Hate Speech in Public Discourse:
A Pessimistic Defense of Counterspeech

Maxime Lepoutre

Abstract: Public hate speech, it has forcefully been argued, assaults the dignity of its targets. Without denying this claim, I contend that it fails to establish that bans, rather than counterspeech, are the appropriate response. By articulating a more refined understanding of counterspeech, I suggest that counterspeech constitutes a better way of blocking hate speech's dignitarian harm. In turn, I address two objections: according to the first, which draws on contemporary philosophy of language, counterspeech does not block enough hate speech; according to the second, counterspeech blocks too much speech. Although these objections should qualify our optimism regarding counterspeech, I demonstrate that each can be turned, with even greater force, against hate speech bans.

Keywords: hate speech, dignity, free speech, democracy, deliberation, philosophy of language, speech acts

1. Introduction

1.1. *Charlie Hebdo*: Three Responses

On January 7th, 2015, the headquarters of *Charlie Hebdo*—a satirical French newspaper frequently charged, despite its self-description as ironic, with diffusing racist depictions of Muslims[1]—were attacked by Islamists. Twelve were killed, eleven injured. Reactions were varied. The most publicized response, spearheaded by the French state's '*Je suis Charlie*' campaign, was to honor the newspaper as embodying the ideal of free speech, fearless and uncompromising.

But not everyone felt this way. Dieudonné M'bala M'bala, a highly controversial French comedian, wrote a Facebook post stating "*Je me sens Charlie Coulibaly*" ("I feel like Charlie Coulibaly"). The message parodied the '*Je suis Charlie'* slogan by referring to Amedy Coulibaly, a gunman closely connected with the *Charlie Hebdo* attack. Accordingly, Dieudonné was widely understood as expressing sympathy with the attackers and their viewpoint. In consequence, Dieudonné received a suspended two-month prison sentence.

A final public reaction, reproduced below, was published in *The Guardian* by Joe Sacco, a journalist and cartoonist renowned for his work in wartime Yugoslavia:[2]

[1] Another possible defense of *Charlie Hebdo* against charges of racism claims that it subjects all of its targets to such satire. This defense remains contested. For criticism by a former *Charlie Hebdo* journalist, see Olivier Cyran, "'Charlie Hebdo', Pas Raciste? Si Vous Le Dites...," *Article 11*, 2013, http://www.article11.info/?Charlie-Hebdo-pas-raciste-Si-vous. My argument, however, will not hinge on whether some of *Charlie Hebdo*'s issues really were racist or vilifying. For present purposes, all that matters is that, in context, it was understood by very many people as promoting vilifying stereotypes. In fact, in Section 5, I return to this example and suggest that the *vagueness* surrounding whether or not it constituted hate speech is no accident, and that this should inform what policy we adopt regarding hate speech.

[2] "On Satire" was first published on *The Guardian* website. Reprinted by permission of Joe Sacco and Aragi Inc. See Joe Sacco, "On Satire" *The Guardian*, 2015, http://www.theguardian.com/world/ng-interactive/2015/jan/09/joe-sacco-on-satire-a-response-to-the-attacks.

Consider several significant features of Sacco's response. First, it points to two groups whose viewpoints and utterances seem extreme, and (arguably) in tension with the commitment to treating others as free and equal citizens: on the one hand, the writers of *Charlie Hebdo*; on the other, those who can't "take the joke" and, in some cases, follow Dieudonné in thinking that the writers got what they deserved. Second, it expresses support for protecting speech even when that speech, like *Charlie Hebdo*'s, may represent abhorrent views. Likewise, it voices reservations about "driving into the sea"—or forcefully excluding—those worldviews that might be broadly sympathetic to Dieudonné's post. Instead, and finally, it calls for

dialogue aimed at criticizing wrongful views, but also at understanding where they come from, and why many have come to find them appealing.

I believe there is something importantly right here, which Sacco poignantly expresses, about democratic public discourse in contexts where some participants possess degrading or hateful viewpoints. In this article, I wish to consider what should be done about the public expression of such hateful viewpoints, or hate speech. An adequate response, I want to suggest, should capture the intuitive appeal of Sacco's reaction. So let us bear it in mind.

1.2. Defining Hate Speech

The definition of 'hate speech' is highly contested. By 'hate speech', I will refer, roughly, to communications that emphatically deny the basic status of other members of society as free and equal citizens. More specifically, I follow most discussions in taking hate speech to deny the basic standing of individuals who belong to vulnerable social groups, in virtue of their belonging to these groups.[3] Two things are worth stressing here. First, hate speech includes not just verbal speech, but all platforms (including printed images) that might be used to communicate these propositional contents. Second, what makes a communication hate speech is not the speaker's emotion, but rather the communication's content.

This abstract characterization is intended only as an approximate guide or heuristic. To make matters clearer, consider a few illustrations of the kinds of speech I am concerned with. Hate speech might include newspaper articles falsely attributing essential or inherent dangerousness to a minority group (e.g., 'Muslims are terrorists') or leaflets portraying group members as subhuman or bestial (e.g., depictions of blacks as ape-like). Sometimes, hate speech more explicitly expresses the exclusion of minority group members from the community of citizens, as in Pegida's infamous posters asserting 'Rapefugees Not Welcome'.

My discussion will focus on hate speech occurring in public discourse within relatively stable democracies, like the US or France.[4] I understand 'public discourse' quite broadly. It encompasses communications occurring in formal political debates, but also those published in newspapers, posted on the internet, or plastered in common spaces. Very often, Waldron notes, such communications become "a permanent or semipermanent part of the visible environment".[5] Moreover, this is commonly speech that is not simply directed at the targets of the hate speech's content, but addresses a broad audience, which includes third parties. Hence, although the distinction between public and non-public speech remains vague, this restriction tends to exclude live, face-to-face encounters where, for example, the speaker directs a slur at a particular person with the deliberate intention to wound, threaten, or intimidate.[6]

---

[3] This definition is largely inspired by Waldron and Brettschneider. In legal scholarship, it echoes the UN's condemnation of speech "based on ideas or theories of superiority of one race or group of persons of one colour". See: Jeremy Waldron, *The Harm in Hate Speech* (Cambridge, MA: Harvard University Press, 2012), pp. 27-28, 34–41; Corey Brettschneider, *When the State Speaks, What Should It Say?* (Princeton, NJ: Princeton University Press, 2012), p. 1; UN, *International Convention on the Elimination of All Forms of Racial Discrimination*, 1965.

[4] For more extensive discussion of these restrictions, see Eric Heinze, "Hate Speech and the Normative Foundations of Regulation," *International Journal of Law in Context* 9 (2013): 590-617, p. 591, and Eric Heinze, *Hate Speech and Democratic Citizenship* (Oxford: Oxford University Press, 2016), pp. 26-30, 69-81.

[5] Waldron, *The Harm in Hate Speech*, p. 37.

[6] Many authors also bracket such face-to-face assaultive forms of hate speech, or treat them separately. For instance: Heinze, "Hate Speech and the Normative Foundations of Regulation," pp. 590–91; Rae Langton, "Beyond Belief: Pragmatics in Hate Speech and Pornography," in Ishani Maitra and Mary Kate McGowan (eds.) *Speech and Harm* (Oxford: Oxford University Press, 2012), pp. 72-93 at pp. 76-77; Waldron, *The Harm in Hate Speech*, pp. 37–38; Brettschneider, *When the State Speaks, What Should It Say?*,

I focus on public discourse for two reasons. The first is dialectical: I will chiefly be engaging with Jeremy Waldron, who takes his argument for banning hate speech to be most compelling with respect to such public discourse. The second is political. The political contrast that is often drawn is between the US, which tends to oppose hate speech bans, and the rest of the world, which tends to embrace them. But even in the US, live, face-to-face "fighting words", as well as targeted threats and harassment, are generally prohibited.[7] So the main debate concerns not *whether*, but *which* kinds of hate speech should be prohibited. In this context, hate speech bans on public discourse are politically more divisive.

This characterization might still seem too broad. Many actual hate speech regulations require not just that the expressed content of the regulated speech profoundly contradict the basic standing of vulnerable groups, but also that it be uttered in circumstances such that it is likely to produce certain effects, like violence or antipathy.[8] Now, there is enormous variation in characterizations of hate speech. In part, this variation is due to the fact that how we define hate speech depends on what kinds of speech-related harms we are concerned with. Accordingly, the reason I adopt this relatively broad account of hate speech is that it is adapted to the particular speech-related harm I will be focusing on (see Section 2).

1.3. Outline

What should be done about hate speech occurring in public discourse within stable democracies? Should we implement legal bans which coercively forbid the public expression of hate speech—that is, should the law threaten to forcibly impose penalties (such as fines or incarceration) on those who publicly engage in hate speech? Or should we tolerate them, and try to address the harm they produce internally to democratic discursive processes? The relevant literature is vast, and spans across philosophical and non-philosophical domains. Consequently, in addressing these questions, I will primarily engage with Waldron's influential defense of hate speech bans in *The Harm in Hate Speech*. Waldron's argument is clear and compelling, not least because it is rooted in an attractive picture of justice and of what it means to respect human dignity. Moreover, it helpfully integrates the insights which law, political philosophy, and feminist thought have contributed to the issue.

While existing responses to Waldron have tended to resist his account of what makes hate speech harmful,[9] I find his account prima facie persuasive. Instead, then, I take issue with the claim that coercive legal bans constitute the best response to this harm. Waldron's argument relies on an overly optimistic picture of his preferred solution (bans), and an overly bleak and conceptually unsophisticated picture of the main alternative (countering hate speech with more speech). By offering a more refined understanding of counterspeech, and a more nuanced vision of bans, I argue that, very often, critiques of counterspeech either fail, or succeed at the cost of undermining hate speech bans as well. In this light, we should be significantly less disposed to ban hate speech than Waldron is. Note, however, that my

---

p. 1; Caleb Yong, "Does Freedom of Speech Include Hate Speech?" *Res Publica* 17 (2011): 385-403, pp. 394–96; Susan Brison, "The Autonomy Defense of Free Speech," *Ethics* 108, no. 2 (1998): 312-339, pp. 313–15.

