Counterspeech

Bianca Cepollaro, Maxime Lepoutre, and Robert Mark Simpson

Forthcoming in *Philosophy Compass*

Abstract. Counterspeech is communication that tries to counteract potential harm brought about by other speech. Theoretical interest in counterspeech partly derives from a libertarian ideal – as captured in the claim that the solution to bad speech is more speech – and partly from a recognition that well-meaning attempts to counteract harm through speech can easily misfire or backfire. Here we survey recent work on the question of what makes counterspeech effective at remedying or preventing harm, in those cases where it is effective, as well as work investigating when and why there is a duty to engage in counterspeech. We suggest that the most fruitful area for philosophical inquiry on this topic, currently, relates to the questions about efficacy. Specifically, we argue that there is a need for better frameworks for conceptualizing the efficacy of counterspeech. Philosophers have collaborative work to do, alongside social scientists, in developing these frameworks.

1. Introduction

Counterspeech has recently come into focus as a subject for philosophical analysis. In essence, counterspeech is communication that tries to counteract potential harm brought about by other speech, either face-to-face or remotely.[1] Suppose someone is spreading false information about the dangers of vaccines. You might try to rebut their claims, and thus discourage listeners from making bad health choices. Or maybe someone is using a slur to attack another person. You might try to jump in and challenge the derogatory force of this utterance. We will say more in §2, about how to draw the boundaries of this concept, but these basic

[1] Much recent work on this topic focuses on social media (e.g. Richardson-Self 2021, Saul 2021, Buerger 2021).

examples illustrate its core. Counterspeech is speech that aims to remedy potential harm brought about by other speech.

How has this concept taken shape? It initially came out of debates about free speech. Many liberals believe that in order to remedy harm done by discriminatory or deceptive speech, we simply don't need to silence people. Some *imminently* dangerous speech needs to be restricted. But for slower-acting harmful speech, we can generally mitigate the harm by speaking back in the right way. In short, as Justice Brandeis said, in the landmark US Supreme Court case *Whitney v. California* (1927) – a case concerning the government's right to quash dissent – "the solution to bad speech is more speech".

But interest in counterspeech is not exclusively driven by concerns about free speech. Some philosophers have investigated this topic not so much because they are wary of legal restrictions on speech, but because of counterspeech's positive potential as a tool for promoting social justice. This kind of inquiry often runs in parallel with work examining the mechanisms via which harmful speech operates. If we can diagnose how, for instance, presupposition – or ambiguity, or the subversion of illocutionary uptake, etc. – functions in harmful speech, then we can strategize about how to prevent the harm, using speech that thwarts the crucial linguistic mechanism.

The pivotal question, for both of these starting points, is what we will call the *Efficacy Question*: when and how is counterspeech actually an effective way to counteract harmful speech? Without a good answer to this, we have reason to resist the 'more speech' dictum, and to question social justice norms that enjoin people to speak back against harmful speech. Some of the nascent literature on this topic engages with the Efficacy Question, with a particular interest in how well-meaning counterspeech sometimes exacerbates the harms that it is trying to remedy. We survey this work in §3. Our main take-away is that more work needs to be done in conceptualizing counterspeech's efficacy. Empirical work that aims to assess counterspeech's efficacy (and its limits) needs sharper definitions of its success conditions, in order to be illuminating.

The other key question about counterspeech that philosophers have taken up, we call the *Deontic Question*. It is costly to engage in counterspeech, and it is unfair, *prima facie*, to ask those who endure the harms of harmful speech to also bear the costs of speaking back against it. The price for free speech may seem intolerable, if it is in fact only being paid by some. Hence the question arises as to how counterspeech-related duties should be allocated. Are duties generated by proximity? Or by expected efficacy? Or by some ideal of distributive fairness? We survey these issues in §4. We conclude in §5 by explaining what we see as the most fruitful lines of inquiry around this topic, arguing for the priority of the Efficacy Question in current philosophical work.

2. What is counterspeech?

2.1 Definition

Counterspeech is communication that seeks to counteract potential harm that is brought about by other speech. The literature generally focuses on verbal counterspeech, but there is no reason to exclude non-verbal expression from the concept's scope. Counterspeech may take the form of assertions, questions, imperatives, or platform-mediated acts of communication, like 'sharing' (Marsili 2021, Michaelson et al. 2021).

