Journal of Ethics and Social Philosophy
VOL. 16, NO. 3 · DECEMBER 2019
https://doi.org/10.26556/jesp.v16i3.682

CAN "MORE SPEECH" COUNTER IGNORANT SPEECH?

Maxime Lepoutre

DEMOCRATIC PUBLIC DISCOURSE is rife with *ignorant speech*—speech that disseminates or promotes falsehoods. Some of this speech comes from speakers who are themselves ignorant and believe the falsehoods they utter. Very often, however, the sources of ignorant speech, as defined above, are not themselves ignorant; rather, they knowingly spread ignorant views and perspectives for personal or political gain.

Ignorant speech is a very broad class. It can spread many different kinds of falsehoods that bear on many different kinds of things. But two species of ignorant speech are particularly noteworthy, as they have proven to be particularly dangerous in contemporary democracies. The first consists in ordinary *political misinformation*, where speakers circulate falsehoods concerning policy-related issues. Whether it is bloggers claiming that the MMR vaccine causes autism, senior politicians asserting that Obamacare would implement life-threatening "death panels," or public figures insisting that exiting the European Union would save the United Kingdom (and its National Health Service) £350 million a week, there is no doubt that policy debates are rife with such misinformation.[1]

The second noteworthy category spreads falsehoods not about policies but about people. Specifically, this species of ignorant speech, which constitutes a type of *hate speech*, puts forward representations of other social groups that falsely deny their basic equality—often by advancing vilifying or degrading characterizations of them. A paradigmatic illustration here is speech that activates distorted stereotypes that present minorities as essentially subhuman or morally inferior: for instance, that refugees are "a plague" of "cockroaches," that Muslims are terrorists, or that Mexican immigrants are "criminals" and "rapists."[2]

The pervasiveness of these two forms of ignorant speech constitutes a serious

1 Godlee, Smith, and Marcovitch, "Wakefield's Article Linking MMR Vaccine and Autism Was Fraudulent"; Holan, "PolitiFact's Lie of the Year"; Sparrow, "UK Statistics Chief Says Vote Leave £350m Figure Is Misleading."

2 Nelson, "Katie Hopkins Reflects on Branding Migrants 'Cockroaches' and 'Feral Humans'";

political problem: policy misinformation and hateful forms of ignorant speech, it has been argued, are both liable to generate grave injustices. The former risks misleading the public into supporting unjust policies or harmful candidates.[3] As for the latter, it is capable of producing a wide array of unjust harms, such as inciting animosity, causing psychological trauma, undermining targets' sense of dignity, subordinating targets, or enacting discriminatory norms.[4]

A celebrated response to this problem recommends countering ignorant utterances with informed and truthful speech. This is of course familiar from debates in democratic theory: "deliberative" theorists of democracy have long advocated inclusive deliberation, and have increasingly done so on epistemic grounds. On their view, exchanging arguments, voicing personal narratives, and soliciting expert testimony are powerful ways of debunking political misinformation and thus correcting false political beliefs.[5]

This response is not specific to ordinary political misinformation: "counterspeech" is also an extremely popular remedy for hateful forms of ignorant speech. In particular, opponents of legal restrictions on hate speech typically insist that we should counter hate speech verbally rather than legally.[6] In this context, counterspeech involves contesting the distorted contents that hate speech asserts, presupposes, or encodes, so as to block whatever harms their diffusion would otherwise give rise to.

Despite this popularity, the counterspeech response to ignorant speech has come under heavy criticism, especially in the context of hate speech. According to one influential objection, advocating counterspeech as the remedy for hate speech is *unfair*, as it "places the burden of the remedy on those targeted (and potentially harmed) by the allegedly harmful speech."[7] Relatedly, others have worried that, because it demands too much of targets of hate speech, counterspeech also risks being *ineffective.* For one thing, there is evidence that targets

---

Asser, "What the Muhammad Cartoons Portray"; Reilly, "Here Are All the Times Donald Trump Insulted Mexico."

3 Brown, "Propaganda, Misinformation, and the Epistemic Value of Democracy."

4 Delgado, "Words that Wound"; Waldron, *The Harm in Hate Speech*; Maitra, "Subordinating Speech"; McGowan, *Just Words*, ch. 7.

5 See, e.g., Young, *Inclusion and Democracy*, chs. 1–3; Anderson, "The Epistemology of Democracy"; and Landemore, *Democratic Reason*.

6 See, e.g., Louis Brandeis's opinion in *Whitney v. California* (274 U.S. 357 (1927)); Post, *Constitutional Domains*; Brettschneider, *When the State Speaks, What Should It Say?* ch. 3; Lepoutre, "Hate Speech in Public Discourse"; and Strossen, *Hate*, ch. 8.

7 McGowan, "Responding to Harmful Speech," 183. See also Maitra and McGowan, "Introduction and Overview," 9.

rarely respond to such speech.[8] And even if targets did speak back, they may lack the authority needed to do so in an effective or compelling way. This is both because targets generally belong to relatively vulnerable social groups, and because, as feminist philosophers of language have argued, hate speech can itself disable (or "silence") its targets' speech.[9]

However, these two preliminary concerns are not decisive. According to an increasing range of defenders of counterspeech, counterspeech can be thought of as a state-driven practice, rather than as something merely to be performed by private individuals. On this understanding, which Corey Brettschneider and Katharine Gelber have prominently articulated, it is primarily, though not exclusively, the state's responsibility to speak out against hate speech. The state has numerous tools for doing this, such as having high-profile politicians denounce vilifying utterances, designing school syllabuses, erecting historical monuments, or funding civil rights groups to challenge hate speakers.[10]

This adjusted conception of who should engage in counterspeech seems a promising way of alleviating the two concerns outlined above. Recommending state-sponsored counterspeech is less *unfair*, as it shifts the burden of responding away from targets of hate speech. Moreover, the state is more likely than vulnerable targets to have the *authority* needed to be taken seriously when it opposes hate speech.

But even with this adjustment, the policy of countering ignorant speech with more speech still encounters a deep problem. In her influential discussion of oppressive speech, Mary Kate McGowan has argued that the counterspeech response operates with a naive conception of how language works. In particular, it overlooks the asymmetric pliability of conversational norms—the phenomenon whereby it is easier to enact conversational norms than it is subsequently to undo those norms. Because of this phenomenon, the damaging effects of ignorant speech on public discourse are difficult to reverse, or "sticky." Accordingly, McGowan concludes, trying to undo the harmful effects of ignorant speech through more speech seems as futile as trying to "unring a bell."[11]

Crucially, this problem of stickiness stems from facts about the dynamics of

8 Nielsen, "Power in Public."

9 On the authority problem and silencing, see Maitra and McGowan, "Introduction and Overview," 10; McGowan, "Responding to Harmful Speech," 183–84; Langton, "Speech Acts and Unspeakable Acts." For a positive account of the authority requirement needed to successfully challenge hate speech, see Barnes, "Speaking with (Subordinating) Authority," 251–56.

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

11 McGowan, "Oppressive Speech," 403.

language and conversational norms, rather than about who is doing the talking. *Prima facie*, then, it is just as applicable to state counterspeech as to non-state counterspeech. In fact, as we will see, there are reasons to think that it is *especially* problematic for state-driven counterspeech. What this means is that an increasingly popular way of defending counterspeech against criticism—namely, focusing on state-driven counterspeech—will not help with this problem.

In this article, I will examine the stickiness of ignorant speech more closely, with two aims. The first is to show that the asymmetric pliability of conversational norms is an even more general problem than McGowan suggests, and that, as a result, it stands in the way of verbally countering both hateful and non-hateful forms of ignorant speech. The second is to articulate a more sophisticated account of counterspeech, which is distinctively suited to overcoming this problem. Doing so will help us arrive at a normative ideal of public discourse that is more finely attuned to nonideal conditions.

My argument will proceed as follows. After introducing the problem of stickiness, I demonstrate that it is a very general phenomenon, which applies not just to the oppressive or hateful speech that McGowan analyzes, but also to ordinary political misinformation (section 1). I then develop a conception of counterspeech that is directly informed by this difficulty. Specifically, I argue that, by distinguishing between positive and negative forms of counterspeech (section 2) and adopting a more nuanced view of counterspeech's temporality (section 3), we can substantially mitigate the problems generated by sticky ignorance. I conclude by examining the implications of this argument for legal responses to ignorant speech (section 4).

Before proceeding, two important clarifications are in order. First, note that my purpose here is *not* to conclude that counterspeech is a sufficient remedy for ignorant speech, so that legal responses to such speech—such as bans on hate speech or on fake news—are unwarranted. I will be exploring and challenging a specific argument against counterspeech. But this does not necessarily entail opposing legal remedies. To begin, there might be other, as yet uncanvassed, problems for counterspeech. Moreover, my argument should also be useful to those who defend legal remedies. Many proponents of legal remedies accept that counterspeech nonetheless also has a role to play in combatting hateful or deceptive speech.[12] To determine what the division of labor should be between legal and speech-based remedies, such theorists need to determine more precisely what problems are faced by counterspeech but avoided by bans. As will become apparent in section 4, my argument will advance this understanding: in

12 E.g., Gelber, "Reconceptualizing Counterspeech in Hate Speech Policy (with a Focus on Australia)," 214; Langton, "Blocking as Counterspeech."

suggesting that a revised conception of counterspeech can substantially weaken the stickiness of ignorant speech, I am debunking one hypothesis regarding what legal remedies can do that counterspeech cannot.

Second, we should pause to examine the appropriateness of considering hate speech and political misinformation together under the heading of ignorant speech. One might be wary of doing so on various grounds. For one thing, one might point out that not all hate speech involves disseminating or promoting falsehoods. In particular, some instances of hate speech attack vulnerable groups via utterances that are true (the leader of a racist party who shouts "We want you out!"), while other instances of hate speech, like imperatives, are not truth-apt ("Get out of this country!"). Thus, strictly speaking, these types of hate speech do not count as ignorant speech, as I define it.

Because my focus is on how we might verbally counter speech that produces injustice or harms in virtue of spreading or promoting falsehoods, I will be bracketing these further types of hate speech for the purposes of this paper. That is, by "hate speech," I will largely be talking about speech that puts forward representations of its targets that incorrectly deny their basic equality (such as racist or xenophobic stereotypes that present their targets as essentially degraded or vile). While this constitutes a limit to the scope of my counterspeech-related recommendations, it does not render that scope insignificant: after all, an important portion of hate speech *does* produce bad effects (e.g., psychological distress, damage to targets' dignity, stirring up animosity) by broadcasting distorted equality-denying stereotypes.[13]

Still, even with this qualification, one might think that hateful and non-hateful forms of ignorant speech remain importantly different: as discussed above, their dissemination of falsehoods may well generate harms or injustices in different ways. Political misinformation typically does so by instilling false beliefs in listeners, which in turn may induce them to embrace unjust policies or dangerous candidates. Although speech that describes vulnerable groups in degrading ways may also induce false beliefs, this is far from the only way in which it can produce injustice: to reiterate, publicizing such hateful representations can also, among other things, assault its targets' dignity, subordinate them, or enact discriminatory norms.