[7] See *Chaplinsky v. New Hampshire*, 315 US 568 (1942).

[8] For instance, the UK's Public Order Act of 1986, prohibits insulting or abusive speech when it is *likely to stir up hatred*.

[9] E.g., Robert Simpson, "Dignity, Harm, and Hate Speech," *Law and Philosophy* 32 (2013): 701–28; Jonathan Seglow, "Hate Speech, Dignity, and Self-Respect," *Ethical Theory and Moral Practice* 19 (2016): 1103-1116; Brian Leiter, "Review of *The Harm in Hate Speech*," *Notre Dame Philosophical Reviews*, 2012. Available online at: http://ndpr.nd.edu/news/32077-the-harm-in-hate-speech/ ; Tariq Modood, "Hate Speech: The feelings and beliefs of the hated," *Contemporary Political Theory* 13 (2014): 104-109.

defense of counterspeech remains "pessimistic": it will proceed not so much by suggesting that democratic counterspeech fully overcomes the objections levelled at it, but rather by showing that those objections are at least as problematic for bans.

Prior to developing this argument, notice one limitation. Although I focus primarily on Waldron's persuasive account of what makes hate speech harmful, this is not to deny that there are other ways in which hate speech can harm. Thus, my argument in this article is limited in scope. Nevertheless, my inquiry is not irrelevant to these other harms: as a further line of investigation, we should explore to what extent the strategy I deploy for defending counterspeech can successfully be extended to other speech-based harms. Because many of the other harms associated with hate speech involve claims about the *causal* effects of hate speech, I expect that this further line of investigation will be in significant part empirical.

The rest of this article is organized as follows. I begin by outlining Waldron's framework for understanding the harm in hate speech, and why he takes it to support bans (Section 2). In turn, I critique this defense of bans by articulating a more sophisticated understanding of counterspeech which, following Corey Brettschneider, emphasizes the state's role in responding to hate speech (Section 3). Finally, although using state-based speech to counter hate speech encounters serious objections, I devote the majority of this paper to demonstrating that the two weightiest objections to such counterspeech can be turned, with greater force, against bans. Not only are counterspeech and bans "companions in guilt" with respect to these objections, but bans are the guiltier party (Sections 4-5).

2. Waldron's Dignitarian Case for Bans

For some, the harm involved in hate speech warrants coercively excluding or banning it from democratic public discourse.[10] According to their opponents, countervailing considerations, such as freedom, democracy or truth, demand that we tolerate it.[11] Waldron articulates a framework for understanding the harm produced by hate speech which, he claims, should incline us to the former option.

What is the harm in hate speech? Waldron starts from an appealing picture of justice, whereby justice requires treating human beings with dignity. In the first place, this means that citizens' status as free and equal members of society in good standing—their dignity—should be recognized and upheld by other citizens. However, drawing on Rawls's idea of a well-ordered society, Waldron argues that justice requires something more. Citizens must also *know* that their peers uphold their good standing. Absent the assurance that their dignity is safe, citizens cannot fully enjoy their good standing. Indeed, this assurance is necessary for citizens to pursue their aims and participate in civil and political life without fear or shame.[12]

Because the public assurance of dignity is an essential component of justice, Waldron continues, we should give greater emphasis to "political aesthetics". We should, in other

[10] E.g., Waldron, *The Harm in Hate Speech*; Richard Delgado, "Words that Wound," in Mari Matsuda (ed.), *Words that Wound* (Boulder, CO: Westview Press, 1993), pp. 89–110; Brison, "The Autonomy Defense of Free Speech,"; Michael Blake, "Liberal Foundationalism and Agonistic Democracy," *Nomos* 46 (2005): 230–43.

[11] On freedom or autonomy, see, e.g., Edwin Baker, "Harm, Liberty, and Free Speech," *Southern California Law Review* 70 (1996): 979–1020; Brettschneider, *When the State Speaks, What Should It Say?*; Ronald Dworkin, "Is There a Right to Pornography?" *Oxford Journal of Legal Studies* 1 (1981): 177-212. On democracy, see, e.g., Heinze, *Hate Speech and Democratic Citizenship*; Ronald Dworkin, "Foreword," in Ivan Hare and James Weinstein (eds.), *Extreme Speech and Democracy* (Oxford: Oxford University Press, 2009), pp. v-x; Robert Post, "Racist Speech, Democracy, and the First Amendment," *William and Mary Law Review* 32 (1990): 267-328. On truth, see, e.g., John Stuart Mill, *On Liberty* (London: Penguin, 2006 [1859]).

[12] Waldron, *The Harm in Hate Speech*, pp. 65–69, 81–89.

words, care about what our society literally looks and sounds like.[13] Imagine what a hate speech-allowing society would look like to targets of hate speech. Think of the "marching feet and chants of neo-Nazis in Skokie".[14] Think of a Muslim citizen confronted with newspapers depicting Muslims as sexually-depraved animals. By our strongest intuitions, this is not what a well-ordered and just society looks and sounds like. Here, it is not public knowledge that all citizens uphold the dignity of others.

We can appreciate, then, a significant harm generated by public hate speech. In publicly denying the status of its targets as social equals, hate speech affects the way the society looks in such a way as to undermine the public assurance that their status is secure. Moreover, public hate speech reaches out to other hateful persons, to assure them that they are not alone.[15] In doing so, it replaces the assurance of dignity with the assurance of hatred.

How exactly does hate speech relate to this harm? Dignitarian accounts of the harm in hate speech commonly argue that hate speech *causes* dignitary harms. On this understanding, hate speech causally influences how other members of society view and act towards the target group, so that they stop upholding their status as equals.[16] While Waldron's account is consistent with the causal claim, he largely focuses on how hate speech *constitutes* an assault on dignity. In publicly saying 'Listen closely: you are not one of us, and we want you out', hate speech is undermining the public assurance that the target group are respected by their peers. At the same time, it is assuring other citizens with hateful views that they are not alone. On Waldron's view, then, public hate speech does not merely cause an assault on its targets' dignity: it *is* an assault on their dignity.[17]

For Waldron, the fact that hate speech constitutes an assault on its targets' effective enjoyment of dignity reveals how fragile this good is. "The flare-up of a few particular incidents can have a disproportionate effect" on the public assurance of dignity.[18] Hearing racist vituperations of far-right sympathizers in a public square or on the radio may well produce fear and a sense of alienation in its targets, even if the views uttered are marginal. This, Waldron concludes, is why coercive hate speech bans are required. By forcefully excluding hate speech, bans prevent such speech from creating an environment of fear and alienation.

Notice three advantages of this account. First, it rests on intuitively appealing moral foundations, according to which justice requires not just respecting the dignity of human beings by upholding their status as equal citizens, but also publicly assuring them that this is so. Second, the framework is methodologically appealing. Waldron embraces non-ideal theory by demanding that we imagine what hate speech is like in the "real world", particularly to those on the receiving end. This commitment to taking the experiences of targets seriously is a welcome corrective to scholarship which, he suggests, frequently adopts an abstract and correspondingly idealized conception of hate speech.[19] Finally, because it focuses on the harms constituted (as opposed to caused) by hate speech, Waldron's framework is dialectically powerful. If Waldron were simply claiming that hate speech

[13] Ibid, pp. 71-77.

[14] Ibid, p. 71.

[15] Ibid, p. 94.

[16] For causal accounts, see, e.g., Steven Heyman, "Hate Speech, Public Discourse, and the First Amendment," in Ivan Hare and James Weinstein (eds.), *Extreme Speech and Democracy* (Oxford: Oxford University Press, 2009), pp. 158–81; and Angelo Corlett and Robert Francescotti, "Foundations of a Theory of Hate Speech Symposium," *Wayne Law Review* 48 (2002): 1071–1100.

[17] Waldron, *The Harm in Hate Speech*, pp. 166–67. For other constitutive accounts, see, e.g., Langton, "Beyond Belief"; Ishani Maitra, "Subordinating Speech," in Ishani Maitra and Mary Kate McGowan (eds.), *Speech and Harm* (Oxford: Oxford University Press, 2012), pp. 94–120.

[18] Waldron, *The Harm in Hate Speech*, p. 94.

[19] Ibid, p. 33.

causes a certain kind of harmful effect, he would need to adduce considerable empirical evidence that hate speech is indeed reliably correlated with such effects. By contrast, if public hate speech *constitutes* a harm—that is, if it is part and parcel of that harm—then the link between hate speech and the harm in question is more immediate, and such empirical evidence is not needed. Thus, Waldron's framework sidesteps the fiercely contested empirical question of what the causal effects of hate speech really are.[20]

Waldron's argument constitutes a promising starting point for thinking about what we should do about public hate speech. But it remains problematic. Although his account of hate speech's harm and of its relation to that harm are convincing, the inference to the endorsement of hate speech bans is too direct, and ultimately unpersuasive.

3. A Fairer Comparison

A key problem with Waldron's argument is that the comparison he presents between the society that bans hate speech occurring in public discourse and the society that allows such speech is overly idealistic about the former and overly uncharitable about the latter. The present section aims to rectify this unfairness. Take the hate-banning society first:

> *The Hate-Banning Society*: The society in which bans on hate speech suppress public occurrences of hate speech by threatening to impose coercive sanctions on such speech.

One of the most commonly invoked worries with this society is that having one's speech suppressed significantly undercuts one's freedom or autonomy. Although this point can be made with different conceptions of autonomy or freedom, consider how it applies with freedom understood as non-domination. Domination, according to Pettit's influential account, is subjection to alien or unchecked power. Accordingly, non-domination is secured insofar as individuals are capable of contesting the power to which they are subjected, so that it is forced to track their avowed interests.[21] Now, the ability of participants to contest and thereby to check political power is enhanced when they can publicly appeal to the considerations they deem most significant. This remains true, the present worry continues, even when those considerations are hateful ones. Hence, banning public hateful considerations diminishes the freedom as non-domination of citizens with hateful political views.[22]

Importantly, Waldron himself recognizes these freedom- or autonomy-based costs. Bans, he claims,

---

[20] For a more detailed discussion of the difference between speech causing and constituting harm, see Jennifer Hornsby and Rae Langton's account of perlocutionary and illocutionary speech acts ("Free Speech and Illocution," *Legal Theory* 4 (1998): 21-37).

