Hateful counterspeech

Ethical Theory and Moral Practice
https://doi.org/10.1007/s10677-022-10323-7

Hateful Counterspeech

Maxime Lepoutre[1]

Accepted: 22 August 2022

Abstract
Faced with hate speech, oppressed groups can use their own speech to respond to their verbal oppressors. This "counterspeech," however, sometimes itself takes on a hateful form. This paper explores the moral standing of such "hateful counterspeech." Is there a fundamental moral asymmetry between hateful counterspeech, and the hateful utterances of dominant or oppressive groups? Or are claims that such an asymmetry exists indefensible? I argue for an intermediate position. There *is* a key moral asymmetry between these two forms of speech. But, this asymmetry notwithstanding, hateful counterspeech is capable of enacting serious harms—and so, contrary to what many legal theorists have argued, it is in principle an appropriate object of legal regulation. I begin by considering the central argument for thinking that hateful counterspeech is not seriously troubling. This argument holds that oppressed groups lack authority—and, by extension, "speaker power." Yet this argument, I suggest, sits in tension with the fact that low-status members of dominant groups can, through their utterances, seriously harm members of oppressed groups. Philosophers of language have developed sophisticated arguments to explain this last phenomenon: they have argued that speaker power is relativised to particular jurisdictions; that it can be acquired dynamically in local settings; and that it is socially dispersed. I argue that, in light of these arguments, it appears that hateful counterspeech, too, can generate serious harms. Nevertheless, I show that this conclusion is compatible with recognising a crucial moral asymmetry between hateful counterspeech and the hate speech of oppressors.

Keywords Hate speech · Counterspeech · Authority · Power · Freedom of speech · Speech-act theory

✉ Maxime Lepoutre
m.c.lepoutre@reading.ac.uk

[1] Politics and International Relations, University of Reading, Reading, UK

Published online: 27 October 2022

1 Introduction

In his 1980s sermons to the Nation of Islam, its controversial leader Louis Farrakhan regularly expounded the theological doctrine of the "white devils." The white man, Farrakhan declared, is "nothing but a devil in the plainest language. The arch-deceivers of the planet earth. The number one hater, murderer, killer, liar, drunkard." He is "100% wicked," Farrakhan continued; "he's a snake of the grafted type, and if he's allowed to live, he'll sting someone else" (cited in Gardell 1996: 56, 148; see also Farrakhan 1989).

These sermons illustrate a morally complex phenomenon. At first sight, they resemble what is commonly called "hate speech." Although definitions of hate speech vary widely, hate speech is commonly characterised as speech that communicates or otherwise promotes the basic inferiority of its targets, due to their race, ethnicity, sexual orientation, or other social group membership, often with a view to stirring up animosity or antipathy towards them.[1] Farrakhan's sermons may seem to satisfy these conditions. They vilify a group based on their race—portraying them as essentially morally inferior ("nothing but a devil," "100% wicked"), and sometimes even as subhuman ("snake"). And, as a result of this vilification, these sermons appear to urge antipathy towards this group ("if he's allowed to live, he'll sting someone else").

Yet Farrakhan's sermons also seem importantly different from paradigmatic cases of hate speech: they target a dominant racial group; and they emanate from a member of a racial group that is systematically marginalised and oppressed by that dominant group. Indeed, these hateful utterances are arguably deployed as a response to, or way of countering, hateful treatment by the dominant group. We might therefore call this kind of speech "hateful counterspeech."

The case of hateful counterspeech is deeply divisive in public debate. For some, it is "absurd" to compare Farrakhan's assertions, morally speaking, with standard cases of hate speech (Wise 2018). To others, *not* treating them as comparable constitutes an unsustainable double standard.[2] While the former position is generally expressed by left-wing commentators, and the latter more often emanates from the right, the disagreement actually crosses partisan divides. In 2019, for example, the Southern Poverty Law Center (SPLC), a left-learning organisation with a long history of successfully prosecuting white supremacist hate crimes, included Farrakhan and the Nation of Islam on its map of hate groups. In doing so, the SPLC suggested that Farrakhan's inflammatory rhetoric could be seriously harmful in a way that rendered it, in some meaningful way, comparable to standard cases of hate speech.

This public debate has a counterpart in legal and philosophical analyses of hate speech. Hate speech, as characterised above, involves communicating or otherwise promoting the basic inferiority of its targets, due to their membership in a particular social group. This raises the question of what counts as a relevant social group. There are several questions here—but the one that will constitute my focus in this essay is whether the definition of hate speech should only refer to *vulnerable* groups (e.g., racial minorities, sexual minorities), or whether it should include *dominant* groups too.[3]

[1] See, e.g., United Nations (1965), Waldron (2012: 56–57), Simpson (2013a: 701–702), Lepoutre (2021: 86–88), Howard (2019a: 94–95).

[2] For examples, see Richardson-Self (2018:2–4).

[3] See Brown (2016: 2017a) for an overview.

Legal definitions typically take the expansive view, which does not distinguish between vulnerable and dominant groups. For example, Canada's *Criminal Code* (1985: s.319) penalises speech that wilfully promotes hatred based on "colour, race, religion, ethnic origin or sexual orientation, gender identity". The same is true in France, the United Kingdom, Australia, New Zealand, and the Netherlands—to cite a few cases. According to this way of defining hate speech, hateful counterspeech such as Farrakhan's could in principle qualify.

But legal theorists and philosophers have increasingly challenged this view. Katharine Gelber (2019: 611), for instance, suggests that hate speech targets individuals based on their "perceived membership of a *marginalised* group." Likewise, Richardson-Self (2018: 2–3) raises concerns about the ostensible neutrality of Australian hate speech law—the fact that it does not distinguish between vulnerable and dominant groups. On this more restrictive view, it is arguably a mistake to classify hateful counterspeech, such as Farrakhan's sermon, as hate speech. Matsuda (1989: 2358) is explicit about this implication: according to her, "hateful verbal attacks upon dominant-group members by victims [are] permissible."

This disagreement is not a superficial debate over terminology. At its heart is a substantive question about the conditions under which speech can harm its targets. As lawyers and philosophers have long suggested, hate speech can give rise to serious harms.[4] Perhaps the most widely cited harm is that it risks inciting antipathy and even violence towards its targets. But there are many others. Hate speech, it has notably been argued, can generate profound psychological distress (Delgado 1982), erode its targets' dignity (Waldron 2012), enact norms that discriminate against them (McGowan 2019: ch.7; Gelber 2018), or even silence their speech (Maitra and McGowan 2012). These serious harms, to many, are sufficiently weighty to warrant legally regulating hate speech—at least, provided we have no better way of responding to them.[5]

Thus, the substantive question underpinning the definitional question is the following: Is speech that communicates or otherwise promotes the basic inferiority of a target group seriously harmful even when it is directed by vulnerable groups at dominant groups (as in hateful counterspeech)? Or is it seriously harmful only when directed by dominant groups at vulnerable groups (as in standard cases of hate speech)?

I will argue for an intermediate position. In agreement with theorists such as Richardson-Self, I will suggest that there *is* a crucial harm-related asymmetry between these two categories of speech, such that they are not morally equivalent. But I will also show that, nevertheless, hateful counterspeech is capable of generating serious harms—and so, in keeping with many existing legal regulations, it remains in principle an appropriate object of legal regulation as a form of hate speech.

My argument will proceed as follows. I begin by introducing an intuitive argument for doubting my thesis that hateful counterspeech can generate serious harms. This argument holds that members of vulnerable or oppressed groups lack the authority of their more dominant oppressors, and therefore have less power as speakers (Sect. 2). Next, I raise a problem for this argument: it seems incompatible with the fact that low-status members of dominant

[4] See Gelber (2018: 394) for a recent statement of this view, and Maitra and McGowan (2012) for an overview.

