil rights laws are equally indirect in their condemnation of the degrading perspectives expressed by hate speech. So,

even if bans are more direct in their condemnation of hateful utterances, they remain equally indirect where it truly matters.

The more general upshot is this: when it comes to the content of their disapproval—what they disapprove of—hate speech laws do not seem importantly more direct than the most intense forms of state-driven counterspeech.

4.2. The overttness of the disapproval

An alternative ‘directness’ argument, which Waldron has influentially articulated, takes a very different tack. It suggests that, in a different respect, hate speech bans are appropriately *less* direct than counterspeech. On this view, even if the disapproval expressed by bans and counterspeech has the same content—what they disapprove of is the same—bans express that disapproval in a way that is less *overt*.⁵⁹ Their message, Waldron claims, is a “low-key background thing”.⁶⁰

The overttness of a statement refers roughly to its visibility. The same statement can be communicated in more or less visible ways. Consider, by way of illustration, how Tom’s friends might express their view that Tom’s nose looks normal. They might post banners throughout Tom’s school, affirming: ‘Tom’s nose is perfectly normal.’ Alternatively, one friend might discretely whisper in Tom’s ear: ‘Say, your nose looks perfectly normal.’ Or, finally, they might communicate their view

⁵⁹ Waldron, *The Harm in Hate Speech*, 87–88, 93–96.

⁶⁰ *Ibid.*, 93.

via an omission,⁶¹ by never mentioning Tom's nose—and hence, never raising the possibility that it might be anything other than normal.

Counterspeech, as Waldron envisages it, is highly overt. It assures citizens of their good standing by publicly and visibly denouncing hateful views, and publicly and visibly affirming social equality instead.⁶² This is congruent with the picture of state-driven counterspeech sketched above, which involves having high-profile politicians condemn hate speech, erecting public monuments, celebrating national holidays, and so on.

By contrast, Waldron repeatedly emphasizes that, ideally, the expressive function of bans operates in a way that is “silent” or “implicit”.⁶³ By either eliminating hate speech or “driving [it] underground”,⁶⁴ bans prevent such speech from publicly stating that some members of society are inferior or unwanted. Thus, instead of overtly telling members of vulnerable groups that they are not inferior and that they are in good standing, bans convey this assurance by silencing claims to the contrary. In this, they resemble the strategy of conveying that Tom's nose is normal by never mentioning it: in a similar fashion, bans convey the social equality of citizens by keeping their standing out of the spotlight.

⁶¹ For discussion of how omissions can perform communicative speech-acts, see Eric Swanson, “Omissive Implicature,” *Philosophical Topics* 45 (2017): 117–37.

⁶² Waldron, *The Harm in Hate Speech*, 87.

⁶³ *Ibid.*, 87, 88, 92, 93, 94, 96.

⁶⁴ *Ibid.*, 95–96.

For Waldron, this difference matters greatly. The issue with overtly condemning hateful views and assuring citizens of their good standing, he suggests, is that doing so is “evidence of a problem”.⁶⁵ In visibly disputing hate speech’s attack on the standing of its targets, counterspeech makes it apparent that their standing is *in dispute*. And insofar as one’s good standing is publicly in dispute, that standing appears less robust. So, an overt message of opposition to hate speech seems partly self-undermining, in the same way that a banner stating ‘Tom’s nose is perfectly normal’ would be. Just as such a banner might inadvertently draw people’s attention to Tom’s nose and make them wonder whether it is indeed normal, so too overtly assuring a group that they are in good standing might draw attention to their standing and make them question whether it is truly secure.⁶⁶

This is why, for Waldron, the implicit expressive function performed by bans is “tremendously important.”⁶⁷ By keeping degrading perspectives out of view, bans avoid conveying the impression that the status of some social groups is in dispute. Instead, they convey the impression that social equality is broadly uncontroversial. Waldron considers this perceived lack of controversy to be crucial: where social equality “can be taken for granted, [...] people who might otherwise

⁶⁵ Ibid., 87.

⁶⁶ Ibid., 88.

⁶⁷ Ibid., 87.

feel insecure, unwanted, or despised in social settings can put all that terrible insecurity out of their minds”.⁶⁸

There is something importantly right about Waldron’s concern that the overtness of counterspeech can defeat its purpose. But the claim that bans avoid this concern seems far more problematic. To begin, there are reasons to think that hate speech laws actually do operate in a highly overt manner. First, insofar as they are publicly enacted by authoritative agents, such laws—and the message they express—tend to be highly salient.⁶⁹ Second, bans often lead to widely publicized trials, where hate speakers’ degrading perspectives are reiterated and amplified.⁷⁰ Finally, Waldron’s argument for thinking that bans contribute non-overtly to upholding a message of social equality is that they suppress public challenges to social equality. But bans suppress or prevent hate speech in virtue of being seen and heeded by prospective hate speakers. So, the case for thinking that bans are not overt paradoxically depends on bans being highly visible.