13 What is the payoff of this restriction? As I will explain below, part of my aim in considering political misinformation alongside hate speech is to bring the rich philosophical analyses of conversational stickiness developed in the context of hateful speech to bear on our understanding of political misinformation. With this aim in mind, it makes sense to restrict my attention to cases of hate speech that are closer in form to political misinformation, in that—*qua* ignorant speech—they involve spreading or promoting falsehoods.

I do not wish to minimize this difference. My point is rather that, nevertheless, hateful and non-hateful forms of ignorant speech remain similar in at least one important way: as I will explain in section 1, their damaging conversational effects are "sticky," in a way that makes them difficult to rebut via counterspeech. And, when exploring this similarity in section 1.2, I will consider how it may be complicated by the fact that hateful and non-hateful forms of ignorant speech produce harms or injustices via different mechanisms.

This similarity is important, and reveals why it is fruitful to consider hateful and non-hateful forms of ignorant speech side by side. On the one hand, the rich theoretical insights that philosophers of language such as McGowan have developed when thinking about the stickiness of hateful speech can help us understand why and when policy-related misinformation may also be resistant to counterspeech. Because they have traditionally been insensitive to this stickiness, democratic theories that advocate policy-related deliberation would particularly benefit from these insights. But the fruitful relationship also goes in the other direction. Social psychologists have conducted numerous experiments on political misinformation and its correction. Insofar as hateful and non-hateful ignorant speech involve a similar kind of stickiness, this empirical evidence can, I will suggest, inform our verbal responses to hateful as well as non-hateful ignorant speech.

1. THE STICKINESS OF IGNORANT SPEECH

1.1. The Asymmetric Pliability of Conversational Norms

Utterances can affect and alter conversations in countless ways. Among many other things, they can share new information (or bring it into the "common ground"); they can call accepted views into question; they can introduce new hypotheses or ideas, thereby making them more salient or relevant to the conversation; and they can explicitly invoke new conversational standards or rules.

Whenever utterances do one of these things, they alter the norms that govern the conversation. This is clearest in the case of speech that explicitly invokes a new rule for the conversation with the intention of enacting it. Suppose a classroom participant says, "From now on, people must raise their hand before speaking." Provided the speaker has sufficient authority (she is the instructor, rather than an overzealous auditor), her utterance can enact a new conversational norm—namely, the norm that people must raise their hand before speaking.[14]

However, this is only a special case of how utterances alter conversational

14 McGowan, "Oppressive Speech," 393–94.

norms. As McGowan has shown, even contributions that neither explicitly invoke nor intend to enact norms nonetheless alter the norms that govern the conversation. When an utterance shares new information, for instance, this makes it more permissible for subsequent utterances to presuppose that information. For example, it is more conversationally appropriate for *A* and *B* to discuss where the ceremony is taking place *after* they have told their interlocutors that they are getting married. Similarly, making a new hypothesis salient partly changes the topic of the conversation: that hypothesis becomes relevant to the conversation, so that participants can appeal to it without seeming out of line. If police officers investigating the disappearance of Sheriff Smith's hat are listing known thieves, mentioning Deputy Jones (who is not a known thief) would seem inappropriate. But doing so becomes more conversationally appropriate if a participant raises the possibility that the theft was an inside job. McGowan's general point, then, is that "our utterances routinely change what is permissible around us."[15]

But even if utterances routinely change conversational norms, not all norm changes are equally easy to bring about. In particular, some norms are easier to introduce than to reverse. More specifically, alongside other philosophers of language, McGowan suggests that the norms enacted by oppressive speech (such as hate speech) are "especially difficult to undo."[16] To illustrate this asymmetry, imagine that a public figure asserts "*X*s are lazy parasites." Verbally countering this assertion—for instance, by saying "That's false! *X*s are *not* lazy parasites!"—may well be ineffective at reversing the initial utterance's nefarious conversational effects. Intuitively, negating the vilifying proposition in this way would only strengthen its conversational relevance: whether or not *X*s are lazy parasites would become even more clearly the topic of conversation and, consequently, the social standing of *X*s would be even more evidently at issue. On the basis of cases like these, McGowan hypothesizes that the norms introduced by oppressive speech are asymmetrically pliable—verbally introducing them is easier than verbally reversing them.

Elaborating on McGowan's hypothesis, Robert Simpson has argued that the asymmetric pliability of the norms enacted by oppressive or hateful speech results from a more general asymmetry, which concerns *salience*. Generally speaking, while we can easily make an association salient by mentioning it and draw-

15 McGowan, "Oppressive Speech," 406; see also 394–97.

16 McGowan, "Responding to Harmful Speech," 189; see also "Oppressive Speech," 403. McGowan is building on Lewis's ("Scorekeeping in a Language Game") observation that raising the standards for using vague terms like "flat" is easier than lowering them. Stanley (*How Propaganda Works*, 157–61) and Langton ("Blocking as Counterspeech," sec. 6) echo the point that oppressive conversational norm changes are harder to reverse.

ing people's attention to it, it is much more difficult subsequently to make it less salient.[17] This is because rejecting an association may inadvertently make one's interlocutors pay even more attention to it.

This salience-based asymmetry helps explain why the distorted representations advanced by oppressive hate speech are so resistant to counterspeech. The initial utterance makes a vilifying association salient to listeners (say, between being an *X* and being a lazy parasite). This increased salience adjusts the bounds of the conversation, so that the hateful association becomes relevant to it. In other words, by making this association salient, the hateful utterance introduces the following conversational norm: it makes it more permissible for people subsequently to debate whether or not *X*s are in fact lazy parasites. In this context, trying to repudiate the association risks amplifying its salience and thereby reinforcing the conversational norm initially introduced.[18] That is, exclaiming "*X*s are *not* lazy parasites!" may challenge the claim *that* X*s are parasites.* But even as it does so, it risks magnifying the salience of this stereotype within the conversation. And, in turn, saliently denying the vilifying stereotype reinforces its centrality as a topic of debate: it implicitly reaffirms the harmful norm that it is permissible publicly to discuss whether or not *X*s are lazy parasites.[19]

In sum, the distinctive properties of conversational salience make the harmful effects of hateful speech sticky, in that they are more difficult to reverse than to enact. The natural response to this problem is pessimism about counterspeech. McGowan, as mentioned earlier, famously concludes that counterspeech is as futile as "trying to unring a bell."[20] Just as trying to unring a bell is likely to make it ring even more loudly, so trying to reverse the effects of hateful speech may simply amplify those effects. Simpson likewise expresses pessimism: for activated associations "to become *un*-salient," he observes, "we will simply have to change the subject … or allow time to elapse."[21]

Importantly, this problem is especially worrisome given how the policy of

17 Simpson, "Un-Ringing the Bell," 572.

18 Simpson, "Un-Ringing the Bell," 569–70. While McGowan and Simpson focus on oppressive speech generally, this phenomenon also arises in more specific debates. When discussing hate speech that takes a coded form, Stanley (*How Propaganda Works*, 159) worries that verbally countering such speech requires exposing the code, and thereby raising the salience of its hateful contents. Moreover, examinations of slurs have shown that slur terms preserve their bad effects even when they are negated: "*A* is not an *X*" (e.g., Langton et al., "Language and Race," 756–57). Here, too, salience may have an explanatory role: negating a slur involves repeating it, which increases the salience of the degrading stereotype it expresses.

19 In section 1.2, I explain in greater detail why this conversational norm is harmful.

20 McGowan, "Oppressive Speech," 403.

21 Simpson, "Un-Ringing the Bell," 569.

countering hateful speech with more speech has recently been defended. As discussed above, in response to concerns regarding the fairness and authority of counterspeech, its proponents have increasingly argued that counterspeech should primarily be performed by the state. However, we can now see that embracing state-driven counterspeech does not alleviate the stickiness problem. If anything, it worsens it. In virtue of being highly authoritative, state speech is capable of significantly influencing the norms of public discourse on a large scale. Precisely because state speech is so powerful, what the state says is typically highly salient. What this means, however, is that having the state repudiate vilifying representations risks increasing their salience all the more. Thus, the objection under consideration is so problematic for state-driven counterspeech because it works by harnessing the power of counterspeech against it: since the state has an unparalleled ability to set the agenda of public discourse, it also has an unparalleled tendency to strengthen the salience of vilifying representations when challenging them.

1.2. Generalizing the Objection

Philosophers have almost exclusively discussed the problem of stickiness within debates about oppressive speech or hate speech—i.e., speech that casts its targets as fundamentally inferior.[22] But the linguistic phenomenon that underpins this problem is quite general. To see this, recall that, according to Simpson's analysis, the difficulty of reversing the conversational norms enacted by oppressive speech is partly explained by the fact that it is easier to make an association salient than unsalient. Now, this asymmetry is not inherently sensitive to the specific content of the association.[23] It stems from properties of salience, and any content can in principle be salient. Accordingly, the salience-related asymmetry Simpson identifies applies not simply to hateful stereotypes, but also to associations that spread misinformation about policy.

As a result, it is also true of counterspeech directed at policy misinformation that it risks amplifying rather than reversing the salience of its target utterance. This observation is fairly intuitive. The more the Centers for Disease Control (CDC) asserts that "vaccines do not cause autism," the more conversationally salient the misleading association between vaccines and autism risks becoming.[24]

22 Simpson does briefly note when concluding that the asymmetric pliability of conversational norms may apply beyond the context of oppressive speech ("Un-Ringing the Bell," 572–74). But his purpose in doing so is not to establish that ordinary political misinformation, specifically, can be sticky; and his focus in the vast majority of the paper is on oppressive speech.

23 Simpson, "Un-Ringing the Bell," 573.

24 Centers for Disease Control and Prevention, "Vaccines Do Not Cause Autism."

Similarly, insisting, as the anti-Brexit "Remain" campaign did, that the UK did not send the EU £350 million each week, risked drawing attention to that rumor.[25] The point, then, is that insofar as the problem of stickiness results from a general salience-based asymmetry, we have every reason to think that this phenomenon applies to ordinary political misinformation as well.

But more needs to be said. So far, I have suggested that, as with hateful forms of ignorant speech, it is easier to make the associations advanced by ordinary political misinformation salient than unsalient. The second step is to suggest that the salience of political misinformation is connected to its morally problematic effects. Indeed, unless this second step holds, a skeptic might raise the following concern: "Even if it is hard to make political misinformation unsalient through counterspeech, this is not morally problematic. After all, the potential harmful or unjust effects of political misinformation are not strongly connected to, and do not readily ensue from, the mere salience of misinformation. So, even if counterspeech exacerbates the salience of misinformation, it might nevertheless succeed in discrediting it and disabling its harmful conversational effects." To establish that the salience-based asymmetry Simpson diagnoses truly is a problem for countering political misinformation with more speech, we must therefore identify a meaningful connection between the salience of misinformation and its harmful or unjust conversational effects.