[5] This last clause is important. Some theorists agree that the harms in hate speech are sufficiently weighty that legal regulation could in principle be justified. But they insist that, in practical terms, non-legal remedies are more effective. My argument takes no position on this "effectiveness" question. See Howard (2019b: 242–54) for extended discussion.

groups can, through their utterances, seriously harm high-status members of vulnerable groups (Sect. 3). The paper then considers three attempts at responding to this problem—i.e., three attempts at explaining how low-status members of dominant groups may have the power to seriously harm with their speech (Sects. 4–6). In each case, I will argue that these considerations *also* suggest that hateful counterspeech can generate serious harms. This conclusion, I will show, is nonetheless compatible with maintaining an important moral asymmetry between hateful counterspeech and the hate speech of dominant groups. Lastly, I will explore what this moral conclusion means for legislative policy (Sect. 7).

Before proceeding, it is important to clarify what this argument does and does not purport to show. As I have just stressed, it does not suggest that traditional hate speech, and hateful counterspeech, are equally harmful. Nor does it entail that they are equally frequent—they are clearly not. Finally, it does not entail that all members of dominant groups who complain that they have been seriously harmed by "hate speech" have actually been seriously harmed. Put differently, it is compatible with thinking, with Richardson-Self (2018: 2–3), that allegations of "white vilification" are often unfounded. But even with these qualifications, an important conclusion remains: namely, that hateful counterspeech *can* produce harms that are sufficiently serious to warrant legal regulation; and this, as we will see, has meaningful implications for hate speech law.

2 The Case for Unequal Speaker Power

Why deny that hateful counterspeech, unlike paradigmatic instances of hate speech, can be seriously harmful? Put differently, why think that hate speech is seriously harmful when directed by dominant groups at vulnerable groups, but not when directed at dominant groups by vulnerable groups?

An influential answer relates to authority. As speech-act theorists have long observed, what exactly one can achieve with one's speech (one's "speaker power") depends importantly on background conditions ("felicity conditions") (e.g., Langton 2009: ch.1; McGowan 2019). One such background condition is the speaker's *authority*.

What is speaker authority? This is a notoriously thorny question. In this paper, I will be relying, in the first instance, on Mary Kate McGowan's influential characterisation. According to McGowan,

> to have authority in some contexts is to be able to do things that other people cannot do and it is to be able to do these things because one socially counts as having the power to do them. In other words, one is able to do these things because one has conferred upon her by others a certain status (McGowan 2019: 65).

There are three important elements to unpack here. The first is that authority is an important *determinant of speaker power*. What a given person can do with her speech depends importantly on her authority.[6] Below, I will elaborate on the more precise senses in which speaker power depends on authority.

Second, authority is a determinant of speaker power *that emerges from a person's socially constituted status*. As McGowan (2019: 66–68) emphasises, authority involves a status conferred by others: it involves having the power to do things with one's words because of what relevant others take one to have the power to do. As we will see, this status can be conferred

[6] See also Langton (2018: 123).

in formal ways (e.g., when one is appointed to an official position) as well as more informal ways.

Third, authority emerges from a *comparatively high* status. McGowan distinguishes authority from "standing." Both authority and standing involve a status conferred by others. But there is a difference in degree: whereas standing is a status that everyone or nearly everyone might enjoy in the relevant social context, authority is a status that is comparatively high.[7] To have authority, then, is to be able to do certain things with one's words that not everyone can do, in virtue of being perceived by relevant others as comparatively empowered.

Consider, for example, marriage. Not everyone has the power to pronounce two people married. Whether someone can pronounce two people married in the Catholic Church depends on their having the authority to do so. If an overenthusiastic wedding guest dashes to the altar and says "I hereby pronounce you husband and wife," this will leave the couple's marital status unchanged. In contrast, things are different if an ordained Catholic priest utters those same words. This is because such a priest has authority that the wedding guest lacks, which itself emerges from his comparative and socially constituted status: the priest, unlike the enthusiastic guest, is formally recognised by the Catholic Church as having the power to pronounce two people married (Langton 2009: 50).

Note that authority needn't always involve such a formal or institutional position. An eight-year-old's parents can order him to bed. But his five-year old sibling arguably cannot. Again, this is because the younger sibling lacks the authority to do so, where that authority is a comparative status conferred by others. Here, the relevant others might simply be the various members of the family.

Feminist philosophers and legal theorists have extended this analysis to the harmful effects of speech—including the harmful effects of hate speech. Whether and how hate speech can harm depends partly on the speaker's authority, which itself depends on how they stand in society (Langton 2018: 123–29).

The precise nature of this dependence varies based on the kind of harm in question. Some harms have a very "direct" connection to hate speech. This is perhaps most notably the case with harms stemming from the "illocutionary" dimension of hate speech: hate speech can *constitute* harms, such as subordination or silencing (Langton 2018). Yet hate speech can also harm more indirectly, via its "perlocutionary" effects. These perlocutionary harms are less direct in that the harm is not literally constituted by hate speech—instead, it is a causal effect resulting from it. In some cases, the causal chain connecting speech and harm is fairly long, and depends on a number of intermediate steps. For example, hate speech can incite or motivate acts of physical violence, via a causal chain that goes through the mental reactions of listeners (Wilson and Kiper 2020).[8]

When discussing the significance of authority, feminist philosophers of language have predominantly focused on more direct harms. In particular, authority is often regarded as

[7] See also Langton (2018: 138–40),for whom authority is "relative to comparative rivals".

[8] I understand the harm of "inciting violence" as a perlocutionary effect of hate speech. That is, I take this to be a causal harm of hate speech: the harm of causing (by encouraging or otherwise motivating) others to engage in violent acts.

a necessary condition for enacting illocutionary harms such as subordination or silencing (Langton 2009: 62; Maitra 2012: 99–100).[9]

This, as Langton (2018: 137–38) notes, is not true of all harms associated with hate speech. When it comes to indirect causal harms, for instance, authority is seldom considered a necessary condition. The thought here is that, even without authority, a hate speaker could conceivably cause psychological trauma, or even induce others to act violently.

Nevertheless, authority remains a significant determinant of harm even in these latter cases. Even when hate speech is not strictly speaking necessary for a causal harm to occur, it can still vastly amplify its likelihood and severity. As Langton (2018: 138) acknowledges, "a hate-filled tirade […] will be all the more wounding […] when hate speech has authority." Likewise, Richard Wilson and Jordan Kiper (2020: 94–100) have argued, drawing on extensive empirical evidence, that when someone incites violence against a group, their socially constituted status matters greatly. Other things being equal, a head of state or charismatic leader's words are more likely to succeed in motivating listeners to engage in violent acts than those of an ordinary citizen.

In sum, the foregoing considerations suggest that the presence or absence of authority is often deeply relevant to whether, and how seriously, hate speech harms its targets. This, to many philosophers and legal theorists, gives rise to an argument for thinking that hateful counterspeech, unlike "standard" hate speech, is not seriously harmful.

Why might this be? As we have seen, authority is a source of power to do things with one's words that derives from one's comparative and socially constituted status. Now, part of what it means for a group to be oppressed or vulnerable *just is* that its members tend to have a lower status than other groups—in particular, than members of the groups that oppress them. Accordingly, members of vulnerable or oppressed groups would appear to lack authority relative to members of dominant groups. For example, when the Grand Wizard of the Ku Klux Klan affirms the essential inferiority of people of colour, he speaks with a socially constituted authority that Farrakhan lacks, when he affirms the essential inferiority of white people.

Insofar as the harmfulness of one's speech depends importantly on one's authority, this in turn suggests that hateful counterspeech may lack the capacity to harm seriously possessed by standard cases of hate speech. More concretely, the idea might be that Farrakhan's speech, unlike the Grand Wizard's, cannot constitute certain illocutionary harms (such as subordination or silencing) because he lacks authority. As for causal or perlocutionary harms, the thought might be that even if it is strictly speaking possible for Farrakhan's speech to provoke hatred or violence towards the target, or to inflict psychological pain on the target, it is exceedingly unlikely that it will succeed in doing so; and even if it does, the harm generated is likely to be far less severe or serious than in "standard" cases of hate speech.[10]

Matsuda (1989) notably exemplifies this argument about severity. She concedes that "white devil" comments such as Farrakhan's (or, in her own example, Malcolm X) could potentially lead to "hurt" (1989: 2361). But, due to differences in "relative power" between

[9] Not all illocutionary acts require authority. More specifically, Langton (2009) argues that verdictive and exercitive speech-acts (such as subordination and silencing) require authority. Now, some philosophers disagree with this last claim: McGowan argues that harmful exercitives can be enacted without authority. I return to this disagreement in Sect. 6. For now, I simply wish to highlight that authority is commonly taken to have a significant influence on whether, and how seriously, hate speech harms.