The harms that counterspeech works against come in various forms too. They include, for instance, physical harms (e.g. borne of false information about vaccine safety), status-related harms (e.g. stigma, loss of dignity, or other social inequalities resulting from expressions of negative affect or hatred), as well as diffuse harms to social and political institutions (e.g. the erosion of democratic norms) (Anderson et al. 2012, Waldron 2012, Bernecker et al 2021, Muirhead and Rosenblum 2020).

We think it makes most sense to characterize counterspeech in terms of a speaker's *intention* to counteract potential harm. 'Accidental counterspeech' is not inconceivable, and is an interesting phenomenon. But our main interest is when people should be trying to counteract speech-caused harm, via their own speech, and what it takes to succeed in this. One upshot of this definitional stipulation is that counterspeech is not a success term. It is defined by the speaker's intention to try to prevent speech-borne harm, not by its success at this. There is always a possibility, naturally, that a counterspeaker will misperceive some speech as harmful, e.g. because she misconstrues her interlocutor's meaning, or because she has a warped value system that interprets some benign states of affairs as harmful. We do not exclude such cases from our concept by definition. But the Efficacy and Deontic Questions are less relevant when we are dealing with, say, a misconstrued remark that is in fact harmless.

One worry about this concept, even with these initial stipulative restrictions, is that it may seem too capacious. If every bit of disputation or verbal disagreement qualifies as 'counterspeech', then the questions that we are raising might not seem distinctive enough, and moreover, they might seem too broad to admit of useful answers.[2]

[2] Some recent work (e.g. Johnson 2018) takes up something like this broader inquiry—i.e., what is the extent of the duty to voice disagreement?

But the concept of counterspeech that we are sketching need not collapse into the very broad idea of communicating disagreement. Sometimes you think another person is mistaken, in the views or values they are espousing, and you feel moved to voice that judgement. However, in a subset of cases – the cases that counterspeech is relevant to – you perceive that what the person is saying is not only mistaken, but also liable to cause some non-trivial harm to others. Our question is: in those harm-involving cases, when should you speak back to try to thwart the harm, and how can you do so effectively? And note that this speaking back might not even involve an articulation of disagreement; it may consist in one's changing the topic, or attacking the speaker, or subverting or 'bending' the target speech's meaning (Caponetto & Cepollaro ms).

2.2 Dimensions of variation

Counterspeech, thus defined, remains a diverse category. It can be performed by different kinds of speakers, who occupy different conversational, social, and institutional roles. Counterspeech can be performed by targets, i.e. the victims of the harmful speech, or by non-targeted addressees of the harmful utterance, or by mere bystanders. It can be performed by ordinary citizens, or by people in authoritative roles, including state actors acting on society's behalf (Brettschneider 2012, Gelber 2012, Lepoutre 2021, Saul 2021). Counterspeakers can speak out on their own, or as part of a coordinated group (Friess et al. 2020, Fumagalli 2021, Buerger 2021).

Counterspeech's audience is also variable, and this matters because it has an effect on its authority and efficacy. The audience may include the person whose speech is being challenged, victims, bystanders, or any combination of these. In many online contexts, counterspeakers cannot know in advance who their actual audience will be (Saul 2021).

Numerous 'tactical variations' are possible, when responding to harmful speech. One tactical consideration is whether to be reactive or proactive. Many analyses highlight how counterspeech works as a post hoc remedy for harmful speech. But it can potentially also be used pre-emptively, to inoculate audiences against future instances of harmful speech. In these cases, counterspeech helps counteract harm that would otherwise have arisen from future speech (Tirrell 2018, Lepoutre 2021). Another consideration is whether to adopt a more negative or positive style. An example of a negative style would be the fact-checking of harmful misinformation. By contrast, positively-styled counterspeech might involve highlighting true information, rather than rebutting falsehoods (Lepoutre 2021). Admittedly, certain instances of counterspeech lie in an ambiguous location, between the negative and positive (Armas and Ruiz 2021). For example, the reclamation of slurs is a tactic that aims to repudiate the derogatory usage of certain

terms, while simultaneously inaugurating a provocative, alternative, positive usage in its place (Brontsema 2004, Cepollaro and Zeman 2020).