Now, because hateful and non-hateful forms of ignorant speech lead to harms or injustices in different ways (and very likely to a different extent), the specific answer we give to this second question is likely to differ from the answer we give in the case of hateful forms of ignorant speech. In the case of hate speech, Jeremy Waldron has powerfully argued that there exists a tight, constitutive connection between conversational salience and injustice. According to Waldron, justice requires that citizens be assured of their standing as social and political equals.[26] Indeed, unless it is visible to citizens that they can count on having their rights as equal members of society respected, they will be unable to participate in social and political life without fear.[27] The problem with hate speech, Waldron emphasizes, is that the mere salience of its vilifying contents in public debate threatens to undermine this very assurance.

To see why, remember that what it means for something to be conversationally salient is that it is relevant to the conversation, or included within the topic of conversation. It is, in other words, partly what the conversation is about. Consequently, when a vilifying stereotype is salient in public debate, this means

25 Sparrow, "UK Statistics Chief Says Vote Leave £350m Figure Is Misleading."

26 Waldron, *The Harm in Hate Speech*, ch. 4.

27 Waldron, *The Harm in Hate Speech*, 85.

that the public conversation is *about* whether its contents are accurate. What is at issue in the conversation, then, is whether or not, say, women are submissive, blacks are subhuman, or immigrants are parasites. Since these propositions are clearly inconsistent with the good status of the targeted groups, these groups are no longer assured of their good standing. Instead, their standing is precisely what is at issue in public debate. Therefore, it is clearly problematic if counterspeech reinforces the conversational salience of hate speech: the mere salience of hateful representations constitutes an injustice.

This, however, does not seem to be the case with ordinary political misinformation. Unlike with vilifying stereotypes, the fact that a falsehood about policy matters is at issue does not inherently constitute an injustice. Indeed, that there is a debate about whether MMR vaccines cause autism does not in and of itself call anyone's standing into question. So, if the salience of ordinary political misinformation is not harmful or unjust in the same way that the salience of hate speech is, what nevertheless makes this salience morally problematic?

In answering this question, philosophers can benefit from contemporary advances in cognitive science. Cognitive scientists have shown that there is an intimate connection between the cognitive fluency of an association—roughly, how familiar one is with that association—and one's disposition to believe it. As Lewandowsky et al. summarize, "in general, fluently processed information . . . is more likely to be accepted as true."[28]

This finding helps appreciate what is problematic about the conversational salience of political misinformation. When an association is made conversationally salient, that association is thereby activated for participants: because the association is newly relevant to the conversation, their attention is drawn to it, and they are prone to represent it. Thus, the salience of an association is tightly linked to participants' cognitive familiarity, or fluency, with it. If, as cognitive scientists argue, the fluency of an association in turn disposes people to believe it, then the problem becomes manifest. The conversational salience of ordinary political misinformation is problematic because it increases the likelihood that participants will accept the misinformation. And this, in turn, is plausibly causally related to harm and injustice. Insofar as political misinformation conceals the vices of dangerous policies and candidates, and obscures the virtues of good policies and candidates, the salience-induced acceptance of misinformation may help produce injustice.

The upshot for counterspeech directed at ordinary political misinformation

28 Lewandowsky et al., "Misinformation and Its Correction," 212. See also Alter and Oppenheimer, "Uniting the Tribes of Fluency to Form a Metacognitive Nation," 228; and Berinsky, "Rumors and Health Care Reform," 245–47.

is troubling: insofar as countering misinformation with more speech magnifies its salience, doing so risks bolstering people's disposition to accept the targeted misinformation, together with the attending dangerous consequences. This worry is not merely theoretical. Social psychologists have widely reported a "continued influence effect," whereby attempting verbally to correct falsehoods is ineffective or worse.[29] In one study, CDC flyers distinguishing myths from facts about vaccines are shown to backfire: people who have read the flyer end up being more likely to misidentify myths as facts—and to oppose vaccination—than people who have not.[30] Similarly, Adam Berinsky finds that, although rehearsing and correcting the Obamacare death-panel rumors initially makes people somewhat more likely to reject those rumors, this effect largely disappears after a few weeks.[31]

Thus, the problem of stickiness applies not merely to hateful forms of ignorant speech, but also to political misinformation. As we have seen, there are two steps to establishing this generalization: first, as with hateful forms of ignorant speech, it is easier to increase the conversational salience of political misinformation than it is subsequently to reverse this salience; second, in light of the foregoing theoretical and empirical considerations, the conversational salience of political misinformation is importantly connected to its damaging conversational effects. Accordingly, here also, counterspeech risks being as futile as trying to "unring a bell": instead of undoing the damaging conversational effects of political misinformation, it may amplify them.

This result constitutes a serious problem for deliberative theories of democracy. This family of theories, recall, insists that exchanging information and arguments helps to eliminate false beliefs and fallacious arguments.[32] As Hélène Landemore acknowledges, this position presupposes that people will recognize correct views and sound arguments when presented with them.[33] If, however, having deliberators challenge falsehoods amplifies their salience—which increases listeners' disposition to believe them—then this presupposition fails. Not only are people not necessarily swayed by sound counterarguments, but such counterarguments may instead entrench their false beliefs.

We have arrived at a deep and general problem for counterspeech. Because of the asymmetric pliability of conversational norms—which itself stems from a broader salience-related asymmetry—tackling hateful *and* non-hateful igno-

29 For a review, see Lewandowsky et al., "Misinformation and Its Correction," 113–21.

30 Lewandowsky et al., "Misinformation and Its Correction," 115.

31 Berinsky, "Rumors and Health Care Reform," 254–60.

32 See note 5 above.

33 Landemore, "Yes, We Can (Make It Up on Volume)," 220.

rant speech risks being ineffective or worse. We have already seen one possible response to this problem: Simpson suggests that, in the face of the problem of stickiness, we may have to change the subject, or wait for salient falsehoods to become unsalient. Unfortunately, neither option is wholly satisfactory: often, salient falsehoods bear precisely on the subject we are interested in, and the issue at hand may be so pressing that there is no time to wait. In what follows, I will offer an alternative response by demonstrating that the objection at hand relies on an insufficiently sophisticated account of counterspeech.

2. POSITIVE COUNTERSPEECH

2.1. Positive and Negative Counterspeech

In the first place, the objection under consideration overlooks the distinction between negative and positive counterspeech. One way to counter ignorant speech is to explicitly negate its ignorant content. "It is false," a politician might announce, "that Obamacare will implement death panels." Similarly, civil rights activists might respond to hate speech by insisting "*X*s are not lazy parasites!" Call this *negative counterspeech.* Because it involves repeating the ignorant proposition in the process of negating it, negative counterspeech risks strengthening its conversational salience. Accordingly, this kind of counterspeech is highly vulnerable to the stickiness of ignorant speech.

But there is another way of verbally countering ignorant speech. As we have seen, the harder-to-reverse quality of some conversational norms results partly from a salience-based asymmetry: it is easier to make associations salient than unsalient. So far, we have focused on how this general asymmetry applies to incorrect associations, hateful (section 1.1) and non-hateful (section 1.2). However, the asymmetry applies to correct associations as well: *ceteris paribus*, it is also easier to make correct associations salient than unsalient.

As a result, the asymmetric pliability of conversational norms does not suggest that all kinds of counterspeech are bound to be ineffective. Rather, it recommends a distinctive kind of counterspeech. To avoid reinforcing the salience of false associations, we should steer clear of counterspeech whose form is primarily negative. Instead, we should privilege *positive* forms of counterspeech, which aim to put forward and render salient correct associations of ideas. On this approach, countering ignorance is less about directly contesting a distorted vision of the world, and more about affirming a correct vision of the world that is inconsistent with the falsehoods at hand.[34] Because positive speech entails or

34 While philosophical and legal debates typically ignore this distinction, it resonates with

implicates, but does not explicitly assert, that the ignorant speech in question is false, it avoids increasing the ignorant proposition's salience—and, by extension, it avoids strengthening the conversational effects of ignorant speech. Moreover, because positive counterspeech makes correct associations salient instead, it harnesses the asymmetric pliability of conversational norms in the interests of truth: once salient, correct representations are comparatively difficult to render unsalient.[35]

---

some discussions in social psychology. Since the pioneering work of Tversky and Kahneman ("The Framing of Decisions and the Psychology of Choice"), social psychologists have investigated how framing messages in different ways—including negative versus positive ways—can alter their impact on listeners. Although Tversky and Kahneman themselves do not explore how framing might affect attempts at correcting previous misinformation, this more specific application has been taken up in the subsequent literature. Lewandowsky et al., for instance, recommend that corrections avoid repeating myths, and suggest that corrections "should be more successful when they can be encoded as an affirmation of an alternative attribute" ("Misinformation and Its Correction," 115–17). Though these comments are brief, and focus exclusively on ordinary misinformation, they seem congruent with my recommendation, inasmuch as they suggest countering falsehoods in an affirmative mode, rather than by negating falsehoods.

35 This last claim might seem too quick. As Leslie has argued, human beings are disposed to generalize strikingly negative information more quickly—and on the basis of less evidence—than neutral or positive information ("The Original Sin of Cognition," 395–99). Accordingly, one might think that negative associations are "stronger" than positive associations, so that positive associations are easier to make unsalient than negative associations. There are two things to say in response to this concern. First, Leslie's argument does not speak directly to the *salience* of positive and negative associations. Put differently, her evidence may indeed show that in *one* respect—the tendency to be generalized—negative associations are stronger than positive associations. But, from this, it need not follow that negative associations are stronger in the respect under consideration—namely, the difficulty of being made unsalient. After all, it could conceivably be the case that (1) *while* negative information is salient, we are more likely to generalize on its basis than on the basis of positive information, but that (2) negative information is no more difficult to render unsalient than positive information. Still, Leslie's argument relating to generalization might be thought to at least suggest the possibility that, by analogy, negative associations might be more difficult to render unsalient than positive associations. However—and this is the second point—even if this hypothesis is true, it remains consistent with my main point here: that, in virtue of how salience works, positive associations are more difficult to render unsalient than to initially render salient. Indeed, we can think that such a salience-based asymmetry applies to positive associations (which creates a hurdle that opposing speech must overcome) while allowing that this asymmetry is even stronger in the case of negative associations. This is congruent with the kind of evidence Leslie finds. She finds that we typically generalize *more quickly* on the basis of negative information than on the basis of positive information, *not* that we do not generalize on the basis of positive information. In fact, she even acknowledges that, although this is less common, the tendency to rapid generalization "arguably applies to strikingly positive information" ("The Original Sin of Cog-

Notice that resorting to positive counterspeech does not amount to simply changing the subject. The truths that positive counterspeech affirms bear on the same issue as negative counterspeech. This is guaranteed by the fact that the content of positive counterspeech is supposed to entail (or at least implicate) the falsehood of the targeted ignorant speech. So, positive counterspeech expresses a similar proposition as negative counterspeech, but under a different mode of presentation. Consider: "The star of *King's Row* was married twice" and "The fortieth US president was married twice" make the same claim about the world (namely, that Ronald Reagan was married twice) while highlighting different facets of that claim (one highlights Reagan's film career, the other his political role). Analogously, positive and negative counterspeech tell us something similar about the state of the world, while emphasizing different features of that state: while negative counterspeech reactivates incorrect associations in the process of negating them, positive counterspeech activates correct associations.