Still, let us assume, for the sake of argument, that bans are nonetheless comparatively less overt than counterspeech. Perhaps, in some contexts, bans are so effective at getting people to refrain from hate speech that criminal trials prosecuting hate speakers are rare. Moreover, one might emphasize that bans’

⁶⁸ Ibid., 88.

⁶⁹ McAdams, *The Expressive Powers of Law*, 123–24.

⁷⁰ See, e.g., Simpson, “Dignity, Harm, and Hate Speech,” 724; Heinze, “Hate Speech and the Normative Foundations of Regulation,” 600; Gelber and McNamara, “The Effects of Civil Hate Speech Laws,” 656–57.

ability to deter hate speech depends on their visibility *to potential hate speakers*. And this, one might think, is not a problematic kind of visibility. Rather, bans' visibility *to potential targets of hate speech* is what risks making targets worry that their good standing is in dispute.

But even if we grant these responses, they highlight a second key problem: the argument for thinking that bans are expressively less overt than counterspeech hinges on bans' success at deterring hate speech. Indeed, bans contribute to assuring citizens of their good standing in a distinctively 'silent' way *only insofar* as they successfully eliminate hate speech from public discourse. Therefore, the present reasons for thinking that bans are expressively distinctive—and, by extension, for thinking that the expressive argument succeeds—are parasitic on the reasons for thinking that the deterrence argument succeeds.

This dependence on the deterrence argument is troubling not just for the claim that the expressive argument is the primary justification for bans, but also for the weaker claim that it is a necessary supplement to other justifications. First, if the expressive argument depends on the success of the deterrence argument, then it risks being justificatorily redundant, rather than necessary: hate speech laws will be expressively distinctive only insofar as hate speech and its attending harms are already suppressed.

Second, if the expressive argument works only insofar as bans deter hate speech, then this argument no longer allows us to circumvent—even in part—the protracted empirical dispute regarding whether bans successfully deter hate speech. This constitutes a major blow: as discussed in Section 2, the expressive argument's

apparent ability to sidestep the deterrence argument's longstanding empirical concerns was one of its central appeals.

More generally, this suggests that considerations of directness fail to establish the expressive superiority of bans. The arguments canvassed in this section struggle to identify a meaningful difference in the expressive contents of bans and counterspeech (4.1) and in the overtness with which they express their contents (4.2). And even to the extent that considerations of overtness might favor bans, they make the expressive argument depend wholly, and problematically, on bans' ability to deter hate speech.

5. Expressive Complicity

I have examined and criticized two sets of considerations for thinking that hate speech laws are expressively superior to counterspeech: considerations of expressive *strength* and considerations of expressive *directness*. Yet even if state-driven counterspeech can in principle match the expressive strength and (in)directness of bans, there is a final potential reason for preferring bans: the reason is that refusing to ban hate speech might send a countervailing message of *complicity* with hate speakers and their perspectives.

In liberal democracies that honor freedom of expression, citizens typically have the right to participate in public deliberation and to engage in public protests, where they voice their opinions and perspectives. In virtue of respecting this right, the state is committed to tolerating public utterances and protests, and to protecting

these public actions from those who would coercively suppress them. Now, in contexts where hate speech is not banned, the state's commitment to toleration and protection extends to hate speech. Besides refraining from coercively interfering with hate speech, this might also involve, say, deploying police officers to prevent the intimidation of hate speakers.⁷¹ For instance, during the infamous 1978 neo-nazi demonstration in Chicago's Marquette Park, substantial police forces were posted to keep counter-protestors away from neo-nazis. This state-provided tolerance and protection, it is sometimes said, expresses a form of complicity with hate speakers. As Matsuda famously puts it, tolerating and protecting hate speech constitutes "a statement of state authorization".⁷²

If this is correct, then banning hate speech may be expressively preferable to simply engaging in state-driven counterspeech. Even if state-driven counterspeech itself could in principle condemn hateful views as effectively as bans, the state toleration and protection of hate speech sends a countervailing message. Accordingly, in the absence of bans, even a state that engages aggressively in counterspeech ultimately sends a mixed message. By contrast, since a state that bans hate speech is by definition committed to not tolerating or protecting hate speech, the condemnation it expresses is not muddled in this way.

The key question, then, is whether state protection and toleration of hate speech really does send out a problematic message. To adjudicate this issue, we need a

⁷¹ Matsuda, "Public Response to Racist Speech," 2375.