[21] Frank Lovett and Phillip Pettit, "Neorepublicanism: A Normative and Institutional Research Program," *Annual Review of Political Science* 12, no. 1 (2009): 11-29, p. 14.

[22] For similar objections, see, e.g., Brettschneider, *When the State Speaks, What Should It Say?*, pp. 76–80; Yong, "Does Freedom of Speech Include Hate Speech?," pp. 390–93; Ronald Dworkin, "Is There a Right to Pornography?" *Oxford Journal of Legal Studies* 1 (1981): 177-212. Baker ("Harm, Liberty, and Free Speech," p. 992) raises a related objection, but which appeals to a substantially different conception of autonomy, the capacity for disclosing oneself to others. Note that this objection cannot defend the protection of all kinds of hate speech. Shouting a slur at someone in the street, for instance, is not an attempt at contesting political power. Nor does it seem a necessary means of self-disclosure. As discussed in the introduction, however, I am bracketing such assaultive, face-to-face forms of hate speech. By contrast, the examples of public hate speech mentioned in the introduction (e.g., 'Muslims are terrorists') sometimes *are* raised to support political demands.

> have a direct bearing on freedom to publish, sometimes on freedom of the press, very likely on freedom of the Internet […] And this matters to individuals. Often the messages that racists or Islamophobes are stopped from expressing are the very messages […] that matter most to them. For them, other aspects of political expression pale into insignificance [by comparison][23]

But he responds that these considerations are outweighed by considerations grounded in the dignity of targets of hate speech. All things considered, then, bans seem morally justified. We could go even further on Waldron's behalf. Lacking assurances that one is a citizen in good standing risks depriving one of the psychological conditions needed to participate politically without fear or shame. By undermining the public assurance of dignity, hate speech may therefore undercut its targets' ability to contest exercises of political power. As such, even *within* the value of freedom (understood as non-domination), one might think that the balance of reasons favors bans. I want to grant these responses to the freedom-based argument, not just for the sake of argument, but also because—unlike some critics[24]—I find such claims about the harmfulness of hate speech persuasive.

Even so, we should still be worried about bans. Although the freedom-based reasons against bans may be outweighed, they remind us that the dignity and freedom of targets of hate speech comes at a significant price. Part of what is non-ideal in actual circumstances is that very many citizens uphold hateful viewpoints. Witness the rise of European far-right parties like the French *Front National*, whose rhetoric is often saturated with xenophobia. In some cases, moreover, supporters of such hateful political ideologies are economically disaffected citizens who feel, sometimes justifiably, that their voices have been ignored by mainstream parties.[25] In such non-ideal conditions, coercively suppressing hate speech seems to come at a substantial moral cost: it further diminishes the freedom of very many of the worst off citizens, and thereby fuels the exclusionary processes that, often, contributed to producing their hateful disaffection.[26]

What this means is that even if Waldron is right that the claims of targets of hate speech outweigh those of citizens who possess and wish to express abhorrent views ('hateful citizens', for short), the trade-off between the two is significantly worse than he acknowledges. This realization should strongly motivate us to seek a solution that weakens the trade-off. In other words, we should seek a solution that blocks harmful features of hate speech while limiting the costs to hateful citizens. Simply suppressing hate speech should be a last resort. Here, we should be reminded of Sacco's compelling reaction, which I introduced at the outset. Given the moral cost of forcibly excluding extreme viewpoints, Sacco suggests that we should instead try, as far as possible, to engage with and understand possessors of such viewpoints, why they feel disaffected, in a way that does not reproduce their sense of being second-class citizens. And doing so may require allowing hateful citizens to publicly disclose their objectionable worldviews. This intuition, I suggest, is what the normative ideal of the hate-allowing society, at least in its more sophisticated variant, aims to capture.

---

[23] Waldron, *The Harm in Hate Speech*, pp. 148–49.

[24] E.g., Dworkin, "Foreword," p. vi.

[25] E.g., Kim Willsher, "Abandoned French Working Class Ready to Punish Left's Neglect by Voting for Far Right," *The Guardian*, 2015, http://www.theguardian.com/world/2015/mar/22/alienated-french-working-class-vote-far-right-claims-analyst.

[26] Why does it matter that those affected by bans are typically relatively badly off? One rationale is consequentialist—on this view, ignoring those who have previously been ignored is problematic because it risks alienating them further, and breeding more hatred. But the rationale needn't be consequentialist. For instance, one might think this matters because one shares the basic prioritarian intuition that, ceteris paribus, losses of welfare or freedom are more morally problematic when they afflict those whose overall levels of well-being are lower.

Before turning to examine the hate-allowing society, a further comment is in order. In observing that those whose freedom would be restricted by bans are often relatively disadvantaged and numerous, my intention is to bring out how crucial it is to find an alternative response to hate speech. But it is important to note that my more fundamental point does not hinge on this sociological observation. Even if hateful citizens tended to be well off (as they sometimes are), there would still be *a* freedom-based cost to restricting their political speech. Accordingly, it would remain true that *if* we can find an alternative solution that blocks the harm in hate speech without imposing this cost on hateful citizens, we should prefer that solution to legal bans.[27]

To this end, let us now consider the hate-allowing society. Waldron's most sustained depiction of this society is unflattering. Opponents of hate speech bans "love the richness and untidiness of the marketplace of ideas: let a thousand flowers bloom, they say, even the poisonous ones".[28] On this crude understanding, the hate-allowing society responds to the interests of hateful citizens by allowing them to publicly disclose their abhorrent worldviews. As for the interests of targets of hate speech, the idea is that since *all* citizens can voice their opinions, hateful views can be countered by more public speech, which challenges hateful utterances.

This 'counterspeech' solution is inadequate. First, it might seem unfair, insofar as it typically places the burden of challenging hate speech on its targets. Second, it seems ineffective at protecting targets of degrading speech. To begin, empirical evidence indicates that targets of public hate speech seldom respond directly to hate speakers. This suggests that the very harmfulness of hate speech makes speaking back on the spot very difficult for its targets.[29]

Now, one might think that this evidence is not exceedingly problematic: perhaps what is needed is not so much that targets of hate speech respond then and there, but rather that over the long run, they vocally challenge these abhorrent views, for instance through mass media campaigns. But the more fundamental problem is that even if targets of hate speech did speak back, it is unclear how effective this would be at preserving the public assurance of dignity. The difficulty, as philosophers of language often put it, is that targets of hate speech frequently lack the authority that is needed to block the harm in hate speech.[30] If what is harmful in public hate speech is that it asserts that a given group are unwanted by the society at large, then being able to refute this assertion requires being able to speak on behalf of the society at large. Since targets of hate speech are typically members of minority groups, they are often not in a position to do this. To put this point more vividly, hate speech says: 'You and your kind are unwelcome here. We want you out.' Intuitively, its targets often cannot refute this by simply saying: 'No, we are not unwanted!'

---

[27] Thanks to a reviewer for raising the issue of whether my defence depends on the sociological claim. As this reviewer also suggested, a deontological argument, which need not involve balancing the interests of hate speakers against those of their targets, might circumvent this issue altogether. See, e.g., Heinze, *Hate Speech and Democratic Citizenship*, chs. 3-4. Here, however, I will continue with the balancing approach because I am trying to show that even on Waldron's own terms, his argument for bans (which involves a balancing of interests) does not succeed. For Waldron's interest-balancing, see *The Harm in Hate Speech*, pp. 147-150.

[28] Waldron, *The Harm in Hate Speech*, p. 67. An unrefined understanding of counterspeech is also presupposed by: Delgado, "Words That Wound" pp. 95, 108; Catharine MacKinnon, *Only Words* (Cambridge, MA: Harvard University Press, 1996), pp. 72-73.

[29] Laura Beth Nielsen, "Power in Public: Reactions, Responses and Resistance to Offensive Public Speech," in Ishani Maitra and Mary Kate McGowan (eds.) *Speech and Harm* (Oxford: Oxford University Press, 2012), pp. 148–73.

[30] Ishani Maitra and Mary Kate McGowan, "Introduction and Overview," in their *Speech and Harm* (Oxford: Oxford University Press, 2012), pp.1-22, at pp.9-10.

However, this picture of the hate-allowing society and of counterspeech is uncharitable. As Corey Brettschneider has shown, setting up a choice between either having the state ban hate speech or having the state step back and let private citizens respond is misleading. Instead, Brettschneider suggests, the state can be active as a speaker endowed with expressive powers. So a more refined alternative to bans is the following:

> *The Hate-Allowing Society*: The society where hate speech is not coercively suppressed, but where, in response to hate speech, the state actively speaks out against occurrences of hate speech.[31]

Consider the wide array of tools states possess for speaking back. The state should speak back by having public officials affirm the ideals of human dignity, not just through public pronouncements in formal deliberative forums, but also by naming public spaces, dedicating monuments, and enacting public holidays. Moreover, state officials should speak back by denouncing particular instances of public hate speech.[32]

This solution is promising, Brettschneider contends, because it seems capable of avoiding two undesirable outcomes. On the one hand, it avoids the freedom-based costs associated with coercively preventing citizens from voicing hateful viewpoints. Thus, it steers clear of what Brettschneider calls the "Invasive State", which monitors and controls its citizens' speech in a way that substantially threatens their freedom. At the same time, because the state speaks out against hate speech, this solution averts the opposite extreme: the "Hateful Society", where hate speech flourishes and succeeds in harming its targets.[33]

To substantiate this promise, let us examine more closely how state-based counterspeech handles the specific dignitarian harm we have been concerned with. Before doing so, however, note a terminological departure from Brettschneider. Whereas Brettschneider most often refers to the policy under examination as "democratic persuasion", I will continue using the broader term "state-based counterspeech". This is to emphasize that, here, I am concerned less with state-based counterspeech's ability to *persuade* hate speakers and others, and more with its ability to block hate speech's attempted assault on the dignity of its targets.[34]

With these clarifications in mind, how does state-based counterspeech avoid the objections raised against more ordinary forms of counterspeech? First, while this solution is consistent with the victims of hate speech also speaking back, it is not primarily victims who are required to respond. Rather, this task primarily devolves to public officials or, as I explain below, third parties. Thus, this solution seems less unfair. Second, state officials, unlike many private citizens, are often capable of authoritatively challenging hate speech's attempted assault on the public assurance of dignity. In decrying detestable speech, the state is *reassuring* its targets. It is saying: 'We are sorry that you had to hear that. It's true, some individuals unfortunately still believe this. But *on behalf of the people generally*, rest assured that we stand by you, that they, not you, are the outliers, and that we *will* ensure that what they say remains an empty threat.'