Consider an example. In his "I Have a Dream" speech, Martin Luther King Jr. takes a stand against a deeply racist public culture, which is rife with statements professing that black Americans are morally, culturally, and intellectually inferior to whites. Famously, King does so by vividly depicting a vision of racial equality. In this vision, the American nation "will rise up and live out the true meaning of its creed: 'We hold these truths to be self-evident, that all men are created equal,'" and, accordingly, "the sons of former slaves and the sons of former slave owners will be able to sit down together at the table of brotherhood."[36] King's utterance is very different from saying: "That's false! Black Americans are *not* immoral, uncivilized, or stupid!" While his vision of racial equality entails—or at the very least, strongly implicates—the falsehood of these racist stereotypes, it is first and foremost a positive representation of racial equality. This difference matters. Because it does not repeat the racist stereotypes, King's positive speech does not enhance their salience in the public sphere. Instead, it draws his audience's attention away from these stereotypes, and toward a different—and incompatible—set of associations: namely, between the American dream and the values of equality and interracial unity.[37]

---

nition," 397n10). So the analogy at hand does not endanger my thesis that the salience-based asymmetry might work for correct and positive associations: it suggests, rather, that this effect might work even more strongly when the associations in question are negative.

36 King, "I Have a Dream."

37 As this example indicates, in advocating positive counterspeech—which, when leveled at hateful speech, may well involve putting forward an egalitarian vision—I am in agreement with McGowan's suggestion that speech can be used to enact egalitarian norms (*Just Words*, 185–89). There are, however, at least two reasons why McGowan's discussion of positive speech needs developing. First, because McGowan only discusses the positive potential

Positive counterspeech can also be used to oppose ordinary political misinformation. In response to a misleading statement *X* (say, that the UK pays the EU £350 million a week, which could benefit the National Health Service instead), one strategy is to say: "*X* is false. We do not pay that amount." But another strategy would be to conduct a public education campaign that avoids mentioning *X* and instead draws the public's attention to facts that entail *X*'s falsehood. Such a campaign might publicize facts about what the UK *receives* from the EU (including, if applicable, how free movement in Europe helps staff the National Health Service).[38] And it might also reveal facts about what the UK actually pays the EU (expressed, perhaps, as a proportion of overall spending, so as to highlight its comparatively small size). Such a public response would entail that the UK does not pay the EU as much as *X* claimed, and that it is misleading to posit a contrast between funding the EU and funding the National Health Service. In doing so, however, it would avoid repeating the £350 million figure, and would therefore avoid reinforcing its salience.

In fact, there is empirical evidence that such positive responses to political misinformation are more likely to succeed. Notice, first, that the pessimistic evidence mentioned in section 1.2 involved paradigmatic instances of *negative* counterspeech. The CDC "myth versus fact" flyers about vaccines worked by reiterating myths and labeling them "false." Moreover, the result that correcting Obamacare death-panel rumors was ineffective obtained primarily in cases where the rumors were explicitly repeated beforehand. By contrast, when subjects did *not* repeat the rumor beforehand, corrections actually tended to reduce support for the rumor.[39] Likewise, Lewandowsky et al. report that people are more likely to reject false reports about how a fire started if, instead of simply being told that the initial report was false, they are given an alternative narrative of how the fire began.[40] On this basis, they hypothesize that, when correcting misinformation, it is better to "emphasize the facts you wish to communicate rather than the myth."[41]

---

of speech by way of concluding *Just Words*, her exploration of this idea is quite brief and suggestive. Second, and more substantively, McGowan does not connect her concluding discussion of the positive potential of speech with her earlier critique of counterspeech earlier in *Just Words*. Specifically, in chapter 5, McGowan criticizes the use of counterspeech to respond to oppressive speech by appealing to the asymmetric pliability of conversational norms (*Just Words*, 11–20). What I am suggesting is that, provided counterspeech takes a "positive" form (as I define it here), it can go some way toward alleviating this concern.

38 Blitz, "Brexit and the 'NHS Dividend.'"

39 Berinsky, "Rumors and Health Care Reform," 258.

40 Lewandowsky et al., "Misinformation and Its Correction," 117.

41 Lewandowsky et al., "Misinformation and Its Correction," 123.

In light of the positive conception of counterspeech, counterspeech seems more promising than trying to unring a bell. The conversational norms enacted by ignorant speech are difficult to reverse primarily when counterspeech takes a negative form, which repeats the ignorant contents to debunk them. But positive counterspeech is less vulnerable to this problem. It draws attention away from false associations, and instead renders salient correct associations that entail or implicate the falsehood of ignorant utterances. In doing so, positive counterspeech repurposes the asymmetric pliability of conversational norms, and makes it work for truth rather than falsehood.

2.2. Concerns with Positive Counterspeech

Even though distinguishing between positive and negative counterspeech helps alleviate the problem of stickiness, difficulties remain. The first problem is that, in some cases, positive counterspeech seems unavailable. Positive counterspeech affirms a view that entails or implicates the falsehood of the targeted ignorant speech *without* repeating or invoking its contents. So, positive counterspeech is available as an alternative to negative counterspeech only insofar as there exists a way of presenting the negation of the ignorant contents that does not explicitly invoke those contents. For instance, in King's "I Have a Dream" speech, positive counterspeech is possible because there is a way of presenting racial equality—namely, King's vision of interracial unity and brotherhood—that is not simply the explicit negation of racist stereotypes.

For some instances of ignorant speech, however, there is no way of saying something that entails or suggests the falsehood of the ignorant speech besides invoking and rejecting its ignorant contents. Conspiracy theories seem especially liable to generate this difficulty. By definition, conspiracy theories propound extremely improbable causal stories, which lack credible evidence. *Because* these stories are so improbable, it is difficult verbally to oppose conspiracy theories in a way that does not inadvertently invoke—or "ring the bell" of—the conspiracy theory.

Consider birtherism, the theory that Barack Obama was born in Kenya and is therefore ineligible for the US presidency.[42] Negative counterspeech—saying that Obama was *not* born in Kenya—clearly reactivates the association between Obama and Kenya. But it is not clear what the positive alternative would be. Vigorously affirming that Obama was born in Hawaii would be a highly unusual thing to do, were it not for the concern that Obama was born outside the US. Put differently, because the conversational salience of the birtherist myth is the main reason why Obama's birthplace is politically worth talking about, it is difficult

42 Smith and Tau, "Birtherism."

to say anything on this topic without inadvertently invoking or reactivating the birtherist myth. Given this background fact about conversational salience, insisting that Obama was born in Hawaii risks, like negative counterspeech, calling attention to the concern that he was born in Kenya.

The problem so far has been that genuine positive counterspeech—which opposes falsehoods without reinvoking them—may sometimes be unavailable. But there is a more fundamental problem. Even when we *can* use positive counterspeech to reverse the conversational salience of ignorant associations, counterspeech cannot reverse the harms that ignorant speech brought about while it was salient. McGowan articulates this concern when discussing sexist speech:

> the mere fact that an act of oppression can easily be reversed does not entirely disqualify it as oppressive.... While shorter-lived [oppressive norms] may be *less* oppressive ... they are still oppressive.[43]

The point is straightforward. Even if counterspeech can deactivate vilifying stereotypes, it presumably cannot change the fact *that they had been activated up to that point.* And, consequently, it cannot change the fact that, while they were activated, their salience gave rise to harms.

This concern is not specific to oppressive or hateful types of ignorant speech. Indeed, Laura Caponetto's taxonomy of the ways in which speech can "undo" the effects of prior utterances lends more general support to the worry. According to Caponetto, speech can *recognize* that a past utterance failed to enact the norms we thought it did. For instance, when a newspaper reveals that Father Tom was not actually a priest, and the Roman Rota (or a similarly qualified institution) responds by declaring that the marriages he officiated are null and void, this declaration constitutes a recognition that the validity of these marriages was only purported all along.[44] Alternatively, speech can *amend* or *retract* conversational norms that were enacted by prior utterances, so that these norms are discontinued.[45] But what Caponetto explicitly refrains from saying is that speech can retroactively make it the case that certain conversational norms were never enacted to begin with.[46]

If Caponetto is right to refrain from attributing this retroactive ability to speech, then this highlights a second sense in which counterspeech might seem as futile as trying to unring a bell. Even if you can eventually stop a bell from continuing to ring, you cannot undo the fact that it rang in the first place. Analogous-

43 McGowan, "Oppressive Speech," 404–5.

44 Caponetto, "Undoing Things with Words," 2, 6–8.

45 Caponetto, "Undoing Things with Words," 9–14.

46 Caponetto, "Undoing Things with Words," 10.

ly, even if positive counterspeech can stop the harmful conversational effects of ignorant speech from *persisting*, it can never change the fact that these harmful effects *occurred up to that point*. So, even in the best cases, positive counterspeech seems a limited remedy for ignorant speech: it cannot fully eliminate its harms.

Still, even if it is true that counterspeech is limited in this way, one might respond that it is unfair to hold this against proponents of counterspeech. After all, asking of counterspeech that it somehow fully eliminate the harms of ignorant speech, or that it somehow make it the case that the ignorant speech never occurred, may seem impossibly demanding. And if this demand is unrealistic or unachievable, then it seems misguided to make it.

What this response reveals is that the objection at hand makes sense only in light of a particular understanding of what the alternative to counterspeech should be. Critics of counterspeech are typically not suggesting that, in place of counterspeech, we do nothing. Rather, they commonly recommend implementing legal restrictions.[47] Crucially, the legal regulation of hate speech, or of fake news, is often justified as a deterrent: by imposing penalties, criminal or civil laws get people to refrain from uttering hateful or deceptive contents. When legal remedies succeed as deterrents, they *fully* eliminate the harms that suppressed utterances would otherwise have occasioned. Indeed, insofar as legal remedies prevent ignorant utterances altogether, they prevent any harms these might have generated. Thus, to the extent that legal restrictions succeed as deterrents—an assumption I will revisit in section 3.2—the objection at hand places counterspeech at a comparative disadvantage relative to legal remedies: whereas counterspeech comes in too late to eliminate all of the harms associated with ignorant speech, legal remedies can in principle fully remove those harms. This, in turn, would constitute a reason for preferring legal remedies to counterspeech.

