2

THE PHILOSOPHY OF COUNTER LANGUAGE

Laura Caponetto and Bianca Cepollaro

Introduction

Speech can be "toxic", as philosopher Lynne Tirrell, among others, has recently emphasized (Tirrell 2017, 2018). Tirrell provided a broad characterization of *toxic speech* as speech that "diminishes democratic participation, undermines civil liberties, compromises the rule of law, and damages human dignity" (Tirrell 2018, 120). We will here use the term to refer to speech that spreads prejudicial stereotypes and/or endorses discriminatory practices (such as sexism, racism, homophobia, transphobia, ableism, etc.). The epidemiological metaphor aptly suggests that toxic speech can operate implicitly, rather than explicitly. Just as toxins silently stockpile, eroding the body little by little, so, too, toxic utterances can cumulatively disrupt the social fabric by propagating discriminatory attitudes in surreptitious ways. Toxic speech, so construed, includes both blatantly hateful utterances and subtler forms of discriminatory discourse and thus forms a broader category than *hate speech*.[1]

Scholars have pointed out that toxic speech has the potential to shape our epistemic and normative landscapes, by changing what we believe and accept, as well as what is permissible in a given context (see, e.g., McGowan 2009, 2019; Langton 2012, 2018; Caponetto and Cepollaro 2021). The claim is supported by psychological evidence showing that derogatory language, and speech toxicity more broadly, reduces the well-being of and increases suicide rates among targeted individuals (Swim et al. 2001; Mullen and Smyth 2004; Leader, Mullen, and Rice 2009), while prompting implicit negative evaluations and dehumanizing attitudes toward them (Carnaghi and Maass 2007; Fasoli, Maas, and Carnaghi 2015; Fasoli et al. 2016; Soral, Bilewicz, and Winiewski 2018). Against this backdrop, the question of how to resist becomes particularly pressing. If toxic speech can have harmful effects on individuals and the social

DOI: 10.4324/9781003377078-4
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

world, then devising strategies to counter it is of paramount importance. This is where the debate on counterspeech starts out.

"Counterspeech" is a term of art introduced in legal theory to pick out a family of measures that are alternative to censorship. As Justice Brandeis famously put it in *Whitney v. California*, the remedy to toxic speech (or "evil" speech, as he called it) should be "more speech, not enforced silence" (Brandeis 1927). But what forms should "more speech" (or counterspeech) take? Philosophy of language has developed tools and drawn distinctions that can shed light on the forms that toxic speech can take and the most suitable ways to tackle them. As we shall see, engineering counterspeech to suit the communicative features of the toxic utterance it responds to may increase its chances to hit the mark.

This chapter provides an opinionated survey of a number of counterspeech strategies that have been variously discussed in contemporary philosophy of language.[2] We will point out that certain strategies are particularly apt to counter toxic contents that are conveyed *implicitly*, whereas others have their best shot with toxic contents that are *explicitly* stated. We will also suggest that the appropriateness and expected outcome of a given strategy importantly vary with the context and that the counterspeaker's role (e.g., their belonging to the targeted group or not; their speaking as a private citizen or a government representative) is one of the major contextual variables.[3] Overall, our goal is to uncover how tools from philosophy of language can illuminate the workings of toxic speech and help us devise strategies to counter it. In this pursuit, we will sketch the foundations of a philosophy of counter-language.

Counterspeech Strategies

Denying

The most intuitive strategy to counter toxic speech through more speech is to reject it as false, possibly by providing reasons or evidence against it. To see the strategy at work, suppose that John, Paul, and Arthur are chitchatting, when they see Sarah, an acquaintance of theirs, in a large SUV across the street. John says,

1 Wow! That's huge. No way she can park it. Women just can't drive.

John's utterance is clearly toxic: it contributes to spreading, and openly endorses, the idea that "women can't drive" – a sexist stereotype that is part and parcel of a system of representations, meanings, and attitudes casting women as unequal to men. Now suppose that Arthur replies,

2 Oh, come on. That's not true,

and goes on by offering experiential evidence, and even research data, showing that women are generally just as good at driving as men. Arthur here engages in

what we call "denying": he issues a direct rebuttal to John's toxic claim, which is deemed false and is rejected on the basis of contrasting evidence.

Direct rebuttals like this perfectly fit the more speech model emerging from Brandeis's words. Here's the quote once again, this time in full:

> If there be time to expose through discussion the falsehoods and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence.
>
> *(Brandeis 1927)*

Denying is the prototypical way of "exposing through discussion the falsehoods and fallacies" propagated by toxic speech. It is aimed at stopping their spread, and, provided the counterspeaker disposes of counterevidence and has good argumentative skills, it may indeed force the toxic speaker to concede that they were wrong and prevent other people in the audience from endorsing their views about the targeted group. This is what our counterspeech strategies should ideally aim for.