This might seem overly abstract. To make matters more concrete, consider how Christiane Taubira (then the French Minister for Justice) intervened in response to David

[31] Brettschneider, *When the State Speaks, What Should It Say?*

[32] Ibid, pp. 95-96. See also Katharine Gelber, "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* (Cambridge: Cambridge University Press, 2012), pp. 198–216.

[33] Ibid, pp. 10-18, 77-81.

[34] Note that although he often focuses on how state speech can persuade listeners (e.g., pp. 6, 86), Brettschneider also acknowledges other uses, such as signalling that the state is not complicit with hateful citizens' vilifying views (e.g., p. 17). This latter function is closer to the one I emphasize.

Rachline (a *Front National* Senator) in the French Senate in 2015. After Rachline criticized Taubira's advocacy of humane treatment for criminals, whom he referred to as "thugs" and "executioners", Taubira replied to his various claims, and concluded as follows: "You belong to a political family […] that systematically seeks scapegoats […] and that looks for invaders everywhere […] [O]ne shudders at the thought of what would happen if it took power—but we will fight to make that absolutely impossible".[35] The overwhelming majority of the Senate rose to their feet to applaud this response. In her response, Taubira does several things: she takes the time to answer; she denounces far-right ideologies and their scapegoating practices; she promises to keep those exclusionary ideas away from power; and she does this in broad daylight, to wide acclaim. If we want hate speakers isolated, so that they cannot undermine the assurance of dignity, public counterspeech like Taubira's seems to be an exemplary realization of those aims. Her response not only contributes to making Rachline's ideology isolated (as a ban might achieve), but also publicly shows it to be isolated.

In response to this example, one might point out that France does have hate speech bans, and that Taubira herself does not repudiate such bans. In fact, in other instances, legal action has been brought against *Front National* officials for levelling racist insults at Taubira.[36] While this is true, it does not undermine the main point of the example: that when abhorrent views *are* aired without being legally punished, as happened in this case, having public officials condemn them seemingly suffices to uphold the public assurance of dignity. State counterspeech thus appears sufficient to block the harm Waldron ascribes to hate speech.

Of course, public officials can only do so much. But Brettschneider contends that the state can also speak back in more indirect ways. First, it can fund private associations, like civil rights groups, whose functions include publicly denouncing occurrences of hate speech. This provides institutional support to targets of hate speech by empowering third parties to block its attempted harms.[37] Second, the state can speak against hateful views through processes of public education, which teach children about historical events like the Holocaust, and expose them to democratic values.[38]

The latter recommendation involves the state coercing individuals as well as speaking to them. Parents are required to send their children to school. And, consequently, they will almost unavoidably have to come into contact with worldviews they disavow. Nevertheless, it does not coercively prevent hateful citizens from publicly expressing their viewpoints. And it is coercively suppressing such contributions which is especially problematic for individual freedom. Recall: doing so is problematic, in part, because it reduces one's ability to publicly contest exercises of political power. But having to send one's children to school, or hear other viewpoints, does not.[39] Accordingly, hardly anyone in this debate denies that the state should exercise coercion in other areas of public life.

Let us take stock. On this refined picture of counterspeech and of the hate-allowing society, the state intervenes to block the dignitarian harms of hate speech. But it does so without coercively suppressing hateful citizens' public utterances. Accordingly, it weakens

[35] Sénat Video, "Questions Au Gouvernement: David Rachline," *Sénat*, 20 October 2015, http://videos.senat.fr/video/videos/2015/ video30342.html.

[36] Le Monde, "Neuf mois de prison fermes pour l'ex-candidate FN ayant comparé Taubira à un singe," *Le Monde*, 2014, http://www.lemonde.fr/societe/article/2014/07/15/taubira-comparee-a-un-singe-9-mois-de-prison-ferme-pour-l-ex-candidate-fn_4457839_3224.html.

[37] Brettschneider, *When the State Speaks, What Should It Say?*, pp. 110–11. However, I wish to remain agnostic regarding Brettschneider's further claim, in chapter 4, that *removing* funds from organizations counts as non-coercive, and should be included under the umbrella of state counterspeech. For criticism of this claim, see Sarah Song "The Liberal Tightrope," *Brooklyn Law Review* 79 (2013-2014): 1047-1057, at pp.1053-1057.

[38] Brettschneider, *When the State Speaks, What Should It Say?*, pp. 96-104.

[39] Ibid.

the aforementioned trade-off between the freedom of hateful citizens and the freedom of their targets. In light of this fairer comparison between the hate-allowing society and the hate-banning society, the former seems more desirable. In what follows, I consider two objections which suggest that my Brettschneider-inspired argument thus far idealizes the notion of state-based counterspeech. In both cases, however, I will show that counterspeech and bans are companions in guilt: these objections can be turned—with greater force—against the hate-banning society.

4. Objection 1. Does Counterspeech Block All Hate Speech?

4.1. The Limited Effectiveness of Counterspeech

One might object that it is unrealistic to think that speaking back can block the dignitarian harm in hate speech. In his very brief consideration of state-based counterspeech, Waldron objects that this solution would provide the wrong kind of public assurance of dignity. In particular, raising "billboards proclaiming the principles of justice" would be "creepy" and "totalitarian".[40] To avoid this, the assurance of dignity should be provided in a way that is more implicit.

This response seems misguided in several respects. To begin, it is unclear in what sense laws banning hate speech constitute implicit assurances. After all, such laws must be publicly known and typically lead to widely publicized lawsuits. But even granting this, we can query the claim that explicit assurances are undesirable. First, Waldron makes a straw man of speaking back. As the Taubira example illustrated, counterspeech often involves public responses to particular instances of hate speech that are not flat-footed in the way Waldron's envisaged billboards are. Moreover, even if state counterspeech does involve explicit pronouncements of values, is that really a bad thing? Does Waldron really think that billboards—like those displayed in football matches—stating 'SAY NO TO RACISM' are "creepy"? And is explicitly erecting monuments honoring the Civil Rights Movement "totalitarian"?

Waldron seems to think such explicit assurances are bad because they are "evidence of a problem".[41] But racism *is* a problem. If Waldron is suggesting that banning hate speech could somehow eliminate racism or make it entirely invisible, this runs counter to his repeated exhortations that we should think of what justice requires in the real world. Moreover, even if banning hate speech could somehow make racism entirely invisible, Section 3 suggested that, given that it is a problem, it seems at least an open question whether it is preferable to recognize the problem and denounce it, or maintain a blissful ignorance of its existence.

A better version of the objection under consideration comes from philosophers of language who argue that degrading speech often takes a form that is very difficult to identify and speak out against. Rae Langton, in particular, has emphasized how degrading representations are often *presupposed* rather than straightforwardly asserted in hate speech.[42] Instead of saying 'Blacks are lazy', someone might say 'Even blacks would do that job', thereby implying that blacks are lazy. Why does this matter? Conversations tend to follow rules of accommodation. Very roughly, this means that the content of presuppositions tends to be added by default to the common background assumptions that the relevant conversation relies on, unless those presuppositions are explicitly challenged. Suppose you say 'I need to walk my dog'. The idea is that it automatically becomes an accepted assumption in our

[40] Waldron, *The Harm in Hate Speech*, p. 87.
[41] Ibid.
[42] Langton, "Beyond Belief," pp. 82–84.

conversation *that you have a dog*, unless I say 'Wait a minute, you don't have dog!'[43] This accommodation of presuppositions means that some contents can be smuggled in more easily to the conversational common ground than if they were straightforwardly asserted.

On its own, this is not immediately problematic for state-based counterspeech. If the problem is that degrading representations in hate speech automatically become common ground unless they are blocked, then it seems this rather *calls* for counterspeech. While one way of stopping the degrading representation of black citizens from becoming common ground is to prohibit utterances of 'Even blacks would do that job', the other is to block it through counterspeech: 'What do you mean, *even blacks*?! We don't condone those views around here!'

It is in light of a further observation, concerning coded speech, that the point about presupposition becomes particularly challenging. Jason Stanley argues that the conventional meanings of words can shift, so that hateful representations become encoded in ordinary speech. For example, repeated associations in the news of ordinary terms, like the association of 'inner cities' with black citizens, eventually makes this term refer to black citizens. Similarly, repeated associations of 'welfare' or 'food stamps' with laziness make 'relying on welfare' or 'relying on food stamps' encode laziness.[44]

When combined with the existence of coded speech, the observation regarding presuppositions is highly problematic. For presuppositions not to be added to the conversational common ground, they must be identified and challenged. But if they are concealed through coded speech, doing so becomes far more difficult. More precisely, when degrading contents are encoded in public speech, the message is typically quite clear to its targets and to other hateful citizens, but less visible to third parties. This, in turn, is particularly worrying for *state-based counterspeech*, since it emphasizes the need for third parties, including public officials, to play a central role in responding to hate speech.