[10] I return to, and critically assess, these claims about the likelihood of provoking violence in Sect. 6.

white people and people of colour in American society, she suggests that the hurt would be of such a "different degree" that it would not warrant legal intervention (1989: 2361, 2362n15).[11]

The general idea is therefore the following. If the capacity to harm seriously with one's speech depends importantly on authority, and if hateful counterspeakers, unlike "standard" hate speakers, lack authority (due to having a comparatively low status relative to the dominant groups they target), we have reason to doubt that hateful counterspeech is liable to generate serious harms—i.e., harms sufficiently significant to warrant legal intervention.

3 The Problem of Low-Status Hate Speech

We have considered an argument, based in the significance of authority, for thinking that hateful counterspeech might not be seriously harmful. This argument, however, faces a problem relating to "low-status" hate speech.

To appreciate this problem, consider that, even within a dominant social group, some individuals can have a low status. In a capitalist and patriarchal society, for example, an impoverished white woman may have a fairly low social status, despite belonging to a dominant racial group. In some cases, that status may even be lower, overall, than that of some members of a vulnerable racial group. For instance, some impoverished white women may have a lower status, overall, than a very influential member of a racial minority such as Farrakhan.[12]

Even so, it is generally held that racist hate speech uttered by "low-status" members of a dominant racial group *can* be seriously harmful. In support of this position, Ishani Maitra (2012: 100-01) offers the following example:

An Arab woman is on a subway car crowded with people. An older white man walks up to her, and says, "F***in' terrorist, go home. We don't need your kind here." He continues speaking in this manner to the woman, who doesn't respond. He speaks loudly enough that everyone else in the subway car hears his words clearly. All other conversations cease. Many of the passengers turn to look at the speaker, but no one interferes.

In this scenario, Maitra (2012: 100-02) observes, the older white man's utterance arguably constitutes the harm of subordinating the target—even though it comes from an "ordinary," or indeed "low-status," speaker.

Maitra is not alone here. The idea that "low-status" racist speech can enact serious harms is in fact a recurrent theme of feminist philosophy of language (e.g., Barnes 2016: 242–45; Langton 2018: 133–38; McGowan 2019: 66–70). Nor is this phenomenon considered unusual. As Langton notes, instances of harmful low-status hate speech are sadly commonplace. She cites, for example, anonymous graffiti accusing Muslims of being terrorists; hate

[11] See also Richardson-Self (2018: 17), who highlights inequalities of authority when explaining why complaints about "white vilification" are often misguided; and Gelber (2018: 401-03), for whom authority considerations yield one reason (albeit not the only reason) to believe that hate speech directed by a white person at a person of colour may generate serious harms in a way that cannot be achieved by hateful counterspeech.

[12] In making these comparisons, I do not mean to suggest that we can easily rank everyone according to their overall status. I am making a more modest point: that, in at least *some* cases, it seems intuitive that a member of a dominant racial group has lower status than a member of a vulnerable racial group. Yet, as I explain below, it also seems intuitive that the former's racist utterances could seriously harm the latter. This is enough to raise the apparent puzzle that I set out to explore.

mail targeting African-American athletes; or, more recently, the torrent of racist abuse Gina Miller received from members of the public following her anti-Brexit legal action (2018: 128–36).

Thus, many cases of hate speech involve ordinary, or even low-status, members of dominant groups; and, importantly, these cases are commonly considered to be seriously harmful—sufficiently harmful, to many commentators, that they could warrant legal regulation (e.g., Brown 2017b: 305).

On the face of it, however, this observation sits in tension with the argument regarding hateful counterspeech introduced in Sect. 2. The argument for thinking that hateful counterspeech is not seriously harmful appeals to the claim that the harmfulness of speech depends importantly on the speaker's authority. Yet, prima facie, low-status members of dominant groups may themselves seem to lack authority. Authority involves having power in virtue of possessing a comparatively high status conferred by others. But, by definition, low-status hate speakers from dominant groups have a status that is comparatively *low*.[13]

The upshot is that, absent further qualification, the argument for thinking that hateful counterspeech is not seriously harmful generates a problem: it threatens to undermine the intuitive and widely held belief that the hateful utterances of low-status members of dominant groups ("low-status hate speakers," for short) can be seriously harmful. *If* the power to harm seriously with one's speech depends on authority (either because it strictly necessitates authority, or because serious harm is possible but very unlikely without authority), and *if* low-status hate speakers lack authority, then there is reason to doubt that low-status hate speech will be seriously harmful.

Where does this upshot leave us? To many people, the idea that low-status hate speech can generate serious harms is extremely intuitive. Indeed, as mentioned above, many feminist philosophers of language are strongly committed to this idea. Accordingly, those who want to argue that hateful counterspeech is not seriously harmful may try to deny this apparent implication. In other words, they may try to show that, notwithstanding the considerations outlined above, low-status hate speakers *do* have the power to generate serious harms with their speech. The rest of this essay will consider the three most promising explanations that have been put forward for why this may be: the first relates to the jurisdiction of speaker power (Sect. 4); the second highlights the dynamics of speaker power (Sect. 4); and the third emphasises the social dispersion of speaker power (Sect. 6).

In each case, I will examine to what extent these explanations also extend to hateful counterspeech. This will yield two conclusions. The first is that, on closer inspection, all three explanations for why low-status speakers from dominant groups can generate serious harms *also* suggest that hateful counterspeech can generate serious harms. However—and this is the second conclusion—this is nevertheless compatible with maintaining an important moral asymmetry between the two, which results from the social dispersion of speaker power.

[13] One might worry that this claim is too quick, and that this reveals an immediate disanalogy between hateful counterspeech and low-status hate speech. According to this worry, hateful counterspeakers have a low status *compared to their target.* By contrast, low-status hate speakers have a low status compared to other members of their dominant group, but a *high* status compared to their target. This, however, needn't be the case. As we saw at the beginning of this section, some low-status hate speakers from dominant groups intuitively have a lower status than some members of the group that they target. And yet, even in such cases, low-status hate speech is still typically taken to harm its targets.

4 The Jurisdiction of Speaker Power

Why might low-status members of dominant groups have the power to harm seriously with their words? One answer relates to the jurisdiction of authority. One possesses authority (and with it, speaker power) for a particular group of people, at a particular time (Langton 2018: 138–41). Thus, even if someone is relatively low-status in most people's eyes, they may nonetheless be high-status, or authoritative, amidst a smaller jurisdiction. This point is familiar from debates about pornography. Producers of pornography are not typically recognised as authoritative by society at large. Still, they may have higher status in a smaller jurisdiction: say, among adolescent men (Langton 2009: 98).

Importantly, a speaker with authority in a particular jurisdiction can, thanks to that authority, use their speech to harm targets outside of their jurisdiction. Specifically, they can do so provided members of its jurisdiction have meaningful and sufficiently empowered contact with the target (Langton 2009: 97–99). Even if many women are outside of pornography's jurisdiction—they reject its claim to authority—pornography can still harm them via its authority over adolescent men. For example, pornography might instil in adolescent men a desire for sexual violence, which some men might act upon.

This analysis carries over to low-status hate speech. As Langton (2018: 139) explains, "'marginal' figures despised by the majority may, for that same reason, be esteemed figures among the minority." An American Neo-Nazi activist may be reviled by most Americans. Yet, partly as a result, his words may carry weight amongst white supremacists. Hence, if he vilifies people of colour, his speech may generate serious harms—for instance, by motivating some of his followers to act violently towards people of colour.