Another tactical consideration is whether counterspeech targets explicit or implicit content. Explicit statements – e.g. "vaccines cause autism", "people in this group are vermin" –tend to elicit direct rebuttals (Ferkany 2021). But some harmful speech functions more indirectly. Remarks like "do your own vaccine research", or "even a woman could do that", convey potentially harmful content implicitly. For these cases, Langton (2018) recommends a form of counterspeech she calls *blocking*, which involves trying to identify and defuse implicit content (see Sbisà 1999), e.g. as conveyed in implicatures, dog-whistles, presuppositions, or other not-at-issue content. Moreover, just as harmful speech can harm via implicitly conveyed content, counterspeech can counteract harm via implicit communication. We tend to think of counterspeech, positive or negative, as a conversational move that openly challenges a problematic utterance. But it can take a more covert form, for example, by presupposing the falsehood of the speech it targets, rather than asserting it, e.g. "we owe the overwhelming safety of vaccines to this ground-breaking discovery" (Fraser ms.), or by surreptitiously changing what prejudiced speakers do with their words (Camp 2018, Caponetto and Cepollaro ms.).

3. The Efficacy Question

3.1 Backfiring, lack of authority, and other problems of efficacy

When and how is counterspeech effective at thwarting harmful speech? Part of what makes counterspeech an intriguing topic, is that we sometimes act like speaking magically undoes the effects of harmful speech. We feel compelled to speak back against harmful speech, as if speech had mysterious curative powers. And yet the disappointing reality is that speaking back often goes awry. Attempts to rebut misinformation can inadvertently encourage audiences to believe it (Nyhan and Reifler 2010). And attempts to communicatively counteract discriminatory speech can perpetuate stigma and exclusion (McGowan 2018), or provoke worse speech as a reaction.

These are not idle concerns. If counterspeech is just as effective as speech restrictions at mitigating harm, then *prima facie*, those restrictions, and the costs that they involve, seem unjustifiable – exactly as the 'more speech' dictum suggests (Lepoutre 2017). On the other hand, if counterspeech is generally ineffective – or worse, prone to backfiring, and amplifying the harms it aims to mitigate – then we should reject the 'more speech' dictum, and focus our efforts on developing suitably-targeted restrictions on harmful speech (Waldron 2012, McGowan 2018).

Are these concerns well-founded? Some recent empirical evidence suggests that the risk of backfiring is often exaggerated (Wood and Porter 2020). But this evidence only pertains to some forms of counterspeech (fact-checking), directed at some forms of harmful speech (misinformation), in some contexts (experimental settings). At present, the empirical data remain insufficient to settle the question.

Given the openness of the question, there are *prima facie* reasons, grounded in philosophy of language, linguistic pragmatics, and cognitive science, to worry that counterspeech (or at least some categories of it) is less effective than we might hope. Worries about backfiring are often explained in terms of *salience*. Some authors claim that counterspeech can increase the salience of the speech it is addressing, or specific pernicious features of that speech (Simpson 2013, McGowan 2018, Saul 2021, Maitra ms.). Different mechanisms can link unintended salience to unwelcome outcomes. Some cognitive science focuses on *fluency*. The more salient some idea is, the more fluent or familiar it becomes, and in turn, the likelier it is to be believed (Lewandowsky et al. 2012). Another mechanism relates to pragmatic conversational norms. The salience of an idea in a conversation can imply that this proposition is credible enough to merit attention. Countering misinformation or hate speech may, by reinforcing their salience, inadvertently give them credibility (Levy 2019). Another mechanism is distraction. Attention is often a scarce resource, and some speech perpetrates harm by distracting us from pressing emergencies, e.g. baseless political slanders, which subvert democratic processes (Williams 2018). Responding to these utterances can make them more salient, and thus amplify, or at least sustain, their distracting-ness.

Even when counterspeech doesn't exacerbate the harms that it aims to prevent, it can easily misfire, or fail to mitigate these harms (Brown 2018; Tirrell 2019). For one thing, many counterspeakers don't have the authority they need to achieve their aims (Langton 2018). For example, fact-checkers are unlikely to dispel false beliefs if they aren't seen as trustworthy by the relevant audiences (Jerit and Zhao 2020). Or in countering discriminatory speech, individual citizens may lack the authority held by the state, to reassure targeted citizens of their civic standing (Gelber 2012; Brettschneider 2012). Granted, this determinant of efficacy is generally hard to evaluate, because different speakers have authority in different domains (Langton 2018). Moreover, in highly polarized societies, it may be that no-one possesses *de facto* discursive authority for the whole society (Hameleers and van de Meer 2019).