The following section will argue that the two concerns just outlined rely on an overly crude conception of the temporality of counterspeech. My focus will primarily be on the second concern—that counterspeech, even if it takes a positive form, cannot reverse the harmful effects that ignorant speech generated prior to being countered. Nevertheless, in closing, I will also indicate how the view of counterspeech's temporality that I develop can defuse the first concern.

3. THE TEMPORALITY OF COUNTERSPEECH

There are at least two potential ways of rethinking the temporality of counterspeech to address the above objections: the first approach emphasizes the retroactive character of counterspeech; the second underscores its diachronic nature.

47 See, e.g., McGowan, *Just Words*, ch. 7.

While the first strategy is fascinating, I will suggest that, even if we grant its theoretical viability, it ultimately cannot overcome the objections at hand (3.1). The second approach, by contrast, meets with greater success (3.2).

3.1. Retroactive Counterspeech

In the first place, one might object that the objections raised in section 2.2 do not take the *retroactive* potential of counterspeech sufficiently seriously.

It seems perfectly natural to think that the past cannot be changed. By extension, the idea that counterspeech could alter the past may, as Caponetto labels it, seem "magica[l]."[48] Accordingly, one might well conclude that speech can at best stop the conversational effects of prior utterances from persisting, but it cannot make it the case that those effects never happened to begin with—and hence, it cannot fully eliminate the harms of ignorant speech.

However, Rae Langton has recently argued that this skepticism is misplaced. Beyond putting a stop to the preexisting conversational effects of past utterances, Langton suggests, counterspeech can retroactively block the occurrence of these conversational effects altogether. More precisely, retroactive counterspeech "changes a past utterance from the unactualized way it would have been, to the way it actually is."[49] In this way, counterspeech "offers a ticket to a modest time machine."[50] Applied to ignorant speech, the idea is that counterspeech can make it the case that prior utterances failed to ever generate certain damaging conversational effects. If so, then one might think that—*contra* the second objection raised in section 2.2—counterspeech can entirely eliminate the harms of ignorant speech.

The immediate question, however, is why we should accept Langton's counterintuitive claim that counterspeech can retroactively alter the nature of prior utterances. Langton offers two reasons. The first is that, in many other domains, we already recognize that an act's character can retroactively be altered by future happenings. "A stabbing may be a killing," Langton observes, "partly in virtue of what happens later."[51] In other words, the fact that a stabbing is also a killing may depend on the fact that the victim *later* succumbed to her wounds. If doctors had managed to save the victim's life, the past stabbing would arguably have constituted a nonlethal assault instead. Thus, if Langton is correct, we cannot dismiss the idea of retroactive counterspeech simply on the grounds that retroactivity seems "magical"—we already recognize retroactive phenomena in everyday life.

48 Caponetto, "Undoing Things with Words," 10.

49 Langton, "Blocking as Counterspeech," 156.

50 Langton, "Blocking as Counterspeech," 146.

51 Langton, "Blocking as Counterspeech," 157.

Nevertheless, even if some facts can retroactively be made true, one might doubt whether this applies to facts about the effects of past utterances. To establish this stronger claim, Langton offers a second argument. Drawing on Austinian speech-act theory, Langton distinguishes two kinds of things utterances can do. Utterances can *cause* things to happen. For instance, delivering a brilliant argument can cause one's audience to affirm its conclusion. This is roughly what Austin refers to as an utterance's perlocutionary dimension. But utterances can also *constitute* certain acts. When a priest says, "I hereby pronounce you husband and wife," she is not merely causing the couple to be married. Instead, her utterance is part and parcel of, or constitutive of, the act of *marrying* them. The speech act(s) constituted in saying something belong(s) to the utterance's illocutionary dimension.[52]

For an utterance successfully to constitute a particular speech act—marrying a couple, declaring war, etc.—certain background conditions ("felicity conditions") must be satisfied.[53] What kinds of felicity conditions are needed will vary from speech act to speech act. According to McGowan, however, most speech acts depend on the speaker having a particular status or social power.[54] Some speech acts (such as marrying, condemning, or ordering) require that the speaker have a comparatively high or special status, which is commonly referred to as "authority."[55] But, McGowan insists, even speech acts that do not require authority nevertheless require that speakers have a minimal and widespread status (which she calls "standing") in virtue of which they are capable of taking part in conversations.[56] Should key felicity conditions not be satisfied, the utterance will misfire, or fail to constitute the relevant speech act.[57] If the officiant for a religious wedding has not been ordained, for instance, then she lacks the relevant authority and her utterance, "I hereby pronounce you married," will fail to marry the couple. Similarly, if a lowly private (rather than a general) shouts "Attack!" to an assembled army, he will (in normal conditions) fail to order that army into battle.

52 Langton, "Speech Acts and Unspeakable Acts," 300–1. See also Austin, *How to Do Things with Words*, 98–108.

53 Austin, *How to Do Things with Words*, 12–49.

54 McGowan, *Just Words*, 15, 63–68.

55 McGowan, *Just Words*, 63–66. See also Langton, "Speech Acts and Unspeakable Acts," 304–5; and Austin, *How to Do Things with Words*, 23–24.

56 McGowan, *Just Words*, 66–68.

57 Note, however, that not all felicity conditions are like this: some are such that, when they fail to be satisfied, the intended speech act is nonetheless constituted, though it is nonideal. See Austin, *How to Do Things with Words*, 39–52.

Crucially, Langton claims, felicity conditions for speech acts can sometimes be satisfied in the future, after the utterance. By way of illustration, she considers the case of marriage: "for 'I do' to count as a marriage," she suggests, "may require felicity conditions in the future, e.g., the consummation of the marriage."[58] This, on Langton's analysis, is what makes retroactively undoing a speech act possible: if a future felicity condition fails to be satisfied, then it follows that the past utterance misfired. Applied to the marriage case, the idea is that, if some "sad events" keep the couple from consummating the marriage, this makes it the case that the earlier "I do" failed to constitute a speech act of marrying.[59]

The final step in Langton's argument is to emphasize that counterspeech can be a way of challenging the future felicity conditions of speech acts. To see this, consider again that many speech acts require that the speaker have authority. Thus, in attempting to perform such speech acts, speakers tend to presuppose that they have the relevant kind of authority. And, except in cases where their authority is already firmly entrenched, they depend on their audience to accept that presupposition. Consequently, a significant way in which counterspeech can block a speech act is by challenging such presuppositions of authority.[60] A colleague tells you, "A double espresso; make it quick!" and you furiously respond, "Who do you think you are? I don't take orders from you!" The colleague's utterance presupposes that they have the authority needed to give you orders. In denying that presupposition, you prevent their speech from constituting an order. At best, it becomes a (very unsuccessful) request.[61] The thought, then, is that by undermining a future felicity condition of the utterance, your response retroactively changes the nature of that utterance.

Returning to ignorant speech, if this theoretical analysis is correct, it suggests that counterspeech can retroactively make it the case that past ignorant utterances failed to enact problematic conversational norms. Imagine that *A* tells *B*: "Your kind are nothing but worthless parasites!" In uttering this falsehood, as we saw in section 1.2, *A* might be attempting to constitute the following harmful speech act: to refute the assurance that *B* is a member of society in good standing. Now, suppose that *A* is part of a minority extremist political group—a group whose authority to determine others' social status is not already firmly established.[62] And suppose, moreover, that *C*—a spokesperson for the majority—in-

58 Langton, "Blocking as Counterspeech," 157.

59 Langton, "Blocking as Counterspeech."

60 Langton, "Blocking as Counterspeech," 150.

61 Langton, "Blocking as Counterspeech," 156.

62 This stipulation is important. As I specified in the previous paragraph, the ability for counterspeech to block a harmful speech act by challenging a speaker's authority is restricted to

terjects: "You don't speak for us, *A*. And we will *never* condone that kind of view." In challenging *A*'s presupposed authority, *C* retroactively makes *A*'s utterance misfire. *A*'s utterance does not refute the assurance of *B*'s standing, because *A* lacks the requisite authority. In sum, if Langton's theoretical account is correct, then—*contra* the second objection outlined in section 2.2—counterspeech can do more than simply stop ignorant utterances from continuing to harm: by challenging the future felicity conditions of prior ignorant utterances, it can retroactively prevent past utterances from ever constituting harmful speech acts.

However, there are problems with the idea of retroactive counterspeech. In the first place, Langton's arguments for its possibility might seem contentious. For one thing, one might deny that the concrete cases Langton points to really do intuitively involve retroactive phenomena. In the stabbing case, for instance, one might give *epistemic* significance, rather than constitutive significance, to the victim's death. On this view, the victim's death merely shows us that the previous stabbing was also a killing. Alternatively, one might think that the death does make the stabbing a killing, while denying that this constitutive function is retroactive: that is, after the death has occurred, the stabbing counts as a killing from the time of death onward, but not before. As for the marriage case, it is unclear that consummation is actually a necessary felicity condition for the success of speech acts of marrying.[63] If so, then the failure to consummate the marriage may not actually retroactively affect its validity.

Moreover, the argument for retroactive counterspeech may come at a theoretical cost. The stabbing case suggests that the future felicity conditions for an utterance to constitute a specific illocutionary act may include some of its perlocutionary effects. As a result, one might worry that, in explaining how retroactive counterspeech is possible, Langton's account ends up blurring the distinction between illocutionary and perlocutionary dimensions of utterances.[64]

Nevertheless, I want to grant, for the sake of argument, that these concerns about the possibility of retroactive counterspeech can be overcome. My main point is that, *even if* we assume that retroactive counterspeech is possible (and

cases where the speaker somehow depends on the audience either accepting or at least not challenging their authority. In cases where the speaker's authority is already firmly established (for example, if the chancellor of the Third Reich said the same thing as *A*) counterspeech may not be able to retroactively block the harmful speech act. In section 4, I briefly reiterate this limitation and its implications for whether we ought to adopt legal restrictions.

63 Indeed, in Catholic doctrine, consummation may be required to make a marriage *indissoluble*. But this is consistent with a marriage being *valid*. Here, and in the previous case, I am grateful to a reviewer for drawing my attention to these complications.

64 I owe this insight to a reviewer. For further criticisms of Langton's account of retroactive blocking, see McGowan, *Just Words*, 48.

that it is realized in the cases Langton points to), it cannot entirely defuse the present objection.

Langton's analysis of how speech can retroactively block earlier utterances focuses exclusively on utterances' illocutionary dimension—i.e., on what acts might be *constituted* by those utterances. But, as explained above, utterances also have a perlocutionary dimension: besides constituting acts, utterances also have *causal* effects, such as getting listeners to believe something. The retroactive picture of counterspeech does not establish, and is not intended to establish, that these causal effects can be undone.