To illustrate, consider Stanley's example of an interview conducted in 2012 with Newt Gingrich, a prominent Republican politician.[45] Gingrich had previously claimed that "black Americans should demand jobs, not food stamps". In response, the interviewer observed that this was offensive to blacks, as Gingrich was clearly invoking negative racial stereotypes—i.e., that blacks are lazy. Gingrich answered "No, I don't see that", and was loudly applauded, while the interviewer was booed.

The example suggests several things. First, it illustrates the use of coded presuppositions expressing degrading contents in public speech. Gingrich's claim presupposes *that black Americans ask for food stamps rather than jobs.* And if 'food stamps' encodes laziness, the presupposition means *that black Americans are lazy*. This is problematic because it makes identifying the degrading stereotype Gingrich expresses more difficult. But the example also reveals a second problem. Even when one has identified the presupposition, speaking back against it risks backfiring. Because the degrading content is encoded, the person who wishes to block it (here, the interviewer) must be the one who first explicitly invokes it. This makes it possible for the person who challenges the hateful content to be blamed for its expression, so that the intervention achieves the opposite of its intended effect. Instead of assuring black citizens of their standing, the would-be challenger is made to seem as though he holds degrading views and expresses them publicly.

What is worse, the above example is hardly the most coded imaginable. Substituting 'inner city' for 'black Americans', Gingrich could have presupposed the same demeaning content by saying 'Inner cities' reliance on food stamps is an issue that needs addressing'.

[43] Ibid., pp. 82–89.

[44] Stanley, *How Propaganda Work* (Princeton, NJ: Princeton University Press, 2015), pp. 123, 138, 156–57.

[45] Ibid, pp. 156-167.

Stanley goes even further, and suggests that the very words 'welfare' or 'food stamps' can sometimes immediately encode *that blacks are lazy*, even in utterances that don't appear to be about black citizens (e.g., 'Food stamps are expensive'). In these cases, what the terms presuppose is their own conventional meaning, which has been altered to encode the degrading stereotype.[46] Hence, the difficulties illustrated in the Gingrich example might easily be amplified. In sum, then, coded speech, which Stanley argues is diffuse, seems to present profound difficulties for state-based counterspeech.

4.2. Two Replies: A Defense and an Attack

A first reply is that we should nonetheless not conclude too much from these difficulties. Stanley at times draws deeply skeptical conclusions regarding counterspeech, including state counterspeech, from the foregoing analysis. At one point, he suggests that the contents encoded in speech are "not negotiable".[47] As a result, the existence of coded presuppositions renders effectively speaking back through public deliberation "impossible".[48] On this pessimistic view, state-based counterspeech is wholly inadequate when it comes to challenging coded vilifying contents.

But this conclusion is too strong. Having authoritative political figures speak back *is* sometimes an effective way of blocking hateful coded presuppositions. Indeed, Stanley's own examples suggest as much. For instance, following US Representative Paul Ryan's comments asserting that "inner cities" lacked a "culture of work", Representative Barbara Lee replied: "Let's be clear, when Mr. Ryan says 'inner city' […] these are simply code words for what he really means: 'black'".[49] Ryan denied having said anything about race, but the response stuck, prompting highly public criticisms in the media.[50]

This last case suggests a more nuanced conclusion. Counterspeech can indeed block the dignitarian harms of coded degrading speech. However, the success of such counterspeech depends on having various skills and abilities, including (amongst other things): the epistemic ability to identify the presuppositions; the clarity to challenge them in a way that makes it more difficult for others to twist one's words; and the authority which enables one to speak publicly and be taken seriously when one battles over words. Indeed, being an elected representative with a record of fighting for civil rights was likely part of what enabled Lee to be more successful than Gingrich's interviewer.[51]

Some might object that these characteristics, which Lee possessed to a significant degree, are rare. Many ordinary citizens might not possess them. However, this objection misses its target. The fact that these conditions are rare is precisely why the form of counterspeech I am defending demands that state officials and their empowered agents speak back. Although this by no means guarantees success, it shows that the solution I am putting forward is sensitive to what is needed for success against coded speech.

Nevertheless, the objection from coded speech remains problematic, and constitutes an apt qualification of Brettschneider's optimism regarding state counterspeech. Quite often, counterspeech may fail to block hate speech's dignitarian harms. Accordingly, my second and more fundamental reply does not deny this worry, but argues that it can be turned, with

[46] Ibid, p. 138.
[47] Ibid, p. 158.
[48] Ibid, p. 161.
[49] Ibid, p. 160.
[50] E.g., Charles Blow, "Paul Ryan, Culture and Poverty," *The New York Times*, 2014, http://www.nytimes.com/2014/03/22/opinion/blow-paul-ryan-culture-and-poverty.html?_r=1.
[51] Cheryl Chumley, "Rep. Barbara Lee Slams Bill O'Reilley for 'Code Words' That Hide Racism," *Washington Times*, 2014, http://www.washingtontimes.com/news/2014/mar/27/rep-barbara-lee-slams-bill-oreilly-code-words-hide/.

greater force, against the hate-banning society. Indeed, Waldron's picture of the hate-banning society is overly idealized if, as we have been assuming, it is taken to infallibly block the dignitarian harms produced by public hate speech.

The idealization at hand is partly empirical. As was noted in Section 2, Waldron sidesteps debates about the causal effects of hate speech. However, he does make empirical assumptions concerning the causal effects of bans—namely, that they are effective at eliminating hate speech, or at least at removing it from public arenas where it can undermine the public assurance of dignity.[52] A number of commentators have argued that these claims are at best contentious, and at worst unfounded.[53]

Even so, because the empirical evidence regarding the overall merits of bans is mixed,[54] we should try to avoid relying on it. For our purposes, then, the main problem is the following. Advocates of bans must show that coercive legal bans are better at handling the cases that proved problematic for state counterspeech. But even in theory, bans seem singularly unpromising with respect to coded hate speech. Recall the coded sentence discussed above: 'Inner cities' reliance on food stamps is an issue that needs addressing'. In context, this presupposed *that blacks are lazy*. However, passing a ban on such speech seems extremely difficult. To begin, and most obviously, the epistemic problems which afflicted counterspeech apply here too: simply identifying instances of coded speech for the purposes of legislating or implementing legal bans can be difficult.

Moreover, two further problems arise specifically for bans. First, legal bans are blunt tools, whereas the very existence of coded hate speech shows how hate speech can easily shift in form. Even if we identified and banned coded hate speech, hate speakers might easily adopt new codes to avoid the ban.[55] The problem is not so acute with counterspeech. It is easier to speak back effectively against a new instance of coded speech than it is to pass yet another law or even to implement a generally-worded law in a significantly new way. At bottom, this problem is moral, not empirical: what makes bans relatively inflexible is that, because they employ coercive force (whereas counterspeech does not), implementing them requires meeting a higher burden of moral justification. This (as I discuss in Section 5) is because coercive measures involve a pro tanto restriction of the coercee's freedom. The point, then, is not just that the institutional processes for implementing bans are in fact more inflexible than counterspeech—instead, it is that they are morally required to be. Even if bans could be just as flexible as counterspeech, they *morally should* be less flexible.

---

[52] If Waldron denied caring about the causal effects of bans on hate speech and asserted that he simply cared about their expressive power, this would generate two problems. First, it would contradict his earlier claim that he wants the public assurance of dignity to be implicit. Second, it would no longer be clear what legal bans offer which counterspeech doesn't, since counterspeech is centrally concerned with expressive powers. In fact, even if one conceded that criminal law has distinctive expressive powers, as Leslie Green ("Should Law Improve Morality?," *Criminal Law and Philosophy* 7 (2013): 473–94) suggests, it would not immediately follow that bans are necessary to play this expressive role. Indeed, as many legal theorists observe, the expressive aim of denouncing the content of hate speech can be performed by many laws besides bans, such as anti-discrimination laws and civil rights laws. See, e.g., Robert Post, *Constitutional Domains: Democracy, Community, Management*, Harvard University Press (Cambridge, MA, 1995), p. 26; Nadine Strossen, "Interview with Nadine Strossen," in Michael Herz and Peter Molnar (eds). *The Content and Context of Hate Speech* (Cambridge: Cambridge University Press, 2012), pp.378-398, at p. 391; Heinze, "Hate Speech and the Normative Foundations of Regulation," pp. 595–96.

[53] For recent negative evidence regarding the effectiveness of bans, see, e.g., Heinze, *Hate Speech and Democratic Citizenship*, pp. 145-153; Jacob Mchangama, "The Harm in Hate Speech Laws," *Policy Review* (2012/2013): 95-102.

[54] For mixed evidence regarding bans, see, e.g., Katharine Gelber and Luke McNamara, "The Effects of Civil Hate Speech Laws: Lessons from Australia," *Law & Society Review* 49, no. 3 (2015): 631–64.

[55] E.g., Heinze, *Hate Speech and Democratic Citizenship*, pp. 145-148.

Second, trying to ban coded speech exacerbates a common charge against hate speech bans, namely that they risk obstructing or 'chilling' democratic public discourse. Coded speech is widespread, and usually involves otherwise innocent and useful terms, like 'welfare, 'food stamps', 'inner city'. As a result, even if we could identify and ban all instances of coded hate speech, including the new ones, this seems undesirable given how much banning this would involve, and how ordinary the banned words would be. This point is especially strong in the more extreme cases of coded hate speech Stanley discussed, where, for instance, a single word like 'welfare' can encode degrading views about black citizens. The counterintuitive implication here would be that 'welfare' should be banned even when used in sentences that otherwise had no bearing on black citizens.