This does not mean that the Neo-Nazi's speech will be *as* harmful as that of someone who commands authority over a larger jurisdiction. Because of his enormous following, Donald Trump can verbally provoke a large-scale assault on the US Capitol. The Neo-Nazi almost certainly cannot. But this does not undermine the more fundamental point. Given his authority over a particular jurisdiction, the Neo-Nazi's speech has the power to generate serious harms—harms, such as provoking violence, that are generally deemed sufficient to warrant legal intervention. This is true even though the jurisdiction may be small, and even if its members are not particularly empowered. Provided members of the jurisdiction have the power to inflict violence on some members of the target group—a relatively low threshold—the Neo-Nazi's vilifying speech can produce serious harms.

Pointing out that authority is relative to a jurisdiction helps explain how low-status hate speech can seriously harm. However, it *also* undercuts the authority-based argument for thinking that hateful counterspeech cannot generate serious harms. If authority is a matter of having a comparatively high status conferred by others in a particular jurisdiction, then hateful counterspeakers can in principle have authority. This is because, even when speakers from vulnerable groups do not have a comparatively high status in the eyes of most people in society, they can have a comparatively high status within a smaller, or more "local," jurisdiction.

Consider again the case of Farrakhan. He does not have a comparatively high status in the eyes of most people in the US. Nevertheless, as the Nation of Islam's leader, Farrakhan has a comparatively high status conferred upon him by this religious organisation and its members. Thus, he appears to have authority within this smaller jurisdiction.

Another example relates to the anti-Western hate propaganda disseminated by the Islamic State in Iraq and the Levant (ISIL). Members of ethnic and religious minorities who spread this propaganda often occupy an overall marginalised position in Western democracies. But they may nonetheless have authoritative status in some jurisdictions—notably in some online sub-communities. Indeed, in his analysis of ISIL's online hate propaganda, Mohamed Badar explains that its purveyors take on, and are informally granted by a set of readers, the status of representatives of the "general consensus of the true [Islamic] practitioners" (Badar 2016: 397).

So, hateful counterspeech can be authoritative in some jurisdictions. But there is a complication here. *Some* jurisdictions are arguably too trivial to result in meaningful forms of speaker authority—i.e., forms of authority significant enough to enable a speaker to generate serious harms with their speech. The president of a five-member board game society may have authority within the jurisdiction constituted by that society. And that authority may allow them verbally to enact game-related rules for its members. But this local authority intuitively seems too trivial to lend the president of this society the power to produce serious harms with their speech.

Could something similar be said of hateful counterspeakers? Is the jurisdiction of hateful counterspeakers' authority too insignificant? One might advance three reasons for thinking so: first, that the "membership" of its jurisdiction is exceedingly small; second, that this membership does not include the targets of hateful counterspeech; and third, that its members are themselves insufficiently powerful.

Neither reason is decisive. To begin, hateful counterspeakers can have authority within fairly large jurisdictions. In Farrakhan's case, the Nation of Islam comprises not five, but fifty thousand, members.[14] This may well match, and even exceed, the jurisdiction of some low-status hate speakers from dominant groups whose speech is considered capable of engendering serious harms (such as the Neo-Nazi activist mentioned above).

Second, we have already seen—when discussing pornography—that authority can facilitate speech-based harm to targets that are outside of its jurisdiction. Accordingly, the fact that white Americans are not part of the jurisdiction of Farrakhan's authority does not necessarily mean that his speech cannot harm them, via its authority over members of the Nation of Islam.

Finally, as discussed above, the case of the Neo-Nazi also showed that local authority can empower the speaker to generate serious harms *even if* members of its jurisdiction are not particularly powerful. We can apply this insight to hateful counterspeech such as Farrakhan's or ISIL's. If hateful counterspeech vilifies its targets in a way that licenses violence against them (say, by likening them to dangerous animals), and if members of its jurisdiction are able to inflict violence on some targets (again, a relatively low standard of empowerment), then it risks producing harms that are generally considered sufficient to warrant legal regulation. This risk is not merely an abstract possibility. ISIL hate propaganda has notoriously succeeded in provoking violent harm. And there is evidence that the same is true—albeit on a lesser scale—of the Nation of Islam's hate rhetoric.[15]

None of this is to deny that there may be an important difference in frequency between low-status hate speech and hateful counterspeech. The hateful utterances of low-status white

[14] MacFarquhar (2007).

[15] See, e.g., Sanders and Cohen (2011) on the 1973 "Zebra murders".

supremacists are considerably more common, and therefore likely to generate serious harms more often, than hateful counterspeech (Lee 2011: 295). But this does not undermine my core contention here: that, if authority is relativised to a jurisdiction, then hateful counterspeech too can count as having authority. Accordingly, this first account of why low-status hate speakers can harm seriously with their speech *also* suggests that hateful counterspeech can be seriously harmful.

5 The Dynamics of Speaker Power

The previous explanation of why low-status members of dominant groups can have authority (and hence, the power to harm seriously with their speech) extends to hateful counterspeech. It therefore undercuts the authority-based argument for thinking that hateful counterspeech cannot harm seriously. There are, however, alternative ways of accounting for the authority of low-status hate speech. One prominent alternative centres on the dynamics of authority. That is, it emphasises the ease with which authority can be acquired in the context of the utterance. Accordingly, even if the low-status hate speaker initially lacks authority, they can readily gain authority as they speak.

This idea has most fully been developed in Maitra's (2012) account of "licensed" authority. Licensing is a phenomenon whereby a speaker's utterances acquire authority due to the audience's reactions to those utterances. Consider again Maitra's subway example. An elderly white man spews racist hatred at a subway passenger. Other passengers hear this, but do not interject. In this scenario, Maitra argues, the speaker has authority, such that his speech can constitute the harm of subordinating the target. And crucially, that authority is licensed by the other passengers' reactions.

For Maitra, this example illustrates several properties of licensed authority. First, the speaker does not *initially* have authority. He acquires authority as he speaks, because of fellow passengers' reactions (2012: 107). Second, the hearers who do the licensing needn't themselves have any special authority. In the subway example, they are ordinary passengers. Third, licensing does not require hearers to actively do anything. An omission (e.g., not challenging the speaker's utterance) is sufficient (2012: 105-06). Finally, this omission needn't be a sign of tacit agreement. According to Maitra, the passengers' failure to interject can license the subway speaker's authority even if they have "strong reservations about what he says" (2012: 166).

If this account is correct, authority is easier to obtain than one might have thought. One can obtain it as one speaks, provided one's audience does not challenge what one says, even if they disagree with what one says, and even if they lack any special kind of authority. On this account, therefore, a low-status hate speaker such as the elderly white man can readily count as having authority—and so, can seriously harm with their speech (in Maitra's example, by subordinating their target).

However, the point that authority can be licensed in this permissive way is not restricted to utterances that have a particular content. It derives from a general observation about how speaker authority depends on dynamic interactions with hearers. Maitra's main example involves a speaker from a dominant group. But, in principle, these dynamic interactions could take place between a hateful speaker from a vulnerable group and their audience.

Consequently, this account could *also* suggest that members of vulnerable groups can readily obtain the authority needed to generate serious harms with their speech.

The point so far is quite abstract. To make it more concrete, notice that we could construct a subway case that parallels Maitra's, but involves hateful counterspeech instead. In this case, the speaker marks the target as essentially inferior due to their race; *but* the speaker belongs to an oppressed racial group, while the target belongs to a dominant racial group. Suppose, for example, that, in the subway, an African American man directs Farrakhan-like rhetoric at a white woman. Like Farrakhan, the speaker accuses his target of being a "white devil" and a "snake," who, in virtue of being white, is essentially wicked and dangerous. Other passengers hear this, but they remain silent.[16] According to Maitra's account of licensed authority, this silence can confer authority upon the speaker. Thus, following the logic of Maitra's case, the speaker's hateful tirade could have the authority needed to subordinate their target.

One might find it counterintuitive to say that the hateful counterspeaker acquires authority (and hence, the speaker power to subordinate) in this revised subway case. There are three things to say in response. First, recall that, as Sect. 4 discussed, authority is indexed to a jurisdiction. Accordingly, the claim at hand is more modest than it might initially seem. It is not that the hateful counterspeaker has authority in society at large. Rather, it is that he has authority in the local speech context, among parties to the conversation (in this case, among passengers in the subway car) (Maitra 2012: 117).