There are other factors, besides these, that may cause counterspeech to misfire in specific instances, or in response to specific kinds of harms. As noted in §2.2, some types of harmful speech, like dog-whistles, are hard to counteract because they convey harmful content covertly, via presuppositional or not-at-issue content (Langton 2012, Stanley 2015, McGowan 2020). And some harmful speech comes equipped with counterspeech-defense mechanisms. Most notably, many conspiracy theories anticipatorily rebut counterspeech challenges, thereby making

themselves "self-sealing" (Cassam 2019). More prosaically, some harmful speech happens in communicative spaces that are simply hard to access and engage with (Sunstein and Vermeule 2009).

3.2 Conceptualizing and assessing efficacy

If we are going to partly rely on counterspeech, as a tool for mitigating the effects of harmful speech, we should be seeking to make our counterspeech effective and reliable to this end. Several things are involved in this. First, we need a systematic overview of what factors (linguistic, cognitive, etc.) impair its efficacy, and how. Second, to determine whether some forms of counterspeech avoid these obstacles, we need a taxonomy outlining what forms of counterspeech there are. As discussed in §2.2 and §3.1, recent philosophical work has made initial in-roads on both of these issues. But we also need something else, which has so far received less attention, namely, a framework for conceptualizing and assessing counterspeech's efficacy. In order to evaluate which forms of counterspeech are most likely to be efficacious, and when, we first need a clearer explanation of what efficacy consists in.

One initial distinction, which is useful in conceptualizing counterspeech's efficacy, is between (1) harms borne of false information, and (2) harms borne of negative affect. Suppose someone makes a public statement that all Muslims support terrorism. You could rebut this by presenting data showing that few Muslims support terrorism, and this may prevent some people from forming a false belief. But it may fail, nevertheless, to undo the stigmatizing effect of that comment – an effect that registers in people's emotions, more than their beliefs, and which can thus persist even once the falsehood is definitively corrected. It may be that a joke at the speaker's expense, or an expression of solidarity with Muslims, would do more than fact-checking to undo the affectively-mediated stigma. Conversely, suppose someone says dieting is futile, and that dieters never achieve any real health benefits. Although this remark may cause negative affect among some people who hear it, what we should mainly worry about, in relation to it, is that people don't acquire false beliefs about the futility of healthy eating, which may lead them to refrain from acting in aid of their own wellbeing. In this case, plausibly, counterspeech should try to address the false piece of information, rather than try to remedy any hurt feelings that the comment might have brought about.

One challenge, then, in developing a good framework for assessing counterspeech's efficacy, is to establish when – in which cases, in connection with which negative outcomes – the efficacy of counterspeech should be judged in terms of *epistemic* results, and when in terms of *affective* results, or other results of a more conative than cognitive nature. This sorting exercise is relatively simple in the above cases, but there are harder cases too. Plausibly, the inefficacy of some counterspeech is due to the speaker not having a clear sense of (i) whether they are

purporting to change people's cognitive or conative attitudes, or (ii) which techniques are likely to be effective either way.

Complications remain even if the counterspeaker is confident about what type of attitudinal change she is trying to achieve. Take the dieting case again, and suppose that, due to someone's misinformative speech, ten people form the false belief that dieting has no health benefits. Now consider two possible responses to this. The first involves a compelling attack on the speaker's credibility, which persuades nine people to drop the false belief, albeit with no real understanding about why it is false. The second involves data refuting the false belief, which persuades five people to abandon it, based on some understanding of evidence of its falsity. Which response is preferable? The answer to this may partly depend on empirics, e.g. which outcome results in a greater overall reduction of harmful eating behavior. But it may also partly hinge on judgements about epistemic axiology, i.e. about whether understanding or 'mere' true belief is a preferable epistemic outcome (see, e.g., Elgin 2017, for discussion).

Or consider the anti-Muslim case again. Suppose that a bigoted statement makes the Muslims in the group feel marginalized, while emboldening other people in their semi- suppressed anti-Muslim attitudes. Now imagine a series of different counterspeech responses: one aimed at undoing the affective hurt to the targets, another aimed at softening the bigoted feelings of those in the anti-Muslim camp, and another that simply tries to shift attention away from the stigmatized group. Which response is best?

To answer this we need some account of what the harm of hate speech consists in, and this is partly a philosophical question – a question about the rights of civic membership. On one influential account, the harm of hate speech – that which primarily justifies its legal restriction – is that it undermines its targets' assurance of their social status (Waldron 2012). Given this account, the efficacy of counterspeech in response to hate speech is principally about its ability to provide reassurance (Tirrell 2018: 138, Lepoutre 2017: 864). But is this the right account? And what affective interventions would it make sense to aim at, instead, given other plausible accounts of the harms of hate speech, or of others kinds of discriminatory speech, like slurs or dog-whistles?