Coded hate speech, then, also poses a problem, and arguably a greater one, for bans. Proponents of bans might perhaps respond that coded speech is less likely to proliferate in the hate-banning society. But this is implausible. First, it is empirically dubious. Notice, in particular, that in many European states that do have bans, extreme-right parties, like the French *Front National* or the German *NPD* have gained public prominence. And they have done so in part by smuggling hateful propositions into public discourse under coded guises. In France, one of Marine LePen's central strategies for mainstreaming the *Front National* has been to co-opt the language of republican values. It has been argued, for instance, that the French far-right has increasingly used words like '*laïcité*' (very roughly, a French concept of secularism) or 'republicanism' to refer specifically to opposition to Islam.[56] Equally, the *NPD* has avoided bans by substituting 'Glory and Honor' for the banned 'Blood and Honor', replacing the swastika with a Nazi-era triangular figure, and using a salute with three-outstretched fingers rather than an outstretched hand.[57] (Notice how the latter example illustrates just how easily citizens evade bans by adjusting their communications). Thus, the advocate of bans' attempted response seems empirically misguided: at first blush, coded hateful speech seems at least as widespread in hate-banning societies.

In fact, this should hardly be surprising, as there are theoretical reasons for expecting coded speech to be *more* prevalent in societies that ban hate speech than in those that do not. First, the aim of avoiding legal penalties without giving up on undermining the public assurance of dignity incentivizes the development of coded hate speech. This is not to deny that some such incentives, like social opprobrium, exist in hate-allowing societies. But, in virtue of operating through the use of coercive sanctions, legal bans produce *further* incentives. Because they do so, Heinze suggests, "bans […] are teaching [hate groups] how to mock and scorn vulnerable targets from within the bounds of the law."[58]

Additionally, bans' principle of operation not only incentivizes, but also facilitates the development of coded hate speech. Robert Talisse hints at this point:

> Where politics is organized around a general principle of non-entanglement designed to avoid confrontations between competing deep commitments, the very content of the concepts that shape and direct our political institutions is left underdetermined; they are up for grabs […] [and] available for easy co-optation[59]

[56] For a sustained analysis, see Aurélien Mondon, "The Front National in the Twentieth-Century: From Pariah to Republican Democratic Contender," *Modern & Contemporary France* 22, no. 3 (2014): 301–20; Hugh McDonnell, "How the National Front Changed France," *Jacobin*, 22 April 2016, https://www.jacobinmag.com/2015/11/marine-jean-marie-le-pen-national-front-immigration-elections/.

[57] Heinze, *Hate Speech and Democratic Citizenship*, pp. 145-148.

[58] Heinze, "Hate Speech and the Normative Foundations of Regulation," p. 600.

[59] Robert B. Talisse, *Democracy and Moral Conflict* (Cambridge University Press, 2009), p. 64.

Put differently, underlying the view that we should ban hateful discourse is the intuitively compelling belief that some propositions are not up for debate.[60] On this view, we need not discuss why, say, 'Immigrants are vermin!' is false—i.e., that all human beings have equal dignity, that this dignity mandates treating them with respect, and so on. We simply ban it. And if some cannot understand why, so much the worse for them. Ironically, however, the upshot is that the relevant values ('dignity', 'respect', 'equality', etc.) become up for grabs. Ceteris paribus, it is easier for hate groups to present their views as egalitarian or concerned with dignity in contexts where no concrete articulation or reiteration of these values has recently been offered. For example, it might be easier for a policy of racial segregation to be publicly supported as egalitarian in circumstances where, because we prefer bans to counterspeech, there has been little public discussion of what equality truly means.

In contrast, democratic counterspeech actively employs powerful tools for authoritatively endowing democratic values with concrete content. It offers arguments and narratives to show precisely what is meant by democratic ideals. Think, for instance, of Martin Luther King Jr.'s "I Have a Dream" oration, where he offers an impassioned vision of a society of equals, which explicitly aims to recapture "the true meaning of equality". Such a narrative contributes to authoritatively clarifying the meaning of equality in a way that makes co-opting this ideal more difficult.

Given these deficiencies with respect to coded speech, advocates of bans might be tempted to offer a final response: that this section's principal line of argument attacks a straw man, as they are in fact not interested in banning coded hate speech.[61] This response is inadequate for two reasons. Firstly, given the logic of Waldron's argument, it is unclear that this response is genuinely available to him. After all, the rationale for banning public hate speech is that it undermines the public assurance of dignity and assures other hateful citizens that they are not alone. But this rationale does not seem inherently sensitive to whether speech takes a coded form or not. Indeed, as the *NPD*'s adjusted Nazi salutes illustrate, coded hate speech can be such that, while it is encoded enough to escape existing regulations, it clearly expresses and diffuses a dignity-denying message. Therefore, this response seems ad hoc or morally arbitrary. Admittedly, Waldron might concede that this arbitrariness is regrettable, but insist that it is inevitable: as a matter of practical feasibility, the law cannot ban all instances of coded speech, and must draw morally arbitrary distinctions between what it does and does not ban.

Nevertheless, there is a second and more fundamental problem with the response under consideration: it is unsatisfactory given the dialectical situation. The discussion of coded speech was initially raised on behalf of critics of state counterspeech to suggest that such counterspeech is less effective than bans. My argument has been that coded speech is at least as problematic for bans as it is for counterspeech. In this context, then, advocates of bans cannot respond that they are not interested in regulating coded speech or are only interested in banning a small portion of such speech without also defusing the original objection to counterspeech.

Let us conclude. I began by considering the objection that counterspeech is often ineffective at blocking hate speech's dignitarian harm, particularly when hate speech takes a coded form. While this objection is forceful, it does not establish that state counterspeech is entirely ineffective in these contexts. Moreover, and more importantly, it can be turned against bans. Not only do bans seem even more ineffective in response to coded speech, but it

[60] Waldron, *The Harm in Hate Speech*, pp. 189–92.

[61] Waldron (ibid, pp. 189-192) might appear to say this, when suggesting that perhaps only intensely degrading speech should be banned. In fact, however, the issue of whether hate speech is coded or not cross-cuts the issue of how degrading its contents are. As illustrated by the *NPD*'s adjusted Nazi symbols, and Holocaust denial, the most abhorrent contents can be put into coded form.

is inherent to their way of functioning that they encourage and facilitate the development of such speech.[62]

5. Objection 2: Does Counterspeech Block Only Hate Speech?

5.1. The Risk of Misdirected Counterspeech

According to the previous objection, having the state speak back does not block *enough* instances of hate speech. By contrast, the present objection claims that state speech typically challenges *too many* instances of speech. That is, it undermines contributions to democratic public discourse that are not hateful (for short, 'non-hateful speech'), and that are even sometimes justice-promoting.

Robin West directs a variant of this objection at Brettschneider, by raising the possibility of the "Hypocritical State". So far, we have been assuming that the state will use its powers of counterspeech against hateful speech. But part of what is non-ideal in the real world is that states and influential political figures sometimes use their expressive powers wrongly. In other words, they misdirect their speech. One explanation for why they might do so, West observes, is that that states are sometimes hypocritical. As a result, they appeal to values like freedom, equality, and justice to advance ends that are contrary to those values, or to denounce public speech that genuinely aims to advance those values.[63]

A few examples might lend support to West's worry that the state may misdirect its expressive powers. Historical cases are common. For example, in *Plessy v Ferguson*, the US Supreme Court publicly invoked the ideal of equality when upholding deeply unjust and inegalitarian practices of racial segregation.[64] Stanley argues that similar cases of misdirected state speech remain widespread in contemporary politics.[65] In the aftermath of the 9/11 attacks, for instance, US Congress passed the Patriot Act, which conferred substantial powers on law enforcement and intelligence authorities. Part of then-President George W. Bush's public justification for this anti-terrorism act, famously, was that "freedom and democracy are under attack".[66] But the act, it has widely been argued, legitimated significant infringements on civil liberties.[67] Several of the most controversial features were therefore revised years later.[68] Here, then, state speech employed the rhetoric of freedom and democracy to authorize a law which arguably threatened those very values. A final example relates to Black Lives Matter. In 2016, the social movement campaigned regularly against structural racism directed at blacks, and notably protested the violent treatment of blacks by law enforcement officers. In response, several prominent Republican politicians, including

[62] This last clause helps pre-empt the response that we can simply *combine* bans and counterspeech. Since bans encourage and facilitate coded hate speech, they impair the effectiveness of speaking back. So it is far from clear that adding counterspeech to bans will necessarily be the best option.

[63] Robin West, "Liberty, Equality, and State Responsibilities," *Brooklyn Law Review* 79, no. 3 (2014): 1031-1045, p. 1039.

[64] *Plessy v. Ferguson* 163 US 537 (1896).

[65] Stanley, *How Propaganda Works,* chs. 2-4.

[66] George W. Bush, "AFTER THE ATTACKS," *New York Times*, 2001, http://www.nytimes.com/2001/09/13/us/after-the-attacks-bush-s-remarks-to-cabinet-and-advisers.html. Thanks to John Filling for suggesting this example.

[67] ACLU, "Myths and Realities About the Patriot Act," *ACLU* , https://www.aclu.org/other/myths-and-realities-about-patriot-act?redirect=myths-and-realities-about-patriot-act; Jim Sensenbrenner, "This Abuse of the Patriot Act Must End," *The Guardian,* 2013, https://www.theguardian.com/commentisfree/2013/jun/09/abuse-patriot-act-must-end

[68] Erin Kelly, "Senate Approves USA Freedom Act," *USA Today*, 2015, http://www.usatoday.com/story/news/politics/2015/06/02/patriot-act-usa-freedom-act-senate-vote/28345747/

Rudy Giuliani, Senator Rand Paul, and Donald Trump, condemned the slogan "Black Lives Matter" as racist on the fallacious grounds that it implied that non-black lives did not matter.[69] Once more, West might want to say, we arguably have a case of misdirected speech: authoritative public figures used the language of equality to denounce a public slogan that was neither hateful nor inegalitarian, and that instead promoted racial equality.[70]

As these cases suggest, the problem is that recommending a policy of state counterspeech in non-ideal conditions risks undermining speech that is not hateful, and that may be justice-promoting. In such conditions, state counterspeech may end up being seriously misdirected. Again, however, I will argue that this objection to state counterspeech can be turned against the policy of banning hate speech. If, as West claims, the state is badly motivated, or hypocritical, then it might misdirect bans as well. It might, say, appeal to the idea of human dignity to ban speech that is not genuinely hateful and that instead aims to advance dignity.