First, even if counterspeech prevents an utterance from constituting a particular speech act, that utterance might still causally influence listeners *as if* the speech act had been successful. This is because listeners may not know that the felicity conditions are not satisfied. Suppose Langton is right that, if *A* and *B* fail to consummate their marriage, their "I do's" fail to constitute the speech act of *marrying*. Even so, if people never find out, the utterance "I do" will retain its habitual causal consequences. Friends will congratulate the couple on subsequent anniversaries, the Internal Revenue Service will tax them as a married couple, immigration services will grant *A* a spousal visa when *B* takes a job in a different country, and so on.

Second, there is no way of undoing these causal effects after the fact. To reiterate, the reason why acts constituted by utterances can retroactively be undone is that their performance depends on felicity conditions, and felicity conditions can sometimes be satisfied after the utterance. By contrast, the causal effects of utterances do not depend on felicity conditions in this way. Therefore, once the causal effects have occurred, nothing can change this fact. As Caponetto nicely illustrates: "one can take back a marriage proposal, but one cannot take back the hearer's excitement at hearing the words 'Will you marry me?'"[65]

This spells trouble for counterspeech directed at ignorant speech, because many of the harms attributed to ignorant speech are causal. For instance, one of the ways in which hate speech harms its targets is by causing them to experience psychological distress, particularly when the hate speaker is perceived to have significant social authority. Now, for Alexander Brown, it is "hard to see how any counterspeech could ameliorate some of the[se] psychological harms."[66] Above, I suggested that counterspeech can undermine hate speakers' authority. If the retroactive blocking account is correct, doing so can retroactively prevent hate speech from constituting certain harmful speech acts (since it undermines a felicity condition for such speech acts). Furthermore, it can help targets feel less

65 Caponetto, "Undoing Things with Words," 4.

66 Brown, "Hate Speech Law," 260.

distressed by the hate speech than they previously did (once they realize that it only represents the views of a small minority). Even so, Brown's point is that counterspeech cannot undo the distress experienced by targets of hate speech *in the intervening period*, before counterspeech told them that the hate speakers lacked social support.

A similar observation applies to ordinary political misinformation. Counterspeech might, by undermining the authority (or even standing) of a source of misinformation, get the audience at large subsequently to reject that source's claims. But it cannot change the causal harms generated by political misinformation *before* it was verbally countered. Imagine that, following an election, an anonymous blogger publicly and falsely claims that the election was rigged. While the vast majority of people withhold judgment as to whether the blog post is reliable, and some express uncertainty about its reliability, a minority flies into a rage upon hearing this, and responds by destroying private property. Even if counterspeech later repudiates the blogger's credibility in a way that all find convincing, it clearly cannot undo the destruction caused in the intervening period.[67]

The upshot is that, at best, the appeal to retroactive counterspeech can only help address the objection that counterspeech cannot undo the harms generated by ignorant speech before it was countered to a limited extent. If Langton is right, counterspeech can retroactively make it the case that a prior ignorant utterance did not constitute a harmful speech act. But it cannot retroactively undo the past causal harms of ignorant speech.

This limitation matters, once more, in light of the comparison between counterspeech and legal remedies. By contrast with counterspeech, insofar as legal remedies successfully prevent ignorant speech from being uttered, they can entirely eliminate its harms, both constitutive *and* causal. So, even if we were

67 A possible complication with this example is that, on Maitra's account of "licensed authority," it is possible for a speaker to acquire authority even if the majority does not accept their claims ("Subordinating Speech," 107). Consequently, one might worry that, in the example, the blogger succeeds in gaining authority prior to the counterspeech, so that the counterspeech fails to retroactively block their authority. A first response is that it is unclear that this example really is a case of licensing in Maitra's sense. In Maitra's discussion, the phenomenon of licensing standardly involves the audience remaining silent, or not making their reservations public ("Subordinating Speech," 107, 116). By contrast, in the above case, even before the challenge to the blogger's presupposed authority occurs, some members of the majority express their uncertainty about the blogger's credibility. Second, and more fundamentally, this concern does not undermine my main point here: that point, once more, is that *even if we assume*, for the sake of argument, that the counterspeech in this example has successfully retroactively blocked the blogger's authority, it still cannot possibly reverse the causal harms produced by the misinformation in the meantime.

to grant Langton's theoretical case for thinking that counterspeech can operate retroactively, counterspeech would still in principle remain at a comparative disadvantage with legal remedies.

3.2. Diachronic Counterspeech

There is a second way in which the objection under consideration—that counterspeech cannot undo the harms generated by ignorant speech before it was countered—depends on an inadequate account of the temporality of counterspeech. Simply put, counterspeech should not merely be interpreted as speech that responds to ignorant utterances after the fact. Rather, counterspeech is better understood as diachronic—as a continuous process, extended over time, that precedes as well as follows ignorant speech. Thus understood, counterspeech might involve educating listeners to inoculate them against future ignorant speech, or preemptively warning listeners about unreliable sources.

A diachronic understanding is implicit in some defenses of counterspeech. When discussing state counterspeech, for instance, Brettschneider recommends having schools teach children about important political matters, such as the Holocaust.[68] This counterspeech is diachronic insofar as it may affect its audience *prior* to their exposure to relevant ignorant utterances: schoolchildren may never have encountered Holocaust denial. Similarly, in her analysis of "toxic speech," Tirrell recognizes that counterspeech can work both as a *post hoc* "antidote" and as a preventive "inoculation."[69]

Nevertheless, the vast majority of discussions of counterspeech articulated by philosophers of language—including the objection currently under examination—cast counterspeech as a *post hoc* response to ignorant utterances. Caponetto's examples of undoing things with words, for example, are invariably verbal attempts at disabling an earlier utterance.[70] Likewise, Maitra and McGowan criticize counterspeech partly by appealing to evidence that targets of face-to-face hate speech rarely respond to that speech then and there. Here too, counterspeech is understood as a *post hoc* response.[71] In recent work, McGowan does briefly acknowledge that counterspeech could take a broader form, such as "a general education campaign." But she insists that her focus, in critiquing counterspeech, is on counterspeech understood as "direct and fairly immediate

68 Brettschneider, *When the State Speaks, What Should It Say?* 96–104.

69 Tirrell, "Toxic Speech," 136–39. See also Richards and Calvert, "Counterspeech 2000," 569–74; and Gelber, "Reconceptualizing Counterspeech in Hate Speech Policy (with a Focus on Australia)," 214–15.

70 Caponetto, "Undoing Things with Words." See also Langton, "Blocking as Counterspeech."

71 Maitra and McGowan, "Introduction and Overview," 10.

responses to allegedly harmful utterances."[72] However, limiting her scope in this way substantially limits the significance of her criticisms. Defenders of counterspeech can reply that they are recommending not simply *post hoc* responses, but counterspeech that may both preempt and follow ignorant speech. Indeed, such diachronic counterspeech, I will now suggest, can go some way toward defusing the objection at hand.

When counterspeech occurs *after* ignorant speech, the problem is that ignorant speech can generate harms prior to being countered, and—at least in the case of causal harms—these harms cannot be undone. How does counterspeech that precedes ignorant speech help? The idea is that preemptive counterspeech can condition the conversational setting to make it inhospitable to ignorant speech. Consequently, when ignorant speech is later uttered, it fails to have its damaging conversational effects. This helps address the problem: if preemptive counterspeech disables ignorant speech before it is uttered, then there is no interval during which ignorant speech can, unopposed, generate irreversible harms.

The basic idea is straightforward. What needs elaboration is how precisely preemptive counterspeech might condition the conversational setting so as to disable subsequent ignorant utterances. There are at least two promising strategies. The first is to expound and widely diffuse important politically relevant facts. Advancing such truths alters the conversation in several ways. First, it renders those truths conversationally salient, and thereby familiarizes listeners with them. Relatedly, it facilitates the introduction of these truths into the common ground of shared beliefs. Indeed, preemptively expounding facts can induce listeners to believe those facts for various reasons: because the speaker is perceived as authoritative; because the speaker offers compelling arguments for them; or simply because, as we have seen, the more familiar propositions are to listeners, the more listeners are disposed to believe them.

This conditioning makes it more difficult for falsehoods to subsequently gain a foothold in the conversation. Insofar as listeners are knowledgeable about a subject matter, they are less vulnerable to being swayed by ignorant speech. Holocaust denial or misleading claims about EU funding are less readily accepted if they contradict propositions on these topics that are already part of the audience's common ground. Moreover, as discussed in section 1.1, it is harder to make ideas unsalient than salient. Accordingly, if true propositions have already been made salient by preemptive counterspeech—which, in turn, disposes the audience to believe them—ignorant speech may struggle to render them unsalient.

72 McGowan, "Responding to Harmful Speech," 183.

A second preemptive strategy is to warn the audience about unreliable sources. By undermining the authority of untrustworthy sources, such preemptive counterspeech prevents those sources' utterances from affecting conversations in damaging ways. This idea is akin to the silencing effect that philosophers of language often ascribe to hate speech and pornography. Hate speech and pornography, it is often said, prevent minorities and women from contributing effectively to conversations.[73] One of the ways in which they do so is by stripping these groups of their authority.[74] When hate speech persistently represents racial minorities as inferior, for instance, they may lose the social standing needed for their speech to be taken seriously.

I am suggesting that this silencing effect can also be used for good, by challenging the authority of prospective promoters of ignorance. This usage applies most clearly to political misinformation: if hearers are credibly warned that a political news site has been systematically wrong, that source's subsequent utterances may be taken less seriously. Nevertheless, it also applies to hateful ignorant speech. Suppose a democratically elected head of state affirms: "We categorically reject racist ideologies. Racism is unwelcome here." The hate speaker who *then* pronounces deeply racist views thereby marks himself as a minority voice, who cannot speak for the majority.

Thus, counterspeech can condition the conversation to protect it from subsequent ignorant speech in at least two ways: by entrenching important facts in the conversation's common ground, and by eroding ignorance-promoting speakers' status, so that they no longer count as authorities.

Importantly, this preemptive conditioning is relevant to the constitutive *and* causal effects of ignorant speech. Like retroactive counterspeech, preemptive counterspeech can undermine the felicity conditions (such as authority) that ignorant utterances need to constitute harmful acts. The only difference with

73 E.g., Langton, "Speech Acts and Unspeakable Acts"; Maitra and McGowan, "Introduction and Overview"; and McGowan, "Responding to Harmful Speech." Note that I am using "silencing" in a loose sense. First, as I am using the idea, silencing includes both perlocutionary silencing (preventing speech from having its intended causal effects) and illocutionary silencing (preventing speech from constituting its intended speech acts). On this distinction, see Langton, "Speech Acts and Unspeakable Acts," 314–15. Second, my use of the term is broader than typical discussions of silencing in the following way. Typically, silencing involves the speaker being prevented from doing certain things with their words as a result of a recognition failure on the part of the hearer. For example, the speaker might have authority, but the hearer fails to recognize this. However, in the cases I am concerned with here, silencing arguably does not involve a recognition *failure*: the speaker actually lacks preestablished authority, and the hearer is right to think that they do not satisfy the authority condition. I am grateful to a reviewer for this insight.