Nevertheless, even if we focus on this particular jurisdiction, one might still find it implausible to say that the hateful counterspeaker obtains authority. This might be because one thinks that obtaining licensed authority requires more than simply silence from the audience. Brown (2019), for example, suggests that Maitra's account of licensed authority is too permissive, and that two further necessary conditions must be satisfied for authority to be licensed. To begin, it must be clear to the audience that remaining silent constitutes, or at least will be understood as, a form of assent, licensing, or complicity. Moreover, the audience's silence must satisfy a threshold of voluntariness: they must not be remaining silent out of fear that speaking out will expose them to significant danger or to unreasonably high burdens (2019: 212–14).

But—and this is the second response—this more demanding understanding of licensed authority can still accrue to hateful counterspeakers. We could stipulate, in the revised subway case, that audience members know that remaining silent may be perceived as assent, and that their silence does not result from fear of exposure to danger or unreasonable burdens. What is more, these further conditions do not necessarily seem harder to satisfy when the speaker belongs to a vulnerable group than when they belong to a dominant group. Take the voluntariness condition. It is possible, of course, that due to intergroup tensions, some audience members might distrust, and perhaps fear, members of vulnerable racial groups. But it is also possible that the speaker's membership in a *vulnerable* group makes them seem less threatening than speakers from dominant groups. So the point is simply that, on the face of it, the more demanding account of licensed authority proposed by Brown seems similarly applicable to hateful counterspeakers.

[16] I have assigned the genders of the speaker and hearer in this way to parallel Maitra's (2012: 100-01) subway case as closely as possible. Maitra does not specify the race of third-party passengers, and I therefore do not do so either.

What if one rejects even this revised account of licensed authority? The final response emphasises that, for our purposes, we can remain agnostic on whether the idea of licensed authority is ultimately convincing. If this account of how authority can be acquired is incorrect, then it fails to explain why low-status hate speakers can harm seriously with their speech. But if it *is* correct, then it also implies that hateful counterspeech can be authoritative—and so, can enact serious harms. Thus, the fundamental point remains that, *insofar* as this explanation succeeds in accounting for the harmfulness of low-status hate speech, it too risks undercutting the authority-based argument for thinking that hateful counterspeech cannot be seriously harmful.

6 The Social Dispersion of Speaker Power

We have considered two possible explanations for how low-status members of dominant groups can have the power to harm seriously with their speech. Both attempt to show that low-status hate speakers can have authority, and both do so by focusing on "local" contexts: the first suggests that authority can be relativised to a particular jurisdiction (Sect. 4); the second, that it can be acquired via speaker/hearer interactions in the specific context of an utterance (Sect. 5). Yet, absent further development, these explanations *also* suggest that hateful counterspeech could be seriously harmful. The final explanation I will consider here suggests that, to block this implication, one needs to examine how local contexts interact with the broader social context.

Before proceeding, note that this final explanation is best understood as complementing the accounts introduced in Sects. 4–5. One can accept that authority is relativised to a particular jurisdiction, and capable of dynamically evolving through local speaker/hearer interactions *while* also accepting that a full understanding of low-status speaker power requires attending to the way in which the broader social context feeds into the local context. Accordingly, this final explanation is typically aimed not at replacing, but rather at supplementing, the previous explanations.[17]

The core insight of this final explanation is the observation, commonly attributed to Foucault, that power is dispersed throughout society (Fricker 2007: 12). On this view, an individual's power in a specific local setting (e.g., the subway, the online thread) does not simply depend on that individual's intrinsic properties. Nor is it fully explained by the attitudes and actions of other individuals in that local context. Instead, the individual's power also depends crucially on facts about the broader social structure: for instance, facts about the broader norms, conventions, and practices constituted by the often uncoordinated actions and attitudes of very many other agents.

McGowan (2019: 4) and Gelber (2018: 407) have influentially drawn on this insight to explain how low-status hate speakers can harm with their speech. In other words, they suggest that low-status hate speakers' ability to generate serious harms in local contexts is crucially facilitated by the norms, conventions, and practices that operate in the broader social context.

[17] One *could* accept the point about the social dispersion of speaker power while rejecting the claims that authority can be licensed and is indexed to a jurisdiction. But in practice, advocates of the "social dispersion" point such as McGowan (2019: 63–65) or Gelber (2018: 401) often combine these insights.

There are at least two complementary explanations of this phenomenon: the former focuses more on the speaker; the latter on their target. According to the former explanation, the broader social context defines harmful conventions which are then available for activation in local contexts. To see this, consider non-harmful conventions first. The meaning of most words in the English language is, in part, a product of conventions that have evolved over time through large-scale linguistic practices.[18] These broad conventions in turn allow individual speakers to do certain things in local contexts. Consider, to illustrate, the word "birds." When we use this word in a particular setting, we activate broader conventions regarding its meaning, which allows us to refer to birds in a way that may not otherwise have been possible.

We can apply this idea to hate speech. Some background conventions are unjust or oppressive. Consequently, these conventions can empower individuals to do harmful things with their speech. Take, for example, racist slurs in a white supremacist society. A slur directed at people of colour has a particular meaning, and a particular force, partly because of the way it has conventionally been used and associated with broader patterns of racial discrimination.[19] Now, a low-status white speaker, in a local context, may not be personally responsible for that racist history. But when they use the slur, they nonetheless avail themselves of the force that results from it.

Thus, attending to the relationship between the broader context and the local context yields a first account of how low-status hate speakers can harm with their speech. As McGowan (2019: 4) summarises, their speech "taps into" the oppressive power of larger social structures. The broader context supplies harmful tools or resources (e.g., racist terms conventionally imbued with derogatory force) that ordinary speakers can activate, or "bring to bear," in local contexts (see also Gelber 2018: 407; Simpson 2013b: 563).

The broader social context can also affect low-status hate speakers' local power to harm in a second way, which relates to the target. When assessing the harmfulness of a hateful utterance in a specific context, it matters whether the target has been targeted before, in other contexts. Indeed, the harm done by a given hateful utterance may be amplified by the fact that the target has previously been exposed to similar utterances. Put differently, the harmfulness of hate speech may be *cumulative*: it depends on, and is exacerbated by, its frequency (Gelber 2018: 400; Bonotti and Seglow 2019: 594–95).

Take, for instance, the psychological harms of hate speech. As Richard Delgado reports,

> human beings […] whose *daily experience* tells them that *almost nowhere in society* are they respected and granted the ordinary dignity and courtesy accorded to others will, as a matter of course, begin to doubt their own worth (1982: 136–37, emphases added).

Put differently, when someone repeatedly encounters hateful utterances, this risks giving the impression that they are disrespected, not just by an isolated hate speaker, but by broad portions of society. This appearance of generalised disrespect is deeply problematic. First, it risks eroding the target's psychological sense of self-worth in a way that no isolated expression of disrespect could. Second, it risks making subsequent hateful utterances more psychologically hurtful: each local instance of hate speech vividly reminds the target of the generalised disrespect they have elsewhere endured.[20] So, if a low-status hate speaker sub-

[18] For a practice-based approach to meaning, see, e.g., Tirrell (2012).

[19] See, e.g., Langton (2018: 137) and Camp (2013: 345).

[20] See, relatedly, Waldron (2012, ch.4) and Bonotti and Seglow (2019: 598).

sequently targets a member of an oppressed group, their speech may be seriously harmful because of the target's past exposure to similar utterances.[21]

In sum, the broader social context can empower low-status hate speakers to produce serious harms in at least two ways: first, by creating harmful conventions that low-status hate speakers can then activate in local settings (e.g., terms imbued, through conventional use, with derogatory force); second, by repeatedly exposing a target to hate speech, such that they are more vulnerable to being harmed by low-status hate speakers' subsequent utterances.