The final challenge is to operationalize the philosophical account of counterspeech's efficacy. The epistemically and affectively mediated harms that counterspeech aims to mitigate cannot always be straightforwardly observed. To provide guidance for empirical research, a further task is to identify measurable proxies for these harms. For instance, instead of directly observing whether targets of hate speech have felt reassured, empirical studies tend to consider whether counterspeech has reduced the frequency or salience of hate speech (see Buerger 2021). In identifying such proxies, we must ensure that our measurements remain connected to our underlying interests. Proxies must track something that is really

linked to the relevant harms, and the relation should not be posited based on pure conjecture (Fumagalli 2022).

In sum, philosophical analysis has several roles to play in assessing the efficacy of counterspeech. It is needed to conceptualize, and help operationalize, what counts as efficacy (§3.2); to theorize the mechanisms that stand in the way of such efficacy (§3.1); to taxonomize the different possible forms counterspeech can take (§2.2); and to generate credible hypotheses, informed by these three functions, regarding which forms of counterspeech are most likely to be efficacious, and in what contexts.

4. The deontic question

Suppose we are in a specific situation where we have good reason to believe that counterspeech will be effective in mitigating some harm. This still leaves a crucial normative question unresolved. Who should bear the costs of engaging in counterspeech? One answer is: everyone, provided they are reasonably able to do so, and the costs or risks are not too high for them. Granted, such duties may be loosened or suspended, given extenuating circumstances. Nonetheless, general circumstantial duties of assistance or basic beneficence may give rise to some kind of universal duty to engage in counterspeech (Goldberg 2020, Howard 2021). If you are on the scene when one person is hurting another person, you should help, provided that you're in a position to do so safely and effectively. Similarly, if you are on the scene when someone is saying harmful things, you should speak up to mitigate the harm.

The idea that everyone has counterspeech duties derives further support from norms of citizenship. The liberal citizen arguably has a civic responsibility to engage and challenge unreasonable conceptions of justice (Clayton and Stevens 2014), and there may be neutrality-based reasons why governments cannot perform this type of engagement (Badano and Nuti 2018). These civic duties become more pressing when unreasonable viewpoints are being espoused in concretely harmful forms of expression. Universal counterspeech duties may also derive support from a 'silence as violence' principle. In some situations, not confronting harmful speech can function as a tacit support for it, thus contributing to the harms that it brings about (Ayala and Vasilyeva 2016).

In contrast to a universal duties view, you may think that the costs of counterspeech should primarily fall on institutions, starting with the state (Brettschneider 2012, Gelber 2012, Lepoutre 2017) and encompassing other organizations (Saul 2021). It is often a matter of public interest that the harms of discriminatory and deceptive speech are mitigated. When those harms befall specific individuals (e.g. through libel, or targeted harassment) criminal law is used to restrict harmful speech. But even when the harms are more diffuse, the state may still be better-equipped than private actors, and more possessed of the requisite authority, to

discursively confront them. An adjacent approach would say that the state must provide support to private actors to engage in counterspeech. For example, it may provide funding to representatives of oppressed groups, to help them build a media profile, or to publish materials aimed at countering derogatory views. Or the state may fund anti-misinformation initiatives through university centres and think-tanks. This approach may potentially give us the best of both worlds. Speakers with situation-specific knowledge can respond to harmful speech 'at the source', but they aren't unfairly burdened in this, since their efforts are sponsored and endorsed by the state (see e.g. Malik 2011; Gelber 2012).

Another view is that the costs of counterspeech should fall on actors who have special discursive influence. Sometimes the most effective counterspeaker won't be the state, but a private individual, like a popular celebrity, or the leader of a respected organisation. Gelber and Bowman (2021) defend a version of this view, arguing that university leaders have a duty to address harmful speech in the university setting.

We can identify cases in which each of the approaches seems most likely to be effective. But do our ethical judgements here adhere to any general ethical principles?

Let's start with a basic consequentialist norm: we should counteract as much harm as possible,[3] and so the costs of counterspeech should be borne by whomever is best-placed to minimize harm. As noted in §1, however, it may seem unreasonable to require those who are harmed by speech to speak out against this harmful speech.[4] Indeed, to do so puts targets at a kind of double disadvantage, relative to non-targets: they must give harmful speech their attention, thus exposing themselves more to its harmful effects; and they must expend effort responding to it and trying to mitigate its impact on others. If this seems unfair, then plausibly there needs to be some kind of fairness-based side-constraint on any consequentialist principle in this area (Schauer 1992, Delgado and Yun 1995, Nielsen 2012, Maitra and McGowan 2012).