This strategy, however, has several limits, the first of which is that it is markedly *confrontational*: in directly rebutting what one's interlocutor has said, one takes an adversarial stance toward them. Sometimes, this is exactly what one should do. Imagine that, during a presidential debate, one of the candidates states that a woman's place is in the home or that black people are violent. It may be not only appropriate but indeed imperative for the other candidate to openly confront them by forcefully denying their statements. Other times, however, barefaced confrontation may not be the best way to go, all things considered. One may have too few chances to make the toxic speaker drop their claim or convince the audience of the falsity of certain toxic views – and face too high a risk of backlash or retaliation. Denying may be dangerous ground, and particularly so for *targets*: a woman who directly rebuts a sexist statement may face a higher risk of being interpreted as overly sensitive, humorless, or a troublemaker than a man who does so; a black man who openly confronts a racist speaker may risk incurring in particularly harsh forms of retaliation, including physical violence (see, e.g., Rasinski and Czopp 2010; Dickter, Kittel, and Gyurovski 2012). Denying may also be dangerous ground when the counterspeaker, independent of their group membership, is subordinate to the toxic speaker – say, because the toxic speaker is their boss, their manager, or their teacher.

Denying aims at falsifying or disproving the toxic utterance it replies to. But often what is problematic with toxic speech has nothing to do with its content being false (Langton 2018; McGowan 2018). When slurs are hurled as epithets ("You S") or used in statements aimed at informing the audience of a (supposed) fact ("That S just moved here"), our concern is not primarily with the contribution they give to the utterance's truth value. Slurs ascribe an inferior status to certain groups of people, function as social mechanisms to push them back "in their place", and undermine their sense of dignity and assurance of

equal standing (Waldron 2012). This – not falsity – is what concerns us the most. Replying, "That's not true" in an attempt to counter a slurring utterance would just be missing the point.

Denying may fail as counterspeech even when falsity *is* an important part of the picture. A mere denial is a statement that what one's interlocutor has explicitly said is not true. As such, it does not address what was implicitly conveyed by the interlocutor's utterance. When falsehoods are not asserted, but merely implicated or presupposed, denying is not sufficient to rebut them.

Consider a revised version of our "women-just-can't-drive" example. As in the original version, John, Paul, and Arthur are chitchatting, when they see Sarah in a large SUV. But this time, John says,

3 Wow! That's huge. No doubt she'll have her husband park it for her.

John does not explicitly say anything sexist, and yet his utterance implicitly conveys the same sexist content as (1), namely,

4 Women are worse at driving than men.

To retrieve this content, one may reason as follows:

i John is saying that, undoubtedly, Sarah will have her husband park her SUV for her.
ii John may be saying so because he thinks that Sarah is worse at driving than her husband.
iii The context provides no elements to infer Sarah's or her husband's actual driving skills.
iv Sarah is a woman and her husband is a man.
v Women are stereotyped as worse drivers than men.
vi John must think and be conveying that Sarah is worse at driving than her husband because of her gender; that is *that women are worse at driving than men*.

Clearly, it is very unlikely that anyone would consciously go through steps (i) to (vi). Typically, we grasp implicit meanings quickly and unreflectively. The above steps offer a rational reconstruction of how one could retrieve (4) – a "conversational implicature", in technical jargon (Grice 1975) – from (3).

There is widespread agreement in philosophy of language that implicitly communicated content tends to enter the conversational common ground by default unless somebody objects. "Common ground" roughly stands for the set of assumptions that participants mutually share for the purposes of the conversation (Stalnaker 2002).[4] If uncontested, (3) will sneak the assumption *that women are worse at driving than men* into the conversational common ground. This does not mean that every participant will necessarily end up believing that

women are worse drivers than men. But from then on, the conversation will proceed under the assumption that it is indeed so: (4) will be accepted as true at least for the sake of the conversation.

As a result, the norms in force in that conversation will shift: certain subsequent moves will become appropriate (or "permissible"), whereas others will become inappropriate (or "impermissible"). Making fun of, or joking about, women drivers, for example, will become more contextually appropriate than it was before. Even more than that, John's move may *encourage* the others to follow suit and play the "gender stereotypes game" as a way of bonding with one another. Conversely, behaviors clashing with what is now common ground will become inappropriate or be discouraged. Suppose that, before seeing Sarah, John, Paul, and Arthur were talking about John's son and his upcoming driving test. Suppose Paul was about to tell the others how he got his driver license thanks to his girlfriend who taught him how to drive. Once (3) is uttered, and if no one objects, it becomes less easy or is no longer appropriate for him to tell that story.

Implicit content tends to get automatically incorporated into the common ground *but only insofar as nobody objects*. Hearers hold in their hands the power to block the process (Langton 2018). Not any objection will do, though. Imagine that, faced with (3), Arthur replies,

5 I don't think so.

Even though (5) denies what John explicitly asserts ("No doubt Sarah will have her husband park her SUV for her"), it lets the implicit content *that women are worse at driving than men* pass. A mere denial is indeed compatible with that content, as proven by the fact that (5) could be fleshed out as

(5*) I don't think so. Sarah is an excellent driver for being a woman!

– which would clearly support, rather than contest, the generic assumption that women are not as good as men at driving.