How does this "social dispersion of power" explanation relate to authority? McGowan (2019: esp. ch.3) denies that being empowered by the broader social contexts constitutes a form of authority. This is because, as we have seen, she takes authority to involve a comparatively high status conferred by others; and, in her view, the mechanisms outlined above do not seem to depend on the hate speaker having such a status conferred by others. Accordingly, she takes the social dispersion of power to support the view that low-status speakers can enact serious harms (including illocutionary harms such as subordination or silencing) without authority.[22]

Yet this interpretation is contested. Some philosophers suggest that we should understand the broader social context, and the way it feeds into local contexts, as *part* of what goes into determining whether a speaker's utterance possesses authority. According to this more expansive view of authority, authority depends not just on jurisdiction, and on local speaker/hearer interactions, but also on broader social norms, practices, and conventions. Barnes, for example, claims that "authoritative speech must be seen to draw upon the social norms of the broader community" (2016: 257). Similarly, Tirrell (2012: 2012) argues that authority can be "diffused" within broad social practices.

While I find McGowan's terminological choice slightly more intuitive,[23] I prefer to remain agnostic on this issue. Even if they disagree on terminology, both sides can and do agree on the substantive point that speaker power can derive from broad social norms, conventions, and practices, in the ways outlined above. It is this substantive point that matters for our purposes: the broad societal sources of speaker power (whether or not we consider them a source of authority) have crucial implications for hateful counterspeech.

As we have seen, the arguments outlined in Sects. 4–5 for why low-status hate speech can seriously harm its targets *also* extend to hateful counterspeech. These arguments are therefore not well-placed to identify a moral asymmetry between hate speech by low-status hate speech and hateful counterspeech. By contrast, it is unclear how the present account could extend to hateful counterspeech. This is because, in an unequal society, the broader

[21] Could frequent exposure instead lead targets to develop a "thick skin," such that subsequent utterances harm them *less*? This is possible. Here, however, I will assume that frequent exposure would usually *amplify* the harm. If members of dominant groups are less frequently targeted, this position bolsters the case for thinking that hateful counterspeech is unable to harm seriously. Since I will be challenging this position, granting this assumption strengthens my overall argument.

[22] See also Gelber (2018: 403).

[23] One reason for this intuition concerns the second mechanism underpinning the social dispersion of power. Prima facie, the fact that a target has endured past exposure to hate speech is more about that target and their social group than about the present speaker's status. Hence, insofar as we find it intuitive to think of authority as deriving from the speaker's status it may seem counterintuitive to describe this phenomenon as a source of authority. Nevertheless, this point is controversial, and the rest of my argument will not depend on it.

social context will typically empower the speech of members of dominant groups, not of vulnerable groups.

To make this more concrete, consider again the case of slurs. As discussed earlier, the historical practices and conventions of a white supremacist society can imbue slurs against people of colour with derogatory force. But it is unclear how this argument could extend to slurs against white people. Historical practices and conventions imbue slurs against people of colour with derogatory force *because* those historical practices and conventions were oppressive and discriminatory towards people of colour. By contrast, in a white supremacist society, historical practices and conventions are characteristically *not* oppressive or discriminatory towards white people. Hence, Mihaela Popa-Wyatt and Jeremy Wyatt (2018: 2899) conclude that when a slur is used against a privileged group (such as "honky" for white people) it cannot mobilise the kind of derogatory force that would attach to a slur against a historically marginalised group (such as the n-word for people of colour). And the same could be said, and for the same reason, regarding Farrakhan's hateful reference to "white devils" in the 1960s.

A similar point applies to the claim that the harmfulness of hate speech is cumulative. In a white supremacist society, people of colour are typically exposed to hateful utterances far more frequently than white people. The case of slurs also helps exemplify this point. Frequency data suggests that, in English, slurs against white people (e.g., "honky," "white devil") are used considerably less often than common slurs against people of colour.[24] Now, racial slurs are just one kind of hateful speech. But this is nonetheless indicative of the idea that hateful counterspeech tends to occur less often than standard hate speech. Insofar as this is the case, the fact that repeated exposure to hateful speech can amplify its subsequent harmfulness, via the cumulative effect Delgado describes, seems less relevant to hateful counterspeech.

The upshot is that, unlike the accounts canvassed in Sects. 4–5, the present account successfully maintains an important moral asymmetry between low-status hate speech, and hateful counterspeech. Nevertheless, as I will now show, it does not follow that hateful counterspeech cannot be seriously harmful—sufficiently harmful, that is, that it could warrant legal regulation.

The reason, in short, is that the present proposal—that the harmfulness of hate speech in a local context depends crucially on the broader context—applies more clearly to some harms than others. We have already seen that the broader social context can amplify some of the perlocutionary harms caused by hate speech. In particular, it can exacerbate its *psychological* harms (e.g., if frequent past exposure makes it more injurious to the target's sense of self-worth; or if the broader context imbues it with especially derogatory meanings). Furthermore, McGowan (2019: 108–113, 140-2, 143−55, 166−83) suggests that numerous illocutionary (or "direct") harms—including *subordination* and *silencing*, as well

[24] Google n-gram data offers tentative support for this claim: in English, between 1960 and 2019, notable slurs against white people (such as "white devil" or "honky") appeared between 30 times and 100 times less often than the n-word. Now, there are limits to what we can conclude from this data: Google n-gram exclusively looks at written uses; and it does not exclude potential non-slurring uses of the search terms (e.g., "honky tonk," or in-group uses of the n-word). Nevertheless, this still offers preliminary evidence of a difference in frequency. Note that I do not mention "white trash" here because it involves a class dimension as well as a racial dimension: typically, it targets *poor* white people. Because of this class dimension, it is not altogether clear that "white trash" targets a dominant group. At any rate, including it would not have meaningfully changed the result: "white trash" appeared between 10 and 25 times less often than the n-word between 1960 and 2019.

as *oppression* and *discrimination*—depend importantly on the local activation of broader social norms and conventions.[25]

But there is at least one important causal harm commonly ascribed to hate speech that seems less dependent on background norms and conventions: namely, the harm of inciting or provoking violence. Hate speech—understood as speech that communicates or otherwise promotes the fundamental inferiority of other groups—is widely taken to have an important connection to intergroup violence. Tirrell (2012: 375), for example, argues that verbally depicting other groups as essentially inferior, and even subhuman ("snakes"; "cockroaches"), may license acts of ethnic violence. Similarly, Wilson and Kiper (2020: 106-07) argue that, because of its capacity to elicit feelings of group-based disgust, hate speech threatens to provoke violence against outgroups.[26]

To see why this matters, compare violence with psychological distress. As discussed earlier, whether the psychological distress inflicted by an instance of hate speech constitutes a serious harm depends on the broader social context. For instance, it notably depends on how frequently the target has endured hateful abuse in other settings. By contrast, physical violence intuitively constitutes a serious harm independently of the broader context. So, the fact that one person inflicted physical violence on another is usually considered, prima facie, to constitute sufficient grounds for legal intervention.

Of course, facts about the broader context can amplify the harmfulness of physical violence. Physical violence motivated by prejudice toward a minority group is arguably more harmful, partly because of its expressive dimension, than physical violence that is not so motivated. But the important point is that even in the latter case, physical violence is typically deemed seriously harmful and (in the absence of excusing conditions, such as self-defence) liable to legal intervention. Thus, if hateful counterspeech—such as Farrakhan's portrayal of whites as "snakes"—motivates violence against some targets, and if the excusing conditions are absent—for instance, the victims were not immediately threatening—then it too seems sufficiently harmful to warrant legal intervention.

Perhaps, however, this is missing the true significance of the broader context for speech-induced violence. I have been considering how the broader context might affect the *harmfulness* of acts of violence. But one might object that its true significance is that it affects the *likelihood* that hateful utterances will provoke violence. As mentioned in Sect. 2, personal charisma or institutional position can increase the likelihood that one's words will successfully provoke violence. One might think that a similar consideration applies to social context. Brown (2017a: 40–41), for example, hypothesises that, in a social context where Muslims are disproportionately targeted by hate speech, individual instances of Islamophobic hate speech could be more effective at increasing hatred against Muslims—and, consequently, at provoking violence against them. The broader suggestion is that, because of the background context, hateful utterances aimed at dominated groups may be more likely to provoke violence than hateful counterspeech.