Alternatively, the side constraint on a consequentialist ethic may not be about fairness, but simply a concern over the magnitude of the burdens that specific individuals are forced to endure. Imagine you are the only vaccine-friendly person in a community of zealous anti-vaccination sceptics. The most effective way of rebutting vaccine misinformation in your community may be for you, as a local

[3] The phrase 'as much harm as possible' is agnostic between different ways of counting or aggregating harm. Thus, if some kinds of harms are more serious than others, this consequentialist norm is compatible with prioritizing them.

[4] Whether it is unreasonable, all things considered, will depend partly on the broader ethical question of whether people have a duty to resist their own oppression. Boxill (1976), for instance, argues that protesting the wrongs one is subjected to is necessary to preserve one's dignity or self-respect, and that, as a result, there is a duty to do so (see Vasanthakumar 2020 for an overview).

person with some ability to 'reach across the aisle', to have regular discussions with your neighbors. But the impact of this may be overwhelming, as you absorb repeat doses of other people's ire and disdain. In online contexts, heavy individual costs can be incurred in unpredictable ways; for instance, one may become the transient target of an online mob (Aly and Simpson 2019, Saul 2021).

In response to this worry, those who posit a universal civic duty to engage in counterspeech (e.g. Howard 2021) can argue that this duty is counterbalanced by each person's right to not compromise their basic interests. The demandingness of these duties will depend on what level of risk or burden we think each speaker is obliged to withstand, but no-one is obliged to martyr themselves. Moreover, when the costs of counterspeech are overwhelming, there may be alternative, lower-risk interventions that speakers can make, to counteract harmful speech (e.g., Fumagalli 2021). Even if the risks of grave personal costs are genuine, each of us might still have a duty to do something.

5. Future directions

Counterspeech admits of investigation from a range of disciplinary angles. In thinking about how to define counterspeech, for instance, we may derive insights from looking at how linguists and anthropologists theoretically define various communicative practices (see Buerger 2020). And the Efficacy Question in particular is one whose answer clearly calls for social scientific methods of inquiry. Philosophical thinking can enter these inquiries at multiple points, as we have noted. But which are the most fertile fields for philosophers to plough, in working on this topic?

In principle, the Deontic Question seems like a natural place to concentrate. Philosophers have disciplinary expertise in thinking through ethical principles, and their application to concrete problems. When combined with our best estimates of when and why counterspeech is effective at remedying harm, this expertise could be used to offer guidance about who ought to engage in counterspeech, and in which situations.

But answers to the Deontic Question are hostage to the Efficacy Question, and thus we believe the latter is for now a better focal point for philosophical inquiry. This is partly for ought-implies-can reasons. As Howard says, "if counterspeech is overwhelmingly likely to be ineffective, what is the justification for obliging agents to engage in it?" (2021: 934; see also Saul 2021: 146-50). Moreover, even if counterspeech *is* effective at undoing harm, the answer to the (deontic) question of who should engage in it, still depends to a significant degree on further (efficacy) questions, about which techniques, actors, and formats, are most likely to succeed in harm-prevention.

Consider counterspeech aimed at remedying harms borne of advocacy of vaccine scepticism. First, there is a question of whether successful counterspeech requires scientific expertise, or some other type of relevant authority. Next, there is a question of whether it is more effective (in terms of eventual harm-mitigation) to try to change the minds of sceptics, or to shore up the beliefs of agnostic onlookers. Next, there is a question of whether certain discursive techniques are required for successful counterspeech – conflict de-escalation, charm, humour, argumentative alacrity, etc. It may be unhelpful to have a norm which says everyone should engage in counterspeech, if in fact only certain speakers, with certain skills, are likely to do so effectively. And a norm that says the state should engage in counterspeech may ring hollow until we have some credible template of how that can be done effectively. A normative constraint which says it is unfair, or too costly, to ask targeted individuals to engage in counterspeech, may seem premature in its conclusions, if it turns out that those speakers are the ones most likely to remedy the harms.