74 McGowan, "Debate," 491–92.

retroactive counterspeech is that, here, the felicity conditions are undermined *before* the problematic utterance.

In addition, preemptive counterspeech alleviates the difficulties that retroactive counterspeech encountered with ignorant utterances' causal harms. The causal effects of speech depend in significant part on what that speech is *perceived as.*[75] The problem with retroactive counterspeech is that, in the interval before ignorant speech is countered, a mismatch might arise between what the utterance actually *is* and what it is *perceived as*. Before retroactive counterspeech established that the purveyor of the vaccines-cause-autism myth was a crank, people may have believed that he was delivering an authoritative verdict, and followed his dangerous prescriptions. Preemptive counterspeech addresses this causal problem by eliminating the interval in which ignorant speech has been uttered but not yet countered. In doing so, it reduces the likelihood of a temporary mismatch between what the ignorance is and what it is perceived as. If listeners already know that A is a crank when A speaks, they are less vulnerable to being duped into harmful actions.

This theoretical account of how counterspeech might preemptively defuse the constitutive and causal effects of ignorant speech is not merely speculative. Cook, Lewandowsky, and Ecker, for instance, find that telling subjects about the scientific consensus on anthropogenic climate change neutralizes the causal impact of subsequent climate-related misinformation.[76] This corroborates the first preemptive strategy outlined above. By giving people authoritative testimony that climate change is real, the preemptive speech introduces this proposition into the common ground of shared beliefs. This makes it harder for subsequent climate-related misinformation to gain assent.

There is also support for the second preemptive strategy: casting doubt on the authority or credibility of unreliable sources. Lewandowsky et al. report that "misinformation effects can be reduced if people are explicitly warned … that information they are about to be given may be misleading."[77] For example, when people are warned that upcoming information about climate change may come

75 Note that, while very many causal effects of utterances depend on how the utterance—and more specifically, its force—is perceived, some are not like this. The utterance "Move aside!" may cause me to move aside if I perceive it as an order. In addition, however, if it is extremely loud, it might also cause my ears to hurt or cause me to be startled. Arguably, neither of these latter two causal effects depends on what I perceive the utterance as. That being said, the harmful causal effects of ignorant speech I am concerned with here generally tend to stem from the perceived force of the utterance.

76 Cook, Lewandowsky, and Ecker, "Neutralizing Misinformation through Inoculation," 10.

77 Lewandowsky et al., "Misinformation and Its Correction," 116.

from nonexperts, and given some indication of how to identify nonexperts, the influence of subsequent misinformation on their beliefs disappears.[78]

So far, I have argued that adopting a diachronic view of counterspeech—where counterspeech can preemptively disarm ignorant speech as well as retrospectively oppose it—helps overcome the main objection considered in section 2.2: that counterspeech can only stop the conversational effects of ignorant speech from persisting, and hence cannot wholly eliminate its harms. But switching to a diachronic view of counterspeech *also* helps counter the other objection raised in section 2.2: that some kinds of ignorant speech—e.g., conspiracy theories—cannot be opposed in a positive way. The only way to counter them, it seemed, is to negate their content. However, doing so risks increasing their salience, and thereby exacerbating their damaging effects.

The idea of preemptive counterspeech reveals a better response. The problem with these utterances is that, once they gain traction, it is difficult to oppose them in a way that does not backfire. To circumvent this difficulty, a more promising approach would be to preemptively condition the conversational context so that they never gain traction to begin with. Now, if the only way to do so were to preemptively deny ignorant propositions ("You may hear next week that Obama was born in Kenya. But rest assured: he was born in Hawaii"), one might worry that this strategy simply reproduces the initial problem: preemptive denial risks inadvertently reinforcing the salience of ignorant associations by repeating them.

Crucially, however, one of the principal preemptive strategies outlined above involves casting doubt on the authority of speakers, rather than directly criticizing the content of their utterances. The recommendation that emerges from the diachronic understanding of counterspeech, then, is that we should preemptively warn people about, say, fake news sites spewing conspiracy theories. We can do this by preemptively identifying specific fake news sites as unreliable; or, alternatively, by warning people that such unreliable sites exist and teaching them how to identify them. As we have seen, such warnings can drastically reduce the audience's vulnerability to misinformation.[79] So, diachronic counterspeech is doubly helpful: not only does it show how counterspeech might wholly forestall the damaging effects of ignorant speech, but it also provides guidance for handling particularly resilient kinds of misinformation.

It is important not to overstate the present section's argument. I am *not* claiming that diachronic counterspeech will always be entirely successful at defusing ignorant speech. Rather, in highlighting diachronic counterspeech, I am making a more restricted point: that, like legal remedies, counterspeech too can

78 Cook, Lewandowsky, and Ecker, "Neutralizing Misinformation through Inoculation," 15.

79 Cook, Lewandowsky, and Ecker, "Neutralizing Misinformation through Inoculation," 11–12.

in principle wholly defuse the harms of ignorant speech, and it can do so even in hard cases where positive counterspeech is unavailable.

That counterspeech may not do this perfectly is a problem, but not a comparative problem. The objections raised in section 2.2 suggested that, *insofar as legal remedies succeed* in deterring ignorant speech, they can wholly eliminate the harms that such speech would otherwise have produced. The degree to which they are in fact successful, however, is limited. In their influential study of hate-speech laws in Australia, for instance, Gelber and McNamara find that the enactment of hate-speech laws did not reduce the incidence of verbal abuse.[80] What this indicates is that, just as diachronic counterspeech is imperfect at disabling ignorant speech, so too legal remedies are imperfect at deterring ignorant speech. Thus, as stated, the objections in section 2.2 fail to provide a principled reason for preferring legal remedies to counterspeech.

In what follows, after summarizing my broader argument, I will briefly examine what this means for the division of labor between legal and speech-based responses to ignorant speech.

4. Conclusion

Whether it takes the form of hate speech or of ordinary political misinformation, ignorant speech tends to be sticky. Because of the distinctive properties of conversational salience, the damaging effects of ignorant speech on conversational norms are typically easier to enact than to reverse. In fact, verbally countering ignorant speech may simply amplify its salience.

Even so, I have argued that refining counterspeech along two dimensions can substantially mitigate this problem. First, we should distinguish between positive and negative forms of counterspeech. Instead of explicitly negating ignorant speech, positive counterspeech affirms a vision of the truth, which entails or implicates the falsehood of the ignorant utterance without repeating it. Hence, positive counterspeech contradicts ignorant speech without magnifying its salience.

The emphasis on positive counterspeech nevertheless cannot fully address the problem at hand. Not all kinds of ignorant speech can be countered in a positive way. And, more fundamentally, even when positive counterspeech helps to

80 Gelber and McNamara, "The Effects of Civil Hate Speech Laws," 644–45. Moreover, while they do find that the language used to express prejudice in newspapers became tamer, this—as Heinze observes—could simply be explained by hate speech taking a more coded form (*Hate Speech in Democratic Citizenship*, 145–48). For more general discussion of this problem, see Heinze, *Hate Speech in Democratic Citizenship*, 145–53; and Mchangama, Review of *The Harm in Hate Speech*, 97.

reverse the conversational effects of ignorant speech, this may not change the fact that, before they were reversed, these effects generated harms.

However, these remaining concerns rely on an unsophisticated understanding of the temporality of counterspeech. One suggestion here is that counterspeech might operate *retroactively* to prevent past ignorant utterances from ever constituting harmful speech acts. But even if we were to grant its theoretical viability, this suggestion is too narrowly focused on the harms constituted by ignorant speech to alleviate the above concern. The more decisive point, then, is that counterspeech should be understood *diachronically*, as an extended process that can both preempt and follow ignorant utterances. Preemptive counterspeech aims to condition the conversational context in a way that disables future ignorant utterances. By disarming ignorant speech before it is uttered, counterspeech can wholly eliminate its attempted harms, causal and constitutive. Moreover, preemptive counterspeech prevents especially resilient forms of ignorance, which could not be countered positively, from taking root.

This, to reiterate, is not to say that counterspeech is infallible. Positive counterparts to negative counterspeech can be difficult to find. Moreover, when an ignorant utterance has already taken root, preemptive counterspeech is no longer an option. And even when preemptive counterspeech remains an option, it is not foolproof. For instance, some speakers may be so authoritative that their verbal influence cannot fully be preemptively disabled.

Nonetheless, the point remains that the refined conception of counterspeech is substantially more resilient in the face of sticky ignorance than it initially appeared. Neither the fact that negating ignorant speech can reinforce its conversational effects, nor the fact that the causal harms of ignorant speech cannot be reversed *post hoc*, suffice to show that counterspeech is inadequate.

Accordingly, while my argument does not preclude thinking that legal remedies may sometimes be warranted to prevent sticky ignorant speech, it does establish that more needs to be said to justify such regulations. First and foremost, advocates of legal regulations need to provide more precise evidence of legal regulations' deterrent effects. If legal regulations are to be defended as a response to the stickiness of ignorant speech, its proponents must (1) offer a more specific account of the contexts in which the refined conception of counterspeech fails, and (2) show that legal regulations could successfully deter ignorant speech in those contexts. For example, they might need to show that, in contexts where the ignorance-promoting speaker's authority is so great that preemptive counterspeech would struggle to disable her speech, legal regulations would by contrast succeed in deterring her utterance. This is a much more specific challenge than simply showing that legal regulations sometimes succeed in deterring hate

speech, and one that has yet to be met. Indeed, as is often observed, even more general evidence of the deterrent effect of legal regulations—i.e., evidence that legal regulations generally tend to deter hate speech—remains sparse.[81]

But that is not all. Even if legal regulations generally succeed in preventing ignorant speech in contexts that prove problematic for counterspeech, it must also be shown that, in the process of doing so, these regulations do not inadvertently amplify the salience of ignorant speech. In other words, it must be shown that legal regulations do not run afoul of the so-called Streisand effect, whereby attempting to hide or censor something unwittingly increases its publicity.[82]

There are at least two reasons why, like negative counterspeech, legal regulations might do so. First, given the powerful expressive force of laws, publicly enacting a law prohibiting, say, degrading speech, risks drawing the public's attention to the fact that there are people who embrace and express degrading views.[83] Second, the enforcement of speech-related legal restrictions often leads to highly publicized trials that put a spotlight on violators, their utterances, and the bad associations they promote. As Simpson notably observes, in the presence of hate-speech laws, the existence of hateful citizens and hateful ideologies is "powerfully conveyed in people's preparedness to express their identity-based contempt even while faced with the threat or reality of prosecution."[84] What this underscores is that the process of implementing and enforcing laws that aim to deter ignorant speech may itself contribute to increasing the salience of such speech.