There are two responses. First, even if hate speech by members of dominant groups proves more likely to motivate violence, it still would not necessarily follow that hateful counterspeech could not be seriously harmful. So long as, in absolute terms, a given

[25] See also Gelber (2018).

[26] See also Howard (2019b: 214n11).

instance of hateful counterspeech remains reasonably likely to provoke violence, this may suffice for it to count as seriously harmful—and so, liable to legal regulation.[27]

Moreover, the exact relationship between social context and success at motivating violence seems empirically debatable. Brown suggests one possible mechanism linking the broader context to the successful incitement of violence. But this is only one possibility among others (2017a: 40–41). Indeed, there are possible mechanisms going in the opposite direction too. For example, if a vulnerable group has persistently been subjected to oppression, some of its members may strongly desire revenge, and feel that they have little to lose in seeking it. In this context, hateful counterspeech directed at their oppressors could conceivably be *more* likely to motivate violence, not less.

This is not a far-fetched possibility. In the "Ballot or the Bullet," Malcolm (1964) described precisely this kind of situation in 1960s America. The persistent oppression of black Americans, he declared, "makes the black community throughout America today more explosive than all of the atomic bombs Russia could ever invent." Nor was the suggestion that persistent oppression rendered the situation "explosive" mere speculation: there is evidence that the hateful rhetoric expounded by Farrakhan and the Nation of Islam sometimes did trigger anti-white violence by its members.[28]

Another example relates to ISIL's terroristic hate propaganda. Venkatesh et al. (2020) have argued that ISIL's success at using hate propaganda to motivate violence is partly explained by the fact that it taps into its audience's feelings of deep injustice. ISIL hate propaganda narratives, they note, "almost exclusively begin with a description of how ISI[L] has been wronged, and therein begins a justification for the violent acts that are meted out in graphic detail via the videos" (2020: 1760). This phenomenon is also reflected in the in-depth interviews Baugu and Neumann (2020) have recently conducted with radicalised Islamist prisoners. Their interview reveals that victimhood stories, which portray injustices allegedly perpetrated by the West, play a key role in motivating violent action (2020: 1580-1).

My point is not that hateful counterspeech is necessarily more likely to provoke violence. It is rather that the relationship between broader social context and speech-induced violence is empirically ambivalent. In some respects, the social context could make hate speech by dominant groups more likely than hateful counterspeech to successfully motivate violence. But in other respects, the opposite could be true.

What this last point suggests is that, even when we factor in the broader social context and its influence on local speech contexts, hateful counterspeech still seems capable of generating *some* serious harms associated with hate speech. Indeed, I have just argued that at least some harms (most notably, that of inciting or otherwise provoking violence) seem in important respects less dependent on the broader social context; and that, even insofar as they are dependent on the broader context, hateful counterspeech still seems capable of producing them. With respect to these harms, there is therefore a meaningful symmetry between low-status hate speech and hateful counterspeech.

Yet, this should not obscure the fact that, with respect to other harms, there remains a moral *a*symmetry between these two categories of speech. As we saw earlier in this section, the broad societal context empowers low-status hate speakers to produce some harms—e.g.,

[27] Relatedly, Howard (2019b: 239–40) argues that speech that imposes a small risk of serious harm can be deeply wrongful (and liable to regulation).

[28] See note 15.

some especially stringent forms of psychological harm, as well as, according to McGowan, some illocutionary harms such as subordination, silencing, oppression, and discrimination—in a way that does not extend to hateful utterances directed by members of vulnerable groups at their oppressors.

Thus, the foregoing examination of the social dispersion of speaker power suggests that low-status hate speech can generate serious harms that hateful counterspeech cannot, but that hateful counterspeech nevertheless remains capable of generating some serious harms. In what follows, after briefly reiterating my argument for this view, I will turn, finally, to the implications of this moral conclusion for policy.

7 Conclusion

Is hateful counterspeech—inferiority-promoting speech directed by vulnerable groups at dominant groups—capable of producing serious harms? I have argued that it is. Yet this, as we have seen, is consistent with maintaining an important moral asymmetry between hateful counterspeech, on the one hand, and the inferiority-promoting speech of dominant groups, on the other.

Recall, briefly, why this is so. The claim that hateful counterspeech is not seriously harmful faces a problem: this claim appears to be in tension with arguments for thinking that hate speech by low-status members of dominant groups ("low-status hate speech") is harmful (Sects. 2–3). Indeed, many of the explanations for the harmfulness of low-status hate speech also seem to extend to hateful counterspeech. For example, attending to the jurisdiction and to the dynamics of authority helps explain why low-status hate speakers may have authority (Sects. 4–5). But it *also* suggests that hateful counterspeakers such as Farrakhan can be authoritative (and so, capable of harming with his words). The insight that a speaker's power to generate harms depends on the broader social context (Sect. 6) is more successful at establishing a moral asymmetry between the two: it explains why low-status hate speakers can generate certain serious harms *and* why hateful counterspeakers cannot. Still, it faces a limitation of scope: it only applies to certain categories of harms. For other serious harms—most notably, the harm of motivating violence against target groups—it remains plausible to think that hateful counterspeech can produce them too.

This moral conclusion has important implications for policy. The first implication is that hateful counterspeech may justifiably be legally regulated as a form of hate speech. As we have seen, hate speech can generate numerous serious harms. Even if hateful counterspeech cannot produce all of them (for the reasons outlined in Sect. 6), the important point remains that it can produce some of these harms (notably, the harm of inciting or otherwise motivating violence), and that the harms that it can produce are sufficiently weighty to warrant legal regulation. Accordingly, existing hate speech laws may be warranted in adopting the "expansive" account of hate speech, which encompasses speech directed at vulnerable groups *and* speech directed at more dominant groups.

This implication matters. As discussed in the Introduction, some legal scholars have influentially criticised the expansive legal definition of hate speech. Matsuda (1989: 2361), for instance, explicitly recommends that "expressions of hatred […] against historically-dominant group members by subordinated group members are not criminalized."[29] My

[29] See also Gelber (2018: 407-08), Richardson-Self (2018: 14).

analysis provides reasons to resist this narrower understanding of hate speech law, and to preserve the expansive account habitually on display in existing legal codes.

Yet my argument has also shown that this first implication requires qualification. The fact that "standard" hate speech can generate many harms that hateful counterspeech cannot also has implications—albeit more distant implications—for law. Specifically, this fact suggests that very fine-grained hate speech laws, which disentangle the different possible harms of hate speech, could sometimes be warranted in focusing narrowly on vulnerable groups. To see why, imagine laws that single out specific speech-based harms—e.g., a prohibition on speech that inflicts *serious psychological harm on groups*; a prohibition on speech that *systemically discriminates against groups*; a prohibition on speech that *facilitates intergroup violence*; etc. The argument offered in Sect. 6 suggests that the first two, but not the third, may be warranted in focusing exclusively on speech directed at vulnerable groups.[30] This, once more, is because, for reasons relating to the social dispersion of speaker power, hateful counterspeech arguably cannot produce these first two harms in the way standard hate speech can.

Having said that, this second implication is less immediately applicable than the first. The reason for this is that existing legal codes relating to hate speech are usually not fine-grained in the way outlined above. That is, existing hate speech laws are very often not focused on a specific harm associated with hate speech. For example, some restrictions target speech whose contents represent other groups as fundamentally inferior—e.g., as essentially vile, subhuman, contemptible, and the like.[31] That inferiority-promoting content, in turn, could contribute to generating multiple kinds of serious harms. For instance, it could subordinate, enact discriminatory norms, cause psychological harms, or even motivate acts of violence against the target groups. A similar observation applies to another common type of hate speech law—laws that prohibit incitement to hatred. Inciting hatred against a group can harm in numerous ways. Among many other things, it can undermine targets' assurance of their civic dignity, as well as facilitate acts of violence against the target group.[32]

The fact that, currently, hate speech laws tend to pick out speech that can harm in multiple ways is important when it comes to defining the "who" of hate speech. Some of these harms (such as constituting subordination, or enacting systemically discriminatory norms) arguably cannot be produced by speech directed by vulnerable groups at dominant groups. But others (most notably, motivating acts of violence) can. And since, to reiterate, these latter harms are sufficiently weighty to warrant legal regulation, legal codes are still warranted in adopting the expansive account of hate speech, which includes speech directed at vulnerable groups and speech directed at dominant groups.