⁷² Ibid., 2378. See also: Brown, *Hate Speech Law*, 263; Galeotti, *Toleration as Recognition*, 156.

clearer picture of what that message might consist in. At its strongest, the ‘complicity’ argument holds that tolerating and protecting hate speech sends a message of *endorsement* of hate speech and its contents. According to Matsuda, for example, protecting the Ku Klux Klan’s speech “means that the state is promoting racist speech”.⁷³ Indeed, she continues, doing so “carries a strong implication that racist activities are supported”.⁷⁴

However, this version of the argument seems too strong. As part of their commitment to freedom of religion, liberal democratic states typically also tolerate and protect people’s right to practice numerous different and conflicting religions. If the mere fact that the state permits and protects conduct implicates that the state endorses this conduct, it would follow that liberal democratic states typically endorse this vast array of religions. But this seems incorrect, for two reasons. First, states often explicitly disagree with, rather than endorse, the tenets of religions that they protect. For example, many states allow and protect religions that forbid members from using contraception or from having abortions, yet are deeply committed to protecting citizens’ rights to have abortions and to use contraception. Moreover, if the state endorsed or promoted all of the conflicting religions that it protects, it would be committed to a bewildering range of inconsistencies. Intuitively, however, the state’s protection of freedom of religion does not commit

⁷³ Matsuda, “Public Response to Racist Speech,” 2378.

⁷⁴ *Ibid.*, 2379.

it to glaring contradictions. In this light, it seems incorrect to think that, in and of itself, the state's protection of hate speech commits it to endorsing hate speech.

Nevertheless, there is a weaker and more plausible way of construing the complicity argument. On this view, the state's toleration and protection of hate speech does not necessarily communicate that it positively endorses or supports hate speech and its contents. But it does communicate that the state only disapproves of hate speech to a limited extent. As Brown suggests, "the mere fact that the state has opted to refrain from legislating against hate speech may send out the message to citizens that it is not as serious about its anti-hate speech message as it purports to be."⁷⁵

Why might this be? An influential justification for tolerating hate speech is that, although hate speech is morally undesirable, it is less undesirable than its suppression would be. This might be, for example, because suppressing hate speech infringes hate speakers' autonomy⁷⁶ or undermines democratic procedures.⁷⁷ Such an approach to justifying the toleration of hate speech, Brown might observe, implies a limit to how "seriously" one disapproves of hateful utterances. After all, this approach is premised on the claim that the reasons for disapproving of hate speech are *outweighed* by other moral reasons.⁷⁸ Consequently, when the state

⁷⁵ Brown, *Hate Speech Law*, 263.

⁷⁶ E.g., Scanlon, "A Theory of Freedom of Expression"; Baker, "Harm, Liberty, and Free Speech."

⁷⁷ E.g., Robert Post, *Constitutional Domains: Democracy, Community, Management*, Harvard University Press (Cambridge, MA, 1995); Heinze, *Hate Speech in Democratic Citizenship*.

⁷⁸ For a similar thought, see Matsuda, "Public Response to Racist Speech," 2377–78.

decides to tolerate and protect hate speech, this might suggest or implicate that there is a limit to how much the state disapproves of hate speech and the perspective it expresses. By contrast, enacting bans arguably does not implicate any limit to the state's disapproval of hate speech.

Although this revised complicity argument is more compelling, it still fails to establish that bans are expressively preferable to counterspeech. To see why, two clarifications are needed. First, while the state's decision to tolerate and protect hate speech may indeed suggest or implicate that the state's disapproval of hate speech is limited, it clearly does not *entail* such a limit. This is because there are many other possible reasons for refraining from banning hate speech, which have nothing to do with the intensity of one's disapproval. For instance, one might believe that the coercive suppression of hate speech is likely to backfire and incite more hateful utterances; that hate speech laws are likely to be used to suppress non-hateful speech; or simply that there are more effective ways of combatting hate speech.

Second, the fact that the state tolerates hateful utterances does not mean that it does nothing else. Matsuda sometimes obscures this point by running together the idea that the state protects hate speech with the idea of "state silence".⁷⁹ But these two ideas are very much distinct. As we have seen, the most compelling version of the anti-ban position holds that, while the state should protect the expression of

⁷⁹ Ibid., 2378.

hateful utterances, it should simultaneously engage in robust counterspeech that intensely and authoritatively condemns those perspectives.