There is *some* empirical evidence of bans being misused. In weakly democratic societies outside the West, Heinze notes, bans are sometimes manipulated "to punish [minorities] for speaking out against the […] dominant ethnic majorities".[71] And although it seems much rarer for bans to be used to punish minorities in stable democracies (which constitute my principal focus), Mchangama does supply evidence that even in countries such as Germany, the Netherlands and France, bans are occasionally misused. For instance, he cites a case, which is not the only one of its type, where a French mayor was fined for advocating a boycott of Israel.[72] Although such a political view is clearly controversial, defenders of bans, including Waldron, are characteristically *not* asking that merely controversial political views be banned.

That being said, I am not committed to the claim that bans are in fact often manipulated or misdirected in stable democracies. Rather, my main claim is a conditional one: *if* there is reason to think that the state's expressive powers will be misused or misdirected, *then*, prima facie, there is reason to think that the state's coercive powers too will be misdirected. After all, if the state is hypocritical, as West suggests, why think that that hypocrisy will not infect its usage of bans as well as its counterspeech? Thus, *insofar* as West's "Hypocritical State" objection is effective against the policy of state counterspeech, it should be effective against the policy of banning hate speech.

Once again, the hate-banning and hate-allowing societies are companions in guilt. Further, the hate-banning society is once more the guiltier party. Indeed, misdirected banning—misusing bans to penalize non-hateful speech—is more morally problematic than misdirected counterspeech—speaking back against non-hateful speech. First, as Brettschneider emphasizes, banning involves coercive power, rather than merely persuasive or expressive power.[73] Coercion operates, roughly, by employing or threatening to employ physical force to make someone do (or not do) something. For this reason, coercion

[69] Megan Twohey, "Rudy Giuliani Lashes Out at Black Lives Matter," *New York Times*, 2016, https://www.nytimes.com/2016/07/11/us/politics/rudy-giuliani-black-lives-matter.html?_r=0; David Weigel, "Three words that Republicans wrestle with," *Washington Post*, 2016, https://www.washingtonpost.com/politics/three-words-that-republicans-wrestle-with-black-lives-matter/2016/07/12/f5a9dfdc-4878-11e6-90a8-fb84201e0645_story.html?utm_term=.0bd766ba0ceb.

[70] I am assuming that segregation was deeply unjust and inegalitarian, that the Patriot Act did curb significant individual freedoms, and that "Black Lives Matter" is not an inherently racist slogan. Should the reader disagree with some of these assumptions, I at least hope to have a given a sense of what the structure of these examples should be, and expect that they will be able to find instances that they find more compelling.

[71] E.g.,Heinze, "Hate Speech and the Normative Foundations of Regulation," p. 592.

[72] Mchangama, "The Harm in Hate Speech Laws," p. 99. For a similar incident, see: Jean-Baptiste Jacquin, "L'appel à boycotter Israël déclaré illégal," *Le Monde*, 2015, http://www.lemonde.fr/police-justice/article/2015/11/06/l-appel-au-boycott-de-produits-israeliens-est-illegal_4804334_1653578.html

[73] Brettschneider, *When the State Speaks, What Should It Say?*, pp. 6–7, 13, 81.

diminishes individual freedom or autonomy in a way that merely expressive or persuasive power does not.[74]

Second, the coercive power involved in misdirected banning reduces individual freedom because of *what* it keeps individuals from doing, namely publicly voicing considerations that matter to them. As discussed in Section 3, being free to publicly raise such considerations is valuable not just intrinsically, but also because it can help one hold exercises of government power accountable. Consequently, by punishing non-hateful public speech, misdirected banning risks significantly undercutting the political freedom of its targets.

Now, we should not overstate the moral asymmetry between misdirected banning and misdirected counterspeech. In other words, it is important to note that although it does not employ coercive force, misdirected counterspeech too can impede the freedom of its targets. To see this, consider that authoritative counterspeech is capable of disabling speech acts. Indeed, in Section 3, the basis for recommending state counterspeech against hate speech was that it can prevent the speech act of refuting the public assurance of dignity. In the case of misdirected counterspeech, state counterspeech diminishes the target's perceived authority, and thereby unfairly reduces her capacity to *convince* her audience through non-hateful speech. Insofar as the ability to publicly raise concerns and convince one's listeners is instrumental to one's effective political freedom, then, misdirected counterspeech is inimical to such freedom.

Nevertheless, even with this qualification, the crucial point remains that misdirected counterspeech does not disable all speech acts associated with the targeted utterance. For instance, in voicing her non-hateful concerns, the speaker can still successfully *disclose* her worldview, *share* her experiences, and *expose* others to new arguments. Moreover, the speaker can *contest* the hypocritical state's claim that her views are unjust or inegalitarian, by offering a rival understanding of these values. For example, when high profile politicians claim that the slogan "Black Lives Matter" is racist, members of Black Lives Matter still can (and do) contest that claim: they can argue, say, that when one racial group faces special injustices, achieving racial equality requires paying special attention to the interests of that group. Such speech acts at least help third parties better understand the speaker's non-hateful concerns, so that some of her discursive power to advance her interests is preserved.[75] Hence, a morally important asymmetry persists: while misdirected counterspeech salvages some valuable speech acts associated with the targeted utterance, misdirected banning suppresses them all.

5.2. Can Bans Avoid the Objection?

In Section 5.1, I considered an objection to state counterspeech, and argued that it can be turned more forcefully against hate speech bans. Because this objection cuts deeper for bans than for counterspeech, proponents of bans might be tempted to deny its basic premise: namely, that conditions are so non-ideal that the state is likely to be ill-motivated, or hypocritical. Waldron voices such a reply. Put differently, he suggests that the objection under consideration—that if we allow some hate speech bans, the state will ban some things

[74] E.g., Christine Korsgaard, *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), pp. 140–43; Joseph Raz, *The Morality of Freedom* (Oxford: Oxford University Press, 1986), pp. 377–78.

[75] Notice a difference with speaking back against *hateful* speech. State counterspeech reduces the persuasiveness of hate speech because it has authority, and because it has the force of better argument on its side. But in the case *misdirected* counterspeech, the speaker against whom the state speaks may well offer more compelling reasons. So she retains some persuasive force in a way the hate speaker does not.

that it should not—is overblown. Like many slippery-slope arguments, it relies on unsubstantiated empirical claims, which are overly pessimistic about states' motivations.[76]

Against Waldron, one line of response asserts that these claims *are* empirically substantiated, and that states do tend to use hate speech bans improperly.[77] To avoid relying on such empirical claims about states' motivations, I will pursue a different response, which focuses on the epistemic difficulties involved in applying hate speech regulations. More specifically, I will argue that there are conceptual reasons, grounded in the vagueness of the harm in hate speech, which explain why states are likely to misidentify instances of hate speech, and hence—if implementing bans—to ban non-hateful speech.

To see this, let us reconsider and examine more closely the harm in hate speech. Waldron is at pains, firstly, to distinguish the harm that warrants legal regulation from the notion of offense. This is because offense, even deep offense, is too widespread. "Especially in a multifaith society, religion is an area where offense is always in the air. Each group's creed seems like an outrage to every other group".[78] Consequently, religious freedom cannot be preserved without freedom to offend.[79] A second desideratum for a speech-induced harm which we can safely recommend banning, Waldron suggests, is that it must not be vague.[80] If it is vague—i.e., if it is often indeterminate whether the harm obtains or not—then there is a conceptual reason to expect that whoever is implementing the ban will often make mistakes, and ban non-hateful things. Put differently, the harm's vagueness constitutes an underlying cause for why legislators and judges will make mistakes.

Putting these two desiderata together, a minimal condition for the harm in hate speech to be one which we can safely recommend banning is that it be clearly, rather than vaguely, distinct from offense. For Waldron, the harm of assault on dignity, where dignity is the status of being an equal citizen in good standing, satisfies this condition. Being offended is a subjective or psychological harm, which manifests itself in felt experiences like shock or distress.[81] By contrast, Waldron claims, assault on dignity is an objective harm, which concerns "how things are with respect to [one] in society".[82] Whether one's dignity is assaulted depends fundamentally on whether others and their utterances are actually rejecting one's good standing. Conceivably, there might be cases where one feels as though this is the case even when it is not. Conversely, while shock and distress often accompany assaults on dignity, they are neither constitutive nor necessary consequences of such assaults. Indeed, some targets of dignity-denying speech may have a thick skin and not feel distressed. But this does not change the fact that they have been harmed, and their dignity assaulted. Hence, because of this qualitative distinction, Waldron is confident that the distinction between offense (which should not be banned) and assaults on dignity (which should be) is clear, not vague.

Ultimately, however, this move fails to dissipate the vagueness of the harm in hate speech. A preliminary worry is that, as discussed previously, Waldron emphasizes how hate speech attacks dignity by undermining the public assurance thereof. But assurance, it might

[76] Waldron, *The Harm in Hate Speech*, p. 202.

[77] Mchangama, "The Harm in Hate Speech Laws," pp. 97–99.

[78] Waldron, *The Harm in Hate Speech*, p. 127.

[79] Ibid, p. 130.

[80] Ibid, pp. 136-143. This may seem inconsistent with Waldron's insistence elsewhere that vague laws can be adequately action-guiding. But there is no inconsistency. When Waldron raises the objection that vague laws might suppress too much behaviour, he replies that this may be a valid consideration with free speech, but not in the cases he discusses there (driving, torture). See Jeremy Waldron, "Vagueness and the Guidance of Action," in Andrei Marmor and Scott Soames (eds.) *Philosophical Foundations of Language in the Law* (Oxford: Oxford University Press, 2010), pp. 58-82, at pp. 75–77.

[81] Here, I use "subjective" and "psychological" interchangeably.