This leaves us with a nuanced conclusion. In cases where hate speech laws pick out speech that can generate multiple serious harms (at least one of which is the harm of motivating or inciting violence) we should prefer the expansive account. This, moreover, is presently the typical case. However, if a state were to adopt a more fine-grained hate speech law, which focuses exclusively on a harm that cannot be produced by hateful counterspeech

[30] In this vein, Gelber's (2018) prescriptive definition of hate speech focuses predominantly on the harm of systemic discrimination. If a hate speech law were to adopt this focus, it would have good reason—as Gelber suggests—to focus on vulnerable groups.

[31] See, e.g., United Nation's (1965) call for punishing the dissemination of "theories of the superiority of one racial group over another".

[32] Brown (2017a: 27–29) discusses numerous social harms associated with incitement to hatred.

(such as systemic discrimination),[33] then it would have good reason to prefer the narrow emphasis on speech that targets vulnerable subgroups.[34]

This conclusion helps steer a middle ground in the polarised debate with which we started. On the one hand, hateful counterspeech *can* generate harms weighty enough to warrant legal intervention—and this may well vindicate the current expansive legal definitions of the "who" of hate speech. But the important truth captured by legal theorists who oppose this expansive definition is that, nonetheless, there remains a moral asymmetry between the harms inflicted by hateful counterspeech, and those inflicted by standard cases of hate speech. And this asymmetry matters, not just because it affects our moral assessment of hateful counterspeech, but also because it suggests that, were we to adopt more "fine-grained" hate speech laws, the narrow definition of the "who," which excludes hateful counterspeech, could sometimes be appropriate.

Acknowledgements For helpful discussions, and/or comments on previous drafts, I am deeply grateful to Aaron Ancell, Sam Berstler, Laura Caponetto, Bianca Cepollaro, Anjalee de Silva, Joanna Demaree-Cotton, Corrado Fumagalli, Nat Hansen, James Hutton, Enes Kulenovic, Robert Simpson, and Suzanne Whitten. I am also grateful to two anonymous reviewers for *Ethical Theory and Moral Practice*, whose generous comments have vastly improved this paper.

Declarations

Competing Interests The author has no competing interests or conflicts of interests to declare.

References

Badar M (2016) The Road to Genocide: The Propaganda Machine of the Self-Declared Islamic State. Int Criminal Law Rev 16:361–411

Barnes MR (2016) Speaking with (Subordinating) Authority. Soc Theory Pract 42:240–257

Baugu P, Neumann K (2020) Online propaganda use during Islamist radicalization. Inform Communication Soc 23:1570–1592

Bonotti M, Seglow J (2019) Self-Respect, Domination, and Religiously Offensive Speech'. Ethical Theory and Moral Practice 22:589–605

Brown A (2016) The "Who?" Question in the Hate Speech Debate: Part 1. Canadian Journal of Law & Jurisprudence 29: 275–320

[33] This, as explained in note 30, is the proposal Gelber (2018) has advanced.

[34] Note that these good reasons could be overridden. Partly for reasons of space, I have focused on the "who" question by examining which subgroups (vulnerable, dominant) are seriously harmed by hateful utterances. But Brown (2016; 2017a) has argued that other factors—including more "pragmatic" factors—should also count. For example, if restricting hate speech law so that it only prohibits speech directed at vulnerable subgroups were likely to lead to a violent backlash against those subgroups, or if it were exceedingly difficult to implement such restricted laws, it may be better to retain the expansive legal definition of hate speech (2017a: 41–42).

Brown A (2017) The "Who" Question in the Hate Speech Debate: Part 2. Canadian Journal of Law and Jurisprudence 30:23–55
Brown A (2017) What Is Hate Speech? Part 2. Law and Philosophy 36:561–613
Brown A (2019) Brison S, Gelber K (eds) (2019) Free Speech in the Digital Age. Oxford: Oxford University Press, pp. 207 – 21
Camp E (2013) Slurring Perspectives. Analytic Philos 54:330–349
Canadian Criminal Code (1985) Public Incitement of Hatred, § 319
Delgado R (1982) Words That Wound. Harv Civil Rights-Civil Liberties Law Rev 17:133–181
Farrakhan L (1989) The Origin of the White Race: The Making of the Devil. Speech Delivered in Chicago, IL
Fricker M (2007) Epistemic Injustice. Oxford University Press, Oxford
Gardell M (1996) In the Name of Elijah Muhammad. Duke University Press, Durham
Gelber K (2018) Differentiating Hate Speech: A Systemic Discrimination Approach. Crit Rev Int Social Political Philos 24:393–414
Gelber K (2019) Terrorist-Extremist Speech and Hate Speech. Ethical Theory and Moral Practice 22: 607–22
Howard J (2019) Free Speech and Hate Speech. Annu Rev Polit Sci 22:93–109
Howard J (2019) Dangerous Speech. Philosophy & Public Affairs 47:208 – 54
Langton R (2009) Sexual Solipsism. Oxford University Press, Oxford
Langton R (2018) Garden J, Green L, Leiter B (eds) (2018) Oxford Studies in Philosophy of Law, Vol.3. Oxford: Oxford University Press, pp.30–53
Lepoutre M (2021) Democratic Speech in Divided Times. Oxford University Press, Oxford
Lee M (2011) The Nation of Islam and Violence. In: Lewis J (ed) Violence and New Religious Movements. Oxford University Press, Oxford, pp 295–305
MacFarquhar N(2007) Nation of Islam at a Crossroad as Leader Exits. *New York Times.* 26 February 2007. https://www.nytimes.com/2007/02/26/us/26farrakhan.html?pagewanted=all
Maitra I (2012) Subordinating Speech. In: Maitra I, McGowan MK (eds) Speech and Harm. Oxford University Press, Oxford, pp 94–120
Maitra I, McGowan MK (2012) Introduction and Overview. In: Maitra I, McGowan MK (eds) Speech and Harm. Oxford University Press, Oxford, pp 1–23
Malcolm X(1964) The Ballot or the Bullet. Detroit. https://www.youtube.com/watch?v=D9BVEnEsn6Y
Matsuda M (1989) Public Response to Racist Speech. Michican Law Review 87:2320–2381
McGowan MK (2019) Just Words. Oxford University Press, Oxford
Popa-Wyatt M, Wyatt J (2018) Slurs, roles and power. Philos Stud 175:2879–2906
Richardson-Self L (2018) Offending White Men. Feminist Philos Q 4:1–24
Sanders P, Cohen B (2011) The Zebra Murders. Arcade Publishing, New York
Simpson R (2013) Dignity, Harm, and Hate Speech. Law Philos 32:701–728
Simpson R (2013) Un-Ringing the Bell. Australasian Journal of Philosophy 91:555–75
Southern Poverty Law Center (2019) Rage Against Change. https://www.splcenter.org/sites/default/files/intelligence_report_166.pdf
Tirrell L (2012) Genocidal Language Games. In: Maitra I, McGowan MK (eds) Speech and Harm. Oxford University Press, Oxford, pp 174–221
United Nations (1965) International Convention on the Elimination of All Forms of Racial Discrimination
Venkatesh V, Podoshen J, Wallin J, Rabah J, Glass D (2020) Promoting Extreme Violence.Terrorism and Political Violence32:1753–75
Waldron J (2012) The Harm in Hate Speech. Harvard University Press, Cambridge
Wilson RA, Kiper J (2020) Incitement in an Era of Populism. Univ Pa J Law Public Affairs 5:56–121
Wise T(2018) No, Farrakhan Is Not the Problem. Medium (blog). https://medium.com/s/story/no-farrakhan-is-not-the-problem-d2d1a37e1162