Using speech to undo epistemically- and affectively-mediated harms is something that in principle anyone can do. Part of the prevalent progressive ethos in our polarized age, is that we should do this. We should speak up to contest harmful falsehoods, and discriminatory sentiments. But actually remedying these harms is both an art and science. Counterspeech is of little value if it is primarily an exercise in virtue signalling. If we actually want to figure out what is needed to undo verbally-mediated harm, we need a well-organized social scientific research program, which is attuned to the dimensions of variation in how those harms are effected. Philosophical insights from pragmatics, and from the intersection of epistemology and value theory, can help that research program succeed. Philosophical work on counterspeech, for now, has useful work to do, then, in being a handmaiden to this emerging social science.

References

Aly, Waleed, and Robert Mark Simpson. "Political Correctness Gone Viral." In *Media Ethics, Free Speech, and the Requirements of Democracy* , edited by Carl Fox and Joe Saunders. New York: Routledge, 2019.

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

Armas, Álvaro, and Andreas Ruiz. "Provocative Insinuations." *Daimon Revista Internacional de Filosofia* 84 (2021): 63-80.

Ayala, Saray, and Nadya Vasilyeva. "Responsibility for Silence." *Journal of Social Philosophy* 47, no. 3 (2016): 256-72.

Badano, Gabriele, and Alasia Nuti. "Under Pressure: Political Liberalism, the Rise of Unreasonableness, and the Complexity of Containment." *Journal of Political Philosophy* 26, no. 2 (2017): 145-68.

Bernecker, Sven, Amy Floweree, Thomas Grundmann (eds). *The Epistemology of Fake News*. Oxford: Oxford University Press, 2021.

Boxill, Bernard. "Self-Respect and Protest." *Philosophy & Public Affairs* 6 (1976): 58-69.

Brandeis, Louis. "Opinion in *Whitney v California*, 274 US 357," 1927.

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

Brontsema, Robin. "A Queer Revolution: Reconceptualizing the Debate over Linguistic Reclamation." *Colorado Research in Linguistics* 17 (2004): 1-17.

Brown, Etienne. "Propaganda, Misinformation, and the Epistemic Value of Democracy." *Critical Review* 30 (2018): 194-218.

Buerger, Catherine. "#iamhere: Collective Counterspeech and the Quest to Improve Online Discourse." *Social Media + Society* (2021): 1-27.

Camp, Elisabeth. "Insinuation, Common Ground." In *New Work on Speech Acts*, edited by Daniel Fogal, Daniel Harris, and Matt Moss. Oxford: Oxford University Press, 2018.

Caponetto, Laura and Cepollaro, Bianca (Manuscript.), *Bending as Counterspeech*.

Cassam, Quassim. *Conspiracy Theories*. Cambridge: Polity Press, 2019.

Cepollaro, Bianca, and Dan Zeman (eds). *Special Issue: Non-Derogatory Uses of Slurs—Grazer Philosophische Studien* 97 (2020).

Clayton, Matthew, and David Stevens. "When God Commands Disobedience: Political Liberalism and Unreasonable Religions." *Res Publica* 20 (2014): 65-84.

Delgado, Richard, and David Yun. "Pressure Valves and Bloodied Chickens: An Analysis of Paternalistic Objections to Hate Speech Regulation." *California Law Review* 82 (1994): 871-92.

Elgin, Catherine. *True Enough*. Cambridge, MA: MIT Press, 2017.

Ferkany, Matt. "How and Why We Should Argue with Angry Uncle: A Defense of Fact Dumping and Consistency Checking." *Social Epistemology* 35, no. 5 (2021): 533-45.

Fraser, Rachel. "How to Talk Back." (Manuscript).

Friess, Dennis, Marc Ziegele, Dominique Heinbach. "Collective Civic Moderation for Deliberation?" *Political Communication* 38, no. 5 (2021): 624-46.

Fumagalli, Corrado. "Counterspeech and Ordinary Citizens: How? When?" *Political Theory* 49, no. 6 (2021): 1021-47.

_____ "Can Rawlsian Containment of Hateful Viewpoints Be Effective." *Social Theory and Practice* (2022): 1-27.

Gelber, Katharine. "'Speaking Back': The Likely Fate of Hate Speech Policy in the United States and Australia." In Maitra and McGowan, Speech and Harm, 50-71.

_____ "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 Kristine Bowman. "Responding to Hate Speech: Counterspeech and the University." *Virginia Journal of Social Policy & the Law* (2021): 248-274.

Goldberg, Sandy. *Conversational Pressure*. Oxford: Oxford University Press, 2020.