To counter implicit contents, one should go beyond mere denials and design one's objection in more sophisticated ways. One such way is what Rae Langton (2018) labeled "blocking".

Blocking

Paradigmatic blocking is a two-step procedure. The blocker *explicitates* and then *denies* the content implicitly conveyed by their interlocutor. "Explicitation" is a term of art introduced by Marina Sbisà (1999) to name the explicit exposure of implicit content. To counter (3), for example, Arthur might say,

6 Are you assuming she can't park it because she's a woman? That's ridiculous.

Once (6) brings to the surface (or explicitates) the implicit content of (3), it can be easily targeted for denial.

Even though Langton mainly focused on *explicitation-plus-denial*, she acknowledged that blocking is defined by its function:

> "Blocking" is a label for a hearer's resistance to what a speaker, or a speech act, presupposes: "Wait a minute –" says the hearer, or "Whadd ya mean – *even* George could win?" […] Blocking interferes with the evolving information taken for granted among participants in a conversation.
>
> *(Langton 2018, 148)*

This suggests that *any* contribution that prevents some implicit content from automatically becoming common ground will count as a blocking maneuver. So, although *explicitation-plus-denial* has been taken to constitute its paradigmatic form, blocking can come in many guises (Cepollaro n.d.). It may consist of explicitation only, for example. A reply like

(6*) Are you assuming she can't park it because she's a woman?

is a blocking maneuver, in that it prevents the common ground from being automatically updated with (4) ("Women are worse at driving than men"). If John still wants (4) to become a shared assumption, he will have to argue for it out in the open.

Sometimes, if one wants to block, it may be enough to stress *that* a certain utterance carries, or may be read as carrying, some implicit content – rather than fully articulating *what* that content is. And, indeed,

7 What are you implying?!

may be just as effective as (6) or (6*) in countering (3).

Blocking is a success term: you cannot block without accomplishing the definitional function of blocking. Turned on its head, this says that, when it comes to blocking, *success* consists in preventing certain implicit contents from entering the common ground by default. This is compatible with such contents eventually making it to the common ground. Suppose that a certain blocking maneuver leads to a discussion surrounding the contested content and that the toxic speaker manages to convince the others of its truth or acceptability. In such a scenario, the contested content eventually becomes a shared assumption. Blocking, however, would still minimally succeed, since the toxic content would not slip into the common ground automatically and unnoticed; that is, *without conversation participants fully realizing it*. Preventing a speaker from smuggling in some controversial content through the "back-door" (as Langton [2018, 152] would put it) and making everyone pay attention to it is per se an important achievement and may serve as counterspeech when that content is toxic.

It is worth pointing out at this point that, according to some, drawing everyone's attention to certain toxic contents is a double-edged sword and risks making paradigmatic instances of blocking backfire.[5] By unpacking and bringing toxic associations to the surface, paradigmatic blocking may make them more contextually *salient* and thus more cognitively available to participants. This is potentially troubling, because empirical studies suggest that the more an association is cognitively available, the more people are disposed to believe it and to act on it (Lewandowsky et al. 2012). So, increasing the cognitive availability of bigoted associations may further bias people toward bigoted choices and behaviors.

Note that paradigmatic blocking (i.e., explicitation-plus-denial) has denying as one of its constituents, and this makes it *confrontational* in character. Furthermore, and related, paradigmatic blocking can be *face-threatening*: it threatens the "positive face" (or reputation)[6] of the speaker, who has not *said* anything bigoted and yet is called out for bigotry. As already pointed out, sometimes, openly confronting a toxic speaker and threatening their face is exactly what one should do. When a politician tries to smuggle in some bigoted assumptions, it is of utmost importance that their attempt be brought to light and that they be forced to take responsibility for the toxic contents their words tacitly conveyed. But threatening another person's face may be perceived as aggressive or uncooperative and may lead to backlash and retaliation. Just as denying, blocking may thus be unsuited or unsafe for counterspeakers who are contextually, socially, or institutionally at a disadvantage in comparison to the toxic speaker.

In the next section, we shall look at a counterspeech strategy that operates in a subtler way and thus may come across as less confrontational and less face-threatening. Elsewhere, we called this strategy *bending* (Caponetto and Cepollaro 2022).

Bending

Consider the revised version of our "women-just-can't-drive" example once again. As you will recall, John, Paul, and Arthur are chitchatting, when they see Sarah in a large SUV. John says,

(4) Wow! That's huge. No doubt she'll have her husband park it for her.

Now suppose that Arthur perfectly realizes that John meant to suggest that women are bad drivers. Yet, he replies as if he interpreted John's remark quite differently:

(8) You're right, she should definitely give him parking lessons! He's so bad at parking. But her SUV is new, I doubt she'll trust him with it.

This is a deviant reply to (4). John intended to suggest that Sarah, as a woman, cannot possibly park a large car. Arthur gets it but replies as if John meant that Sarah should let her husband park her SUV as a way for him to practice driving. Arthur *bends* John's move by treating it as conveying a different, less toxic content. Not only will the assumption that women are worse drivers than men fail to enter the common ground by default but if John does not retort, the conversation will proceed under the assumption that women sometimes are *better* drivers than men, as proven by Sarah and her husband's case.

Bending consists in distorting a certain toxic contribution into an innocuous (or at least less toxic) one. It is a form of *acting as if*: the counterspeaker realizes that a given utterance implicitly conveys that p (a toxic content) but acts as if they took it to implicitly convey that q (an innocuous or less toxic content). In doing so, they prevent p from being incorporated into the common ground by default and attempt to make q enter the common ground instead – something they will manage to do if the toxic speaker plays along.

Since bending partly relies on toxic speakers playing along, the question as to why they would do so arises. Our answer appeals to social norms of equality. Many ordinary social contexts are governed by norms prescribing people, for example, not to be racist or sexist. Clearly, and problematically, such norms do not preclude people from engaging in everyday racism or sexism. Still, they do pressure people not to do so openly (Saul 2018; Mendelberg 2001). This partly explains why everyday bigotry often (although not always) takes implicit, rather than explicit, forms. Bending distorts an implicitly toxic utterance by making it better aligned with equality norms and gives the toxic speaker a sense that bigotry may not be well received by the audience. By distorting John's contribution, Arthur makes it better aligned with the norm of gender equality, and this may lead John to think that open sexism would not be well received in that context. Faced with (8), John *could* in principle retort and openly commit to the content *that women are bad drivers* ("What?! How could *she* teach *him*? Women just can't drive"). But this would be an open violation of the norm of gender equality – which is generally socially risky, and particularly so after Arthur's countermove.

Bending plays the same function as blocking: it prevents a certain implicit content from entering the common ground by default. In this sense, bending is a form of blocking. However, it is a *distinctive* form of blocking that also attempts to sneak in an "ameliorated content"; that is, a less toxic content than the one conveyed by the speaker's utterance. Qua blocking move, bending succeeds when it fulfills blocking's function. Qua distorting move, it succeeds when, in addition, it manages to make the ameliorated content enter the common ground in place of the toxic content. This is the case when the toxic speaker plays along and does not retort by explicitly asserting the toxic content they were implicitly conveying.

Interestingly, bending operates in a covert manner: the one who bends does not point out *what* was wrong, and not even *that* something was wrong, with the toxic speaker's utterance. This usually makes bending maneuvers *less confrontational* and *less face-threatening* than paradigmatic blocking. By acting as

if John's move did not carry anything sexist, Arthur gives him a chance to tacitly disavow his sexist assumption – to carry on as if he never meant to make it. Arthur gives John an opportunity to preserve his "face". Bending may thus be preferable to blocking when taking a confrontational stance toward one's interlocutor would be too risky, unwise, or otherwise undesirable. (Conversely, blocking may be preferable to bending when one wants to force the toxic speaker to take responsibility for their sneaky suggestions.)

If carefully crafted, bending maneuvers can also avoid raising the contextual salience of prejudiced associations. In saying,

(8) You're right, she should definitely give him parking lessons! He's so bad at parking. But her SUV is new, I doubt she'll trust him with it,

Arthur does not make the association between women and poor driving skills any more contextually salient – something he would have done had he opted for paradigmatic blocking instead:

(6) Are you assuming she can't park it because she's a woman? That's ridiculous.

Note, however, that this virtue of bending is conditional on how bending maneuvers are packaged. Suppose that, instead of (8), Arthur uttered,

(9) You're right. It's so sad and enraging to see how skillful women are made insecure by a patriarchal society, and think they need to rely on men to carry out basic stuff. I mean, she obviously knows how to park her SUV, she's an excellent driver. And yet, I'm sure she doubts her capabilities and has her husband do it for her.

This reply is an instance of bending: Arthur acts as if John were expressing disappointment at certain gender stereotypes, rather than endorsing them. Yet, it *does* contribute to raising the contextual salience of stereotypes against women (and women drivers in particular).

The characteristic features of bending make it an interesting counterspeech strategy, which may be particularly well suited when the counterspeaker has an interest in not being perceived as too confrontational or in not openly threatening the toxic speaker's face. Bending may also have mitigated salience-raising effects than alternative strategies, although, as we have pointed out, this will depend on how it is crafted.

Saying Nothing

Much of the literature on counterspeech seems to operate under the assumption that remaining silent in the face of a given discursive move is to accept it, at

least for the sake of the conversation. If silence equals acceptance, then saying nothing can never serve as a counterspeech strategy: to counter toxic speech, one necessarily has to speak out against it.

A number of scholars have argued that silence entitles one's audience to presume that one accepts or approves of what has been said[7] (see, esp., Pettit 1994). While some scholars have explored potential defeaters to the silence–acceptance equivalence (Langton 2007; Goldberg 2018, 2020; Lackey 2018), others have gone as far as to claim that silence can even be expressive of dissent. We will here draw upon recent work by Alessandra Tanesini on eloquent silence as a way of expressing dissent, with the aim of assessing whether silence can, in certain circumstances, constitute a form of counterspeech.[8]

Tanesini (2018, forthcoming) maintains that silence cannot be presumed to communicate acceptance by default. Taking the default (though defeasible) interpretation of silence to be acceptance is to fail to appreciate the distinction between eloquent silences and failures to object. "Eloquent silences" are deliberate silences that are intended to communicate. An eloquent silence is an act and can be a *speech* act (i.e., a communicative means); a failure to object, by contrast, is best thought of as an omission. To substantiate the point, consider the Gricean case of a person who, at a tea party, states that "Mrs. X is an old bag". Grice (1975, 54) imagines the statement to be followed by an "appalled" silence, after which one interlocutor changes the subject to a discussion about the weather. Unlike Grice, who is primarily interested in the change of subject and how it flouts the conversational maxim of relevance, Tanesini is interested in the silence that precedes it and how it can itself communicate disapproval. Far from being a failure to object, an appalled silence in the face of an inappropriate claim can clearly communicate that "the speaker's comment should not be dignified with a response" (Tanesini 2018, 118). In a somewhat similar vein, suppose my partner asks me, "Are you still mad at me?" and I deliberately remain silent. My silence is *eloquent*: it communicates an affirmative answer to my partner's question. By remaining silent, I intend that they believe that I am still mad at them and that they recognize that I have this intention (Tanesini 2018, 114; example adapted from Saville-Troike 1995, 9).

So, not only can silence fail to communicate acceptance, it can even be a way of expressing disapproval or dissent. Interestingly for our purposes, Tanesini (2018) suggested that silence is paradigmatically communicative of dissent when verbal behavior on the part of the silent person would be *expected*. When athletes remain silent as the national anthem plays, their silence clearly communicates disapproval, as they would be expected to sing along. And when a political activist remains silent during an interrogation, their silence is not a mere failure to provide information but communicates their deliberate refusal to do so, as interrogation questions make informative answers expected. In the vocabulary of conversation analysts, "question–answer" is an *adjacency pair*; that is, a two-part exchange in which the first part makes the second relevant

and expected (Schegloff and Sacks 1973; see also Levinson 1983). Other examples of adjacency pairs include "greeting–greeting", "congratulations–thanks", "offer–acceptance/refusal", etc. When silence occurs in place of the second part of an adjacency pair, it *overtly* violates an expectation of verbal behavior. Speakers may exploit this to make their disapproval of something manifest – to communicate a silent implicature of dissent (Tanesini 2018).

This makes room for the possibility that when toxic speech makes verbal behavior of some sort expected, silence can serve as counterspeech. Consider yet another version of our "women-just-can't-drive" scenario. John, Paul, and Arthur are chitchatting, when they see Sarah getting in a large SUV. John says,

(10) So guys, comments about the gal on her way to smash that car for good?

Suppose (10) is met with silence. No one comments. No one laughs or smiles. Embarrassed, John eventually changes the subject. By remaining silent, Paul and Arthur defeat an expectation of verbal behavior set by John's question and thereby manage to successfully communicate disapproval. Their silence can be cast as a form of *blocking*: John's question carries the sexist assumption that women are bad drivers; by keeping silent, Paul and Arthur block its way into the common ground.

Generalizing from this, saying nothing can serve as counterspeech, at least when it defies an expectation of verbal behavior set by the toxic speaker's move. Clearly, this is not but an initial step into an exploration of silence as counterspeech. Albeit sketchy, however, it interestingly goes against the tide in undermining the assumption that counterspeech requires one to verbally step in.

When silent counterspeech is an option, it may be a particularly well suited one for those who occupy disadvantaged social positions, feel relatively powerless, or have been variously silenced. When speaking up would be an unpromising way to go – for example, because one would not be given the credibility one would deserve – or would be too risky, silence may provide the best (if not the only) shot one has at counterspeaking (Tanesini 2018). Notice, moreover, that silence has no salience-raising effects: since it fails to engage with the toxic content *at all*, it entirely avoids the risk of raising the contextual salience of the prejudicial associations conveyed by the speaker.

Admittedly, though, silence can be employed as counterspeech only in a specific (and perhaps very limited) set of contexts; for example, when a toxic utterance is the first part of an adjacency pair. Furthermore, since the very same action of keeping silent can constitute several different speech acts, eloquent silences seem to be highly vulnerable to be misunderstood, wrongly interpreted as noncommunicative, or distorted into communicative contributions other than those the silent person intended to make (Klieber 2021).

Preemptive Moves

Let us conclude our overview with what may be called "preemptive counterspeech". Counterspeech moves are prototypically *reactive* or *post hoc*: the counterspeaker par excellence is a speaker who reacts to a given toxic utterance by speaking back against it to remedy or mitigate its harmful effects. Another way to put it is to say that counterspeech is prototypically a *second-turn intervention*: a response to a first-turn contribution conveying, either explicitly or implicitly, something toxic. In the recent philosophical literature, however, the label "counterspeech" is increasingly being used in a broader sense to also include *anticipatory* or *preemptive* moves. The basic idea behind preemptive counterspeech is that one can use language to condition the conversational context *in advance* and in such a way as to make it inhospitable to toxic speech (Tirrell 2018; Lepoutre 2019, 2021). Suppose that the state, through its officials, repeatedly affirms ideals of equality and mutual respect. In doing so, the state would contribute to enacting norms of equality that may render more socially costly, and thus less likely, for citizens to publicly say toxic things.

Preemptive moves of this sort count as counterspeech only where toxic speech is an existing problem. In an ideal world where toxic speech does not exist, promoting ideals of equality would not count as a form of counterspeech. There is thus a sense in which preemptive counterspeech, albeit temporally prior to (potential) toxic utterances, remains a second-turn intervention. Following Tirrell's (2018) epidemiological metaphor, preemptive counterspeech would be analogous to vaccines: measures introduced *in response* to certain existing diseases, even if they operate *ex ante*.

Maxime Lepoutre (2019, 2021) has recently argued that preemptive moves can alleviate a number of drawbacks often associated with *post hoc* counterspeech. For example, some scholars have expressed the worry that (*post hoc*) counterspeech, even when successful, may ultimately be unable to *undo* the harms of toxic speech: by the time the former comes into play, the latter may have already taken root in a way that cannot be easily reversed (McGowan 2009; Simpson 2013). Preemptive counterspeech is immune to this worry: when successful in conditioning the conversational context, it prevents toxic utterances from even being made.

Lepoutre (2019) also suggested that, *if positively framed*, preemptive counterspeech can avoid salience-raising effects. "Negative counterspeech", in Lepoutre's parlance, is counterspeech that repeats a certain toxic content in the process of negating it. Denying and paradigmatic instances of blocking fall under this category. "Positive counterspeech", by contrast, affirms egalitarian worldviews that implicate the falsity or untenability of certain toxic contents. While repeating a bigoted association to reject it may reinforce its contextual salience, affirming an egalitarian vision of the world that implicates its untenability without repeating it bypasses the problem.

It is no coincidence that our previous example involves state actors. Preemptive moves indeed seem to provide for an especially well-suited form of state-sponsored counterspeech. While state actors rarely find themselves in the position of resisting toxic speech on the spot, they can (and perhaps should) consistently promote education and awareness-raising campaigns, thus serving as preemptive counter-speakers. Lepoutre's (2021) recommendation is that they do so by carefully crafting those campaigns in positive (as opposed to negative) terms.

Preemptive moves are, however, potentially available to ordinary speakers as well. As we saw, toxic speech may shift the norms operative in a given context in harmful ways. As Mary Kate McGowan (2009) has argued, toxic speech may enact oppressive norms by rendering discrimination permissible or (more) appropriate in a given context. More recently, McGowan (2018) suggested that the same norm enactment mechanisms deployed by toxic speech can be used by ordinary speakers to enact egalitarian norms and promote justice. To see how this may be so, an example may help. Until relatively recently, scholars (including philosophers) used to refer to a generic individual by using the male pronoun "he". In this way, they contributed (often unwittingly) to reinforcing the assumption that maleness is the norm.[9] When feminist scholars started to refer to a generic individual by using the female pronoun "she" or the singular gender-neutral "they" in place of "he", their intervention contributed to the erosion of that normative assumption. Although McGowan does not use this example and does not talk about preemptive moves, we think the case nicely illustrates both what McGowan hints at and preemptive counterspeech. Indeed, not only did feminist scholars turn an androcentric assumption to its head by appropriating the mechanism responsible for its diffusion; they also conditioned the context (e.g., the philosophical arena) by making it gradually less hospitable to moves carrying that assumption.

Conclusion

In this chapter, we went through a number of counterspeech strategies discussed in contemporary philosophy of language. Our investigation makes it clear that the general question "What's the best counterspeech strategy?" is ill formed. When it comes to countering toxic speech, there is no one-size-fits-all solution. As we saw, each of the discussed strategies is promising under certain circumstances and unpromising, and even liable to backfire, under others. Philosophy of language may help us identify the main predictors as to which strategies may be better suited to respond to a given toxic utterance and most likely to succeed in a given conversational context.

Notes

1 The definition of *hate speech* is controversial; see, for example, A. Brown (2015) and Anderson and Barnes (2022). See also Lepoutre et al. (2023) for the first corpus-based analysis of the ordinary meaning of hate speech.

2 This chapter adopts a philosophy of language angle. For a general overview of the issues that counterspeech raises in philosophy, including moral and political philosophy, see Cepollaro, Lepoutre, and Simpson (2023).
3 The importance of contextual factors is equally recognized and discussed by Zollner (chapter 1) in this volume.
4 See also Lewis (1979), Langton (2018), and McGowan (2019) for versions of the idea that, insofar as nobody objects, implicit content tends to become common ground by default.
5 Lepoutre (2019, 160ff) provides a general discussion of this objection (see also the references therein).
6 The *loci classici* for the notion of "face" are Goffman ([1955] 1972) and P. Brown and Levinson (1978, 1987).
7 "What is said" captures explicitly conveyed content. This focus on explicit, rather than implicit, communication (and on assertion in particular) is not surprising, since the philosophical debate on the meaning of silence is not mainly concerned with silence in response to toxic speech. We think, however, that some considerations made within this debate can be easily adjusted to address the broader question as to whether silence communicates acceptance of a speaker's overall contribution; that is, of what they have explicitly and implicitly conveyed.
8 The question whether silence equals acceptance is closely tied to a different but related question; that is, whether people have a *duty* to manifestly express their dissent. We will not be concerned with this question here, but see Maitra (2012), Lackey (2018), Langton (2018), and McGowan (2018) for discussion. See also A. Brown (2019) and Saul (2021) on silence and dissent on social media.
9 On the false gender-neutrality of "he", see Moulton (1981).

References

Anderson, L, and M Barnes. 2022. "Hate Speech." In *The Stanford Encyclopedia of Philosophy*, edited by EN Zalta and U Nodelman. Stanford, CA: The Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2022/entries/hate-speech/. Accessed 6 September 2023.

Brandeis, Louis. 1927. "Concurring Opinion in Whitney v California." http://www.columbia.edu/itc/journalism/j6075/edit/readings/brandeis_concurring1.html. Accessed 6 September 2023.

Brown, A. 2015. *Hate Speech Law: A Philosophical Examination*. New York: Routledge.

Brown, A. 2019. "The Meaning of Silence in Cyberspace: The Authority Problem and Online Hate Speech." In *Free Speech in the Digital Age*, edited by SJ Brison and K Gelber, 207–223. Oxford: Oxford University Press.

Brown, P, and SC Levinson. 1978. "Universals in Language Usage: Politeness Phenomena." In *Questions and Politeness: Strategies in Social Interaction*, edited by E Goody, 56–289. Cambridge: Cambridge University Press.

Brown, P, and SC Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Caponetto L, and B Cepollaro. 2021. "'Discrimination Preferred': How Ordinary Verbal Bigotry Harms." *Australasian Philosophical Review* 5 (2): 189–195.

Caponetto, L, and B Cepollaro. 2022. "Bending as Counterspeech." *Ethical Theory & Moral Practice*. doi:10.1007/s10677-022-10334-4

Carnaghi, A, and A Maass. 2007. "Derogatory Language in Intergroup Context: Are 'Gay' and 'Fag' Synonymous?" In *Stereotype Dynamics: Language-based Approaches to the Formation, Maintenance, and Transformation of Stereotypes*, edited by Y Kashima, K Fiedler, and P Freytag, 117–134. New York: Lawrence Erlbaum.

Cepollaro, B.n.d. "A Taxonomy of Blocking Strategies." Unpublished manuscript.
Cepollaro, B, M Lepoutre, and RM Simpson. 2023. "Counterspeech." *Philosophy Compass* 18 (1): e12890.
Dickter, CL, JA Kittel, and II Gyurovski. 2012. "Perceptions of Non-Target Confronters in Response to Racist and Heterosexist Remarks." *European Journal of Social Psychology* 42 (1): 112–119.
Fasoli, F, A Maass, and A Carnaghi. 2015. "Labelling and Discrimination: Do Homophobic Epithets Undermine Fair Distribution of Resources?" *British Journal of Social Psychology* 54 (2): 383–393.
Fasoli, F, MP Paladino, A Carnaghi, J Jetten, B Bastian, and PG Bain. 2016. "Not 'Just Words': Exposure to Homophobic Epithets Leads to Dehumanizing and Physical Distancing from Gay Men." *European Journal of Social Psychology* 46 (2): 237–248.
Goffman, E. (1955) 1972. "On Face-Work: An Analysis of Ritual Elements in Social Interaction." In *Communication in Face-to-Face Interaction*, edited by J Laver and S Hutcheson, 319–346. Harmondsworth: Penguin.
Goldberg, SC. 2018. "Dissent: Ethics and Epistemology." In *Voicing Dissent. The Ethics and Epistemology of Making Disagreement Public*, edited by CR Johnson, 40–60. New York: Routledge.
Goldberg, SC. 2020. *Conversational Pressure: Normativity in Speech Exchanges*. Oxford: Oxford University Press.
Grice, HP. 1975. "Logic and Conversation." In *Syntax and Semantics: Vol. 3. Speech Acts*, edited by P Cole and JL Morgan, 41–58. New York: Academic Press.
Klieber, A. 2021. "'Your Silence Speaks Volumes': Silent Implicature and Its Political Significance." PhD thesis, University of Sheffield. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.855701. Accessed 6 September 2023.
Lackey, J. 2018. "Silence and Objecting." In *Voicing Dissent. The Ethics and Epistemology of Making Disagreement Public*, edited by CR Johnson, 82–96. New York: Routledge.
Langton, R. 2007. "Disenfranchised Silence." In *Common Minds: Themes from the Philosophy of Philip Pettit*, edited by M Smith, R Goodin, and G Brennan, 199–214. Oxford: Oxford University Press.
Langton, R. 2012. "Beyond Belief: Pragmatics in Hate Speech and Pornography." In *Speech and Harm. Controversies over Free Speech*, edited by I Maitra and MK McGowan, 72–93. Oxford: Oxford University Press.
Langton, R. 2018. "Blocking as Counter-Speech." In *New Work on Speech Acts*, edited by D Fogal, DW Harris, and M Moss, 144–164. Oxford: Oxford University Press.
Leader, T, B Mullen, and D Rice. 2009. "Complexity and Valence in Ethnophaulisms and Exclusion of Ethnic Out-Groups: What Puts the 'Hate' into Hate Speech?" *Journal of Personality and Social Psychology* 96 (1): 170–182.
Lepoutre, M. 2019. "Can More Speech Counter Ignorant Speech?" *Journal of Ethics and Social Philosophy* 16: 155–191.
Lepoutre, M. 2021. *Democratic Speech in Divided Times*. Oxford: Oxford University Press.
Lepoutre, M, S Vilar-Lluch, E Borg, and N Hansen. 2023. "What Is Hate speech? The Case for a Corpus Approach." *Criminal Law and Philosophy*. doi:10.1007/s11572-023-09675-7
Levinson, SC. 1983. *Pragmatics*. Cambridge: Cambridge University Press.
Lewandowsky, S, U Ecker, C Seifert, N Schwarz, and J Cook. 2012. "Misinformation and Its Correction: Continued Influence and Successful Debiasing." *Psychological Science in the Public Interest* 13 (3): 106–131.
Lewis, D. 1979. "Scorekeeping in a Language Game." *Journal of Philosophical Logic* 8 (3): 339–359.

Maitra, I. 2012. "Subordinating Speech." In *Speech and Harm. Controversies Over Free Speech*, edited by I Maitra and MK McGowan, 94–120. Oxford: Oxford University Press.

McGowan, MK. 2009. "Oppressive Speech." *Australasian Journal of Philosophy* 87 (3): 389–407.

McGowan, MK. 2018. "Responding to Harmful Speech: More Speech, Counter Speech, and the Complexity of Language Use." In *Voicing Dissent. The Ethics and Epistemology of Making Disagreement Public*, edited by CR Johnson, 182–199. New York: Routledge.

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

Mendelberg, T. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton, NJ: Princeton University Press.

Moulton, J. 1981. "The Myth of the Neutral 'Man.'" In *Feminism and Philosophy*, edited by M Vetterling-Braggin, FA Elliston, and J English, 124–237. Lanham, MD: Rowman & Littlefield.

Mullen, B, and JM Smyth. 2004. "Immigrant Suicide Rates as a Function of Ethnophaulisms: Hate Speech Predicts Death." *Psychosomatic Medicine* 66 (3): 343–348.

Rasinski, HM, and AM Czopp. 2010. "The Effect of Target Status on Witnesses' Reactions to Confrontations of Bias." *Basic and Applied Social Psychology* 32 (1): 8–16.

Pettit, P. 1994. "Enfranchising Silence: An Argument for Freedom of Speech." In *Freedom of Communication*, edited by T Campbell and W Sadurski, 45–55. Aldershot, UK: Dartmouth.

Saul, J. 2018. "Dogwhistles, Political Manipulation, and Philosophy of Language." In *New Work on Speech Acts*, edited by D Fogal, DW Harris, and M Moss, 360–383. Oxford: Oxford University Press.

Saul, J. 2021. "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: Dupery by Design*, edited by A MacKenzie, J Rose, and I Bhatt, 139–157. Cham: Springer.

Saville-Troike, M. 1995. "The Place of Silence in an Integrated Theory of Communication." In *Perspectives on Silence*, edited by D Tannen and M Saville-Troike, 2nd ed., 3–18. Norwood, NJ: Ablex.

Sbisà, M. 1999. "Ideology and the Persuasive Use of Presupposition." In *Language and Ideology*, edited by J Verschueren, vol. 1, 492–509. Antwerp: International Pragmatics Association.

Schegloff, EA, and H Sacks. 1973. "Opening Up Closings." *Semiotica* 8 (4): 289–327.

Simpson, RM. 2013. "Un-Ringing the Bell: McGowan on Oppressive Speech and the Asymmetric Pliability of Conversation." *Australasian Journal of Philosophy* 91 (3): 555–575.

Soral, W, M Bilewicz, and M Winiewski. 2018. "Exposure to Hate Speech Increases Prejudice through Desensitization." *Aggressive Behavior* 44 (2): 136–146.

Stalnaker, R. 2002. "Common Ground." *Linguistics and Philosophy* 25: 701–721.

Swim, JK, LL Hyers, LL Cohen, and MJ Ferguson. 2001. "Everyday Sexism: Evidence for Its Incidence, Nature, and Psychological Impact from Three Daily Diary Studies." *Journal of Social Issues* 57: 31–53.

Tanesini, A. 2018. "Eloquent Silences: Silence and Dissent." In *Voicing Dissent. The Ethics and Epistemology of Making Disagreement Public*, edited by CR Johnson