[82] Ibid, p. 106.

seem, is a feeling. So undermining the assurance of dignity might appear to be a psychological or subjective harm after all: it consists in undercutting citizens' confidence in their good standing. This initial worry is too quick. 'Assurance' is ambiguous between the feeling of confidence and the communicative act that aims to give confidence. It is plausible that Waldron is concerned with the latter. Thus, he might say: 'A speech act undermines the public assurance of dignity when, by publicly denying the standing of others, it refutes communicative acts which state that all citizens are in good standing. And it is an objective matter whether it succeeds in doing so.'

The main problem lies elsewhere. Even if assaults on dignity are objective harms, and therefore of a different ontological kind from offense, the distinction remains *epistemically* vague. Put differently, even if they *are* two different kinds of things, the way in which we find out whether they are occurring often do not allow us to distinguish the two.[83] This is because, intuitively, the way to determine whether an utterance assaults the assurance that all citizens are in good standing is to consider how its targets experienced it. Did they feel a combination of, say, shock and outrage, shame and humiliation, which combined to produce an overall feeling of not belonging, of being excluded? And using these subjective indicators simply is looking at whether someone has been offended.

One might think that the subjective hurt resulting from dignitarian assaults is phenomenologically distinct from mere offense. But this is implausible. It is perfectly conceivable that someone whose dignity has not been assaulted can have a qualitatively similar psychological response to a target of dignity-denying speech. For instance, a believer confronted with an offensive image of Jesus may well experience shock, outrage, shame and humiliation, and feel that being humiliated thus denigrates her social standing.[84]

Sensing this problem, Waldron suggests using another epistemic indicator. Instead of trying to distinguish assaults on dignity from offense based on the experiences they induce, the idea seems to be that we should look more directly for objective markers of assaults on dignity.[85] These markers would help demarcate reasonable feelings of having one's standing attacked from unreasonable ones. But what would these markers be? Waldron suggests that merely offensive speech denigrates a person's beliefs, whereas assaults on dignity denigrate the person herself.[86]

Although Waldron finds this distinction clear, it is difficult to sustain. In particular, it struggles to handle the case of religious offense, which he was centrally concerned with distinguishing from assaults on dignity. Indeed, religious believers might respond that their religious beliefs are constitutive of their identity, so that deriding those beliefs is an attack on their persons. Consider the famous 2005 Danish Cartoons, which represented the prophet Muhammad as a bomb-throwing terrorist, or *Charlie Hebdo's* representation of Muhammad as a child-molester. By attacking central figures of Islam, the response would go, these attack the dignity of profound adherents to Islam. So, this objective criterion fails to clearly rule out religious offense.

Now, Waldron mentions a similar response, but dismisses it as "largely an irresponsible attempt" by its advocates "to claim more […] than they are entitled to" by "recklessly

[83] For discussion of ethical epistemic vagueness, see Tom Dougherty, "Vague Value," *Philosophy and Phenomenological Research* 89 (2014): 352–72.

[84] Waldron (*The Harm in Hate Speech*, pp. 113–14) himself concedes this.

[85] Ibid, pp. 114-115.

[86] Waldron (e.g., pp. 86-87) sometimes suggests that this distinction maps on to Darwall's influential distinction between appraisal respect (which varies based on one's estimation of the targets' virtues and vices) and recognition respect (which is owed to everyone, whatever their vices and virtues). Langton, however, argues that some racist speech acts Waldron would want to count as hate speech are consistent with recognition respect. See Stephen Darwall, "Two Kinds of Respect," *Ethics* 88 (1977): 36-49; and Rae Langton "Hate Speech and the Epistemology of Justice," *Criminal Law and Philosophy* 10 (2014): 865-873.

presen[ting] claims about offense as though they were non-negotiable".[87] And when confronted with the Danish cartoons, he (grudgingly) concludes that it is merely a bad offense: attacking Muhammad's dignity is not an attack on Muslims' dignity.[88]

This response is unsatisfying. First, it seems overly uncharitable: we can easily imagine a reasoned case for thinking that attacks on Muhammad also constitute attacks on Muslims. 'Since part of what it is to be a Muslim is to hold Muhammad's life up as an exemplar', someone might say, 'such representations entail that we Muslims *aspire* to be terrorists or child-molesters!' Arguments like this suggest that, even if Waldron were right that many religious beliefs are not constitutive of believers' identity, there is nothing clearly absurd or "irresponsible" in thinking so. At least in religious contexts, then, the distinction between attacks on persons and attacks on beliefs (and accordingly between reasonable and unreasonable claims to having one's dignity assaulted) often remains epistemically vague.[89]

Second, Waldron's response jars with one of the more appealing features of his framework: his methodological commitment to asking what it is like for those on the receiving end of potentially-harmful speech to be in this position. By introducing criteria of reasonableness to weed out unreasonable feelings of exclusion, and dismissing interpretations of these criteria voiced by self-described targets of religious hate speech, Waldron seemingly departs from this commitment. The point is not just to indicate an internal inconsistency. As I argued earlier, this methodological commitment is attractive. It is sensitive to the fact that targets of speech may have a privileged perspective on what may be harmful or not about it. Furthermore, it is also appealing because, even if targets are mistaken regarding its harmfulness, it helps us better understand how they experience the world, and thus enables participants to better engage with each other.

In this section, I have examined the objection that state counterspeech may be misdirected in non-ideal contexts. This objection, I have shown, applies even more forcefully against bans. Moreover, Waldron's attempt to deny this fails. Even if those in charge of banning hate speech are not badly motivated, the basis for banning hate speech remains epistemically vague. Hence, the present objection stands, and continues to stand more strongly against bans. To make this more vivid, let us end with our running example. When it is vague whether citizens' dignity has been assaulted—as for example with the Danish cartoons or *Charlie Hebdo* cartoons—the hate-allowing state may make mistakes regarding whether to speak back or not. Suppose it does not speak against *Charlie Hebdo*. Instead, it denounces those who publicly claimed that 'those racists got what they deserved'. And suppose this is a mistake. Even then, those against whom the state speaks will not, by contrast with the banning case, be entirely silenced. In such cases, we can at least hope to achieve a better understanding of why some citizens, to use Sacco's phrase, can't "take the joke". And if some of *Charlie Hebdo*'s cartoons did constitute assaults on dignity, it may help us appreciate that this was so.

6. Conclusion

What should we do about public expressions of degrading or hateful views in stable democracies? (Section 1) Waldron's account, which favors bans, constitutes a promising starting point for answering this question (Section 2). Nevertheless, his argument relies on an

[87] Ibid, p. 131.

[88] Ibid, p. 126.

[89] For similar points, see Melissa Williams, "The Uneasy Alliance of Group Representation and Deliberative Democracy," in Will Kymlicka and Wayne Norman (eds.) *Citizenship in Diverse Societies* (Oxford: Oxford University Press, 2000), pp.124-152, at pp. 137–38; Gelber and McNamara, "The Effects of Civil Hate Speech Laws: Lessons from Australia," p. 337.

idealized picture of bans, and an overly bleak understanding of counterspeech. I have aimed to remediate this shortcoming, and thereby offer a qualified defense of counterspeech.

On Brettschneider's more refined understanding of counterspeech, the state authoritatively speaks back against hate speech despite not banning it. Given the freedom-related costs involved in banning hate speech, especially in conditions of widespread political alienation, such counterspeech seems a preferable way of promoting dignity and justice (Section 3). Nonetheless, advocating state counterspeech might seem naïve. In non-ideal conditions, it seems likely to achieve both too little (Section 4) and too much (Section 5). My strategy in responding to these objections has been not just to qualify them, but also, and primarily, to argue that counterspeech and bans are companions in guilt. First, bans too have trouble blocking the dignitarian harms produced by coded hate speech. Moreover, bans seem just as liable to be misdirected as state counterspeech. In fact, in both cases, these objections are more problematic for bans than they are for counterspeech. Bans, put differently, are the guiltier party.

In light of this argument, we should be far less enthusiastic than Waldron at the prospect of banning public hate speech. However, this conclusion must be qualified. First, as I stated at the outset, my defense of counterspeech remains a pessimistic one. As the companions-in-guilt structure of my arguments suggests, we should not be overjoyed about accepting state counterspeech. One of the unhappy conclusions of this article is that both options under consideration face serious limitations.

Second, as I noted earlier, I do not purport to have shown that bans on public hate speech are never warranted. Arguing that the costs typically associated with counterspeech also arise more powerfully against bans generates a case against bans, and in favor of counterspeech. But this case can be challenged. First, there is enormous variation in the forms that abhorrent speech takes, and in the contexts where it occurs. Inevitably, I have focused on a restricted set of cases and have done so against the background of simplified political contexts—in particular, cases of public discourse (as defined in the introduction) in stable democracies. For some forms of abhorrent speech, or some types of contexts, the companions-in-guilt arguments I have supplied may not apply. Moreover, I have been focusing on how hate speech constitutes an assault on the public assurance of dignity. But for some of the causal harms of hate speech, analogous companions-in-guilt arguments, which would rely on substantial empirical investigation, may be less successful. Further research is therefore needed to better understand the precise conditions under which state-based counterspeech is effective.

Still, an important upshot remains: insofar as we are concerned with upholding the assurance of dignity in stable democracies against the challenge posed by public occurrences of hate speech, there are good reasons for thinking that state counterspeech is a preferable solution to bans.[90]

Philosophy Department, University of Cambridge
ml759@cam.ac.uk

[90] Many thanks to Clare Chambers, Joanna Demaree-Cotton, Steven Gubka, Jess Kaplan, Rae Langton, Tamas Szigeti, Jens van 't Klooster, and two anonymous reviewers from *Social Theory and Practice* for extensive comments on previous drafts of this paper. The paper also benefited from helpful discussions with Fabien Cante, Karamvir Chadha, John Filling, Ralph Weir, and participants at the Cardiff Workshop on Religion, Hate, and Offence in a Changing World and at the Cambridge Graduate Seminar. This paper is based on doctoral research that was funded by an Arts and Humanities Research Council (AHRC) studentship. I gratefully acknowledge this support.