Howard, Jeffrey. "Terror, hate and the demands of counter-speech." *British Journal of Political Science* 51, no. 3 (2021): 924-939.

Jerit, Jennifer, and Yangzi Zhao. "Political Misinformation." *Annual Review of Political Science* 23 (2020): 77-94.

Hameleers, Michael, and Toni van der Meer. "Misinformation and Polarization in a High-Choice Media Environment: How Effective Are Political Fact-Checkers." *Communication Research* 47, no. 2 (2019): 227-50.

Johnson, Casey. "For the Sake of Argument: The Nature and Extent of Our Obligation to Voice Disagreement." In *Voicing Dissent*, edited by Casey Rebecca Johnson, 97-108. Abingdon: Routledge, 2018.

Langton, Rae. "Blocking as Counter-speech." In *New Work on Speech Acts*, edited by Daniel Fogal, Daniel Harris, and Matt Moss, 144–62. Oxford: Oxford University Press, 2018.

_____ 'The Authority of Hate Speech'. In *Oxford Studies in Philosophy of Law, Vol. 3*, edited by John Gardner, Leslie Green, and Brian Leiter, 123–52. Oxford: Oxford University Press, 2018.

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

_____ *Democratic Speech in Divided Times*. Oxford: Oxford University Press, 2021.

Levy, Neil. "No-Platforming and Higher-Order Evidence, or Anti-Anti-No Platforming," *Journal of the American Philosophical Association* 5, no. 4 (2019): 487-502.

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 (2012): 106–31.

Maitra, Ishani, and Mary Kate McGowan. "Introduction and Overview." In *Speech and Harm*, edited by Ishani Maitra and Mary Kate McGowan, 1-23. Oxford: Oxford University Press, 2012.

Maitra (ms), *Unsettling Speech*.

Malik, Maleiha. "Religious Freedom, Free Speech and Equality: Conflict or Cohesion?," *Res Publica* 17, no. 1 (2011): 21-40.

Mallett, Robyn, and Margo Monteith (eds.) *Confronting Prejudice and Discrimination*. London: Academic Press, 2019.

Marsili, Neri. "Retweeting: its linguistic and epistemic value." *Synthese* 198, no. 11 (2021): 10457-10483.

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

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

Michaelson, Eliot, Jessica Pepp, Rachel Sterken. "Online Communication." *The Philosophers' Magazine* 94, no. 3 (2021): 90-95.

Muirhead, Russ, and Nancy Rosenblum. *A Lot of People Are Saying*. Princeton: Princeton University Press, 2020.

Nielsen, Laura Beth. "Power in Public." In *Speech and Harm*, edited by Ishani Maitra and Mary Kate McGowan, 148-73. Oxford: Oxford University Press, 2012.

Nyhan, Reifler, and Jason Reifler. 'When Corrections Fail'. *Political Behavior* 32 (2010): 303–30.

Parker, Laura, Margo Monteith, Corinne Moss-Racusin, Amanda Van Camp. "Promoting concern about gender bias with evidence-based confrontation." *Journal of Experimental Social Psychology* 74 (2018): 8-23.

Richardson-Self, Louise. *Hate Speech Against Women Online: Concepts and Countermeasures*. Maryland: Rowman & Littlefield, 2021.

Saul, Jennifer. 'Someone Is Wrong on the Internet: Is There an Obligation to Correct False and Oppressive Speech on Social Media?' In *The Epistemology of Deceit in a Postdigital Era*, edited by Alison MacKenzie, Jennifer Rose, and Ibrar Bhatt, 139–57. New York, NY: Springer, 2021.

Schauer, Frederick. "Uncoupling Free Speech," Columbia Law Review 92, no. 6 (1992): 1321-57.

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

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

Sunstein, Cass, and Adrian Vermeule. 'Conspiracy Theories: Causes and Cures'. *Journal of Political Philosophy* 17, no. 2 (2009): 202–27.

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

_ "Toxic misogyny and the limits of counterspeech." *Fordham Law Review* 87 (2019): 2433–2452.

Vasanthakumar, Ashwini. "Recent debates on victims' duties to resist their oppression." *Philosophy Compass* 15, no. 2 (2020): 1-8.

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

Williams, James. *Stand Out of Our Light*. Cambridge: Cambridge University Press, 2018.

Wood, Thomas, and Ethan Porter. 'The Elusive Backfire Effect'. *Political Behavior* 41 (2020): 135–63.