Accordingly, to establish that legal remedies are more effective ways than counterspeech of overcoming the problem of stickiness, proponents must establish not simply that their deterrent effect applies to the specific contexts where counterspeech fails, but also that the process of enacting and enforcing these laws will not amplify the salience of ignorant speech.

In this light, developing a theoretically refined understanding of counterspeech, whereby counterspeech is positively framed and extended over time, is beneficial in two respects: it offers guidance concerning how verbal responses to ignorant speech might succeed, notwithstanding its stickiness, and it clarifies

81 See note 80 above.

82 I am grateful to a reviewer for pressing me on this point.

83 For discussion of the point that laws are expressively loud, see, e.g., McAdams, *The Expressive Powers of Law*, 123.

84 Simpson, "Dignity, Harm, and Hate Speech," 724. For a similar observation, see Heinze, "Hate Speech and the Normative Foundations of Regulation," 599–600.

the empirical challenge that must be met if we are to ascertain the correct division of labor between counterspeech and legal remedies.[85]

Nuffield College, University of Oxford
maxime.lepoutre@nuffield.ox.ac.uk

REFERENCES

Alter, Adam, and Daniel Oppenheimer. "Uniting the Tribes of Fluency to Form a Metacognitive Nation." *Personality and Social Psychology Review* 13, no. 3 (August 2009): 219–35.

Anderson, Elizabeth. "The Epistemology of Democracy." *Episteme* 3, nos. 1–2 (June 2006): 8–22.

Asser, Martin. "What the Muhammad Cartoons Portray." BBC News, January 2, 2010. http://news.bbc.co.uk/2/hi/middle_east/4693292.stm.

Austin, J. L., *How to Do Things with Words.* Oxford: Oxford University Press, 1962.

Barnes, Michael Randall. "Speaking with (Subordinating) Authority." *Social Theory and Practice* 42, no. 2 (April 2016): 240–57.

Berinsky, Adam. "Rumors and Health Care Reform: Experiments in Political Misinformation." *British Journal of Political Science* 47, no. 2 (April 2017): 241–62.

Blitz, James. "Brexit and the 'NHS Dividend.'" *Financial Times*, March 14, 2018. https://www.ft.com/content/1fcaf8b0-277e-11e8-b27e-cc62a39d57a0.

Brettschneider, Corey. *When the State Speaks, What Should It Say?* Princeton: Princeton University Press, 2012.

Brown, Alexander. *Hate Speech Law: A Philosophical Examination.* Routledge: Abingdon, 2015.

85 For comments on written drafts, I am deeply grateful to Rachel Bernhard, Asli Cansunar, Joanna Demaree-Cotton, Rachel Fraser, Jess Kaplan, Cécile Laborde, Jonas Markgraf, Cathy Mason, Barbara Piotrowska, Sole Prillaman, Tamás Szigeti, Ralph Weir, Andreas Wiedemann, and two extremely helpful anonymous reviewers. This paper has also benefited greatly from discussions with Sam Berstler, Laura Caponetto, Charlie Crerar, Michael Lynch, Lynne Tirrell, Lani Watson, and audiences at the Princeton Workshop in Social Philosophy, the UConn Social Epistemology Working Group, the Oxford Ethics of Online Interaction Workshop, and the Society for Applied Philosophy Conference. This research was generously supported by the Society for Applied Philosophy as well as by a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation.

Brown, Étienne. "Propaganda, Misinformation, and the Epistemic Value of Democracy." *Critical Review* 30, nos. 3–4 (2018): 194–218.

Caponetto, Laura. "Undoing Things with Words." *Synthese*. Published ahead of print, May 25, 2018. https://link.springer.com/article/10.1007%2Fs11229-018-1805-9.

Centers for Disease Control and Prevention. "Vaccines Do Not Cause Autism." Centers for Disease Control and Prevention, October 27, 2015. https://www.cdc.gov/vaccinesafety/concerns/autism.html.

Cook, John, Stephan Lewandowsky, and Ullrich Ecker. "Neutralizing Misinformation through Inoculation." *PLOS One* 12, no. 5 (May 2017): 1–21.

Delgado, Richard. "Words that Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling." *Harvard Civil Rights-Civil Liberties Law Review* 17 (1982): 133–82.

Gelber, Katharine. "Reconceptualizing Counterspeech in Hate Speech Policy (with a Focus on Australia)." In *The Content and Context of Hate Speech*, edited by Michael Herz and Peter Molnar, 198–216. Cambridge: Cambridge University Press, 2012.

Gelber, Katharine, and Luke McNamara. "The Effects of Civil Hate Speech Laws: Lessons from Australia." *Law and Society Review* 49, no. 3 (September 2015): 631–64.

Godlee, Fiona, Jane Smith, and Harvey Marcovitch. "Wakefield's Article Linking MMR Vaccine and Autism Was Fraudulent." *BMJ* 342, no. 7788 (January 2011). https://www.bmj.com/content/342/bmj.c7452.

Heinze, Eric. "Hate Speech and the Normative Foundations of Regulation." *International Journal of Law in Context* 9, no. 4 (December 2013): 590–617.

———. *Hate Speech in Democratic Citizenship*. Oxford: Oxford University Press, 2016.

Holan, Angie. "PolitiFact's Lie of the Year: 'Death Panels.'" Politifact, December 18, 2009. https://www.politifact.com/truth-o-meter/article/2009/dec/18/politifact-lie-year-death-panels.

King, Martin Luther. "I Have a Dream." Speech presented at the March on Washington for Jobs and Freedom, Washington, DC, August 1968. https://americanrhetoric.com/speeches/mlkihaveadream.htm.

Landemore, Hélène. *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton: Princeton University Press, 2013.

———. "Yes, We Can (Make It Up on Volume)." *Critical Review* 26, nos. 1–2 (2014): 184–237.

Langton, Rae. "Blocking as Counterspeech." In *New Work on Speech Acts*, edit-

ed by Daniel Fogal, Daniel Harris, and Matt Moss, 144–62. Oxford: Oxford University Press, 2018.

———. "Speech Acts and Unspeakable Acts." *Philosophy and Public Affairs* 22, no. 4 (Autumn 1993): 293–330.

Langton, Rae, Sally Haslanger, and Luvell Anderson. "Language and Race." In *Routledge Companion to Philosophy of Language*, edited by Gillian Russell and Fara Delia Graff, 753–67. Abingdon: Routledge, 2012.

Lepoutre, Maxime. "Hate Speech in Public Discourse: A Pessimistic Defense of Counterspeech." *Social Theory and Practice* 43, no. 4 (October 2017): 851–83.

Leslie, Sarah-Jane, "The Original Sin of Cognition: Fear, Prejudice, and Generalization." *Journal of Philosophy* 114, no. 8 (August 2017): 393–421.

Lewandowsky, Stephan, Ulrich Ecker, Colleen Seifert, Norbert Schwarz, and John Cook. "Misinformation and Its Correction: Continued Influence and Successful Debiasing." *Psychological Science in the Public Interest* 13, no. 3 (December 2012): 106–31.

Lewis, David. "Scorekeeping in a Language Game." *Journal of Philosophical Logic* 8, no. 1 (January 1979): 339–59.

Maitra, Ishani. "Subordinating Speech." In Maitra and McGowan, *Speech and Harm*, 94–118.

Maitra, Ishani, and Mary Kate McGowan. "Introduction and Overview." In Maitra and McGowan, *Speech and Harm*, 1–23.

Maitra, Ishani and Mary Kate McGowan, eds. *Speech and Harm*. Oxford: Oxford University Press.

McAdams, Richard. *The Expressive Powers of Law.* Cambridge, MA: Harvard University Press, 2015.

McGowan, Mary Kate. "Debate: On Silencing and Sexual Refusal." *Journal of Political Philosophy* 17, no. 4 (2009): 487–94.

———. *Just Words: On Speech and Hidden Harm.* Oxford: Oxford University Press, 2019.

———. "Oppressive Speech." *Australasian Journal of Philosophy* 87, no. 3 (2009): 389–407.

———. "Responding to Harmful Speech." In *Voicing Dissent*, edited by Casey Rebecca Johnson, 182–200. Abingdon, UK: Routledge, 2018.

Mchangama, Jacob. Review of *The Harm in Hate Speech*, by Jeremy Waldron. *Policy Review* 176 (2012): 95–102.

Nelson, Sara. "Katie Hopkins Reflects on Branding Migrants 'Cockroaches' and 'Feral Humans.'" *Huffington Post*, July 27, 2015. https://www.huffingtonpost.co.uk/2015/07/27/katie-hopkins-reflects-migrants-cockroaches-feral-humans_n_7877124.html.

Nielsen, Laura Beth. "Power in Public." In Maitra and McGowan, *Speech and Harm*, 148–73.

Post, Robert. *Constitutional Domains: Democracy, Community, Management*. Cambridge, MA: Harvard University Press, 1995.

Reilly, Katie. "Here Are All the Times Donald Trump Insulted Mexico." *Time*, August 31, 2016. https://time.com/4473972/donald-trump-mexico-meeting-insult.

Richards, Robert, and Clay Calvert. "Counterspeech 2000: A New Look at the Old Remedy for 'Bad Speech.'" *Brigham Young University Law Review* 2000, no. 2 (2000): 553–86.

Simpson, Robert Mark. "Dignity, Harm, and Hate Speech." *Law and Philosophy* 32, no. 6 (November 2013): 701–28.

———. "Un-Ringing the Bell: McGowan on Oppressive Speech and the Asymmetric Pliability of Conversations." *Australasian Journal of Philosophy* 91, no. 3 (2013): 555–75.

Smith, Ben, and Byron Tau. "Birtherism: Where It All Began." Politico, April 24, 2011. https://www.politico.com/story/2011/04/birtherism-where-it-all-began-053563.

Sparrow, Andrew. "UK Statistics Chief Says Vote Leave £350m Figure Is Misleading." *The Guardian*, May 27, 2016. https://www.theguardian.com/politics/2016/may/27/uk-statistics-chief-vote-leave-350m-figure-misleading.

Stanley, Jason. *How Propaganda Works*. Princeton: Princeton University Press, 2015.

Strossen, Nadine. *Hate: Why We Should Resist It with Free Speech, Not Censorship*. Oxford: Oxford University Press, 2018.

Tirrell, Lynne. "Toxic Speech: Inoculations and Antidotes." *Southern Journal of Philosophy* 56, no. S1 (December 2018): 116–44.

Tversky, Amos, and Daniel Kahneman. "The Framing of Decisions and the Psychology of Choice." *Science* 211, no. 4481 (January 30, 1981): 453–58.

Waldron, Jeremy. *The Harm in Hate Speech*. Cambridge, MA: Harvard University Press, 2012.

Young, Iris Marion. *Inclusion and Democracy*. Oxford: Oxford University Press, 2000.