Put together, these two points pose a problem for the revised complicity argument. State protection of hate speech only suggests or implicates—but it does not entail—that the state’s disapproval of the perspectives expressed by hate speech is limited. Now, if state protection of hate speech is merely suggestive of limited disapproval, and can be explained in other ways, it is unclear why we cannot dispel or cancel this problematic suggestion.⁸⁰ In other words, it seems *prima facie* possible to disambiguate the meaning of state protection by specifying which of its possible explanations is correct. This, in turn, is precisely what the state-driven counterspeech that accompanies state protections of hate speech is intended to do. Insofar as it expresses authoritative (3.1) and unreservedly intense (3.2) disapproval for hateful utterances, counterspeech rules out one possible explanation for the protection of hate speech: namely, that the state does not unreservedly disapprove of its message. Additionally, to further disambiguate the meaning of its protections, state-driven counterspeech can also explicitly articulate *why* the state is protecting hate speech. For example, it can clarify that it tolerates hate speech because it believes that the most effective way of discrediting hateful worldviews is to publicly expose and condemn them.

⁸⁰ This is congruent with philosophical discussions of implicature. As mentioned in note 58, some kinds of implicature—namely, conversational implicatures—are ‘cancellable’ by context or further utterances.

To salvage the complicity argument in the face of this problem, its proponents must reject the expressive effectiveness of state-driven counterspeech. If state-driven counterspeech really can express the state's unreserved disapproval effectively—and, in particular, if it can do so as effectively as bans would have—then it can rule out the 'limited disapproval' explanation of state toleration, in favor of an alternative explanation. So, the complicity argument implicitly depends on the claim that state-driven counterspeech is not effective at expressing unreserved disapproval—or, at the very least, that it is less effective than bans.

The problem is that, in the present context, this response begs the question. What proponents of the expressive argument for hate speech laws are trying to establish is that counterspeech is less effective than bans at expressing condemnation. The complicity claim—that refusal to ban hate speech signals complicity with hate speakers—was introduced to justify this thesis. As we have just seen, however, this complicity claim is compelling *only if* counterspeech is less expressively effective than bans. So, the complicity argument ultimately begs the question: it implicitly relies on the conclusion it aims to establish.

Thus, to vindicate the complicity argument for bans, we would first need to identify respects in which the expressive dimension of bans is distinctively effective compared to counterspeech. But whether there are any such respects is precisely what we have been investigating all along, with little success.

6. Conclusion

Can expressive considerations give us reasons to supplement counterspeech with hate speech laws? By comparing the expressive strength, directness, and complicity of bans and counterspeech, I have argued that such reasons are in fact highly elusive.

First, given an appropriately expansive understanding of counterspeech, the disapproval counterspeech conveys can be comparable in authority and intensity—the main determinants of expressive strength—to the disapproval conveyed by bans (Section 3). It is also unclear in what sense considerations of directness might tell in favor of bans: both the content and the overtness of hate speech laws' message are difficult to distinguish meaningfully from the content and overtness of counterspeech. And, more importantly, even insofar as bans may be desirably less overt than counterspeech, this is entirely in virtue of their success at deterring hate speech (Section 4). Finally, the view that allowing and protecting hate speech sends a message of complicity to hate speakers is unhelpful in our context. It cannot contribute to establishing that bans are expressively superior to counterspeech because, at bottom, it presupposes this claim (Section 5).

Thus, attempts at showing that bans are expressively distinctive either overlook the full expressive potential of counterspeech, or are parasitic on the success of a separate argument for bans, the deterrence argument. This conclusion, I have argued, is problematic for two reasons. First, it threatens to make the expressive argument for hate speech laws redundant: at best, it succeeds only insofar as hate

speech is already suppressed. Second, if the success of the expressive argument depends on the success of the deterrence argument, then, contrary to what is often assumed, it does not allow us to avoid the longstanding empirical challenges that stand in the way of establishing that bans deter hate speech.

What does this mean for hate speech laws? The broader upshot is not necessarily that they are unjustified. It is, instead, that to justify such laws, its advocates should focus less on their symbolic importance, and more on their causal ability to suppress hateful utterances. Even if unearthing systematic evidence that bans deter hate speech is extremely difficult, the shortcomings I have diagnosed with the expressive argument suggest that doing so nonetheless constitutes a more promising, and a more fundamental, justificatory strategy.

Notice, finally, that my argument has revisionary implications not only for the justification of hate speech laws, but also for the practice of counterspeech. Although I have argued that counterspeech can largely disapprove of hate speech as powerfully as bans, my argument also highlights that not just any kind of counterspeech will do. To match the expressive strength of bans, counterspeech must be driven not merely by private actors, but also by the state; and it must include not only verbal denunciations, but also a host of expensive and taxing measures, such as erecting historical monuments and enacting aggressive civil rights legislation. What this shows is that rejecting the expressive argument for hate speech laws is not a license for expressive *laissez-faire*: on the contrary, engaging with the expressive argument helps bring into view the arduous and costly efforts that speaking out against hate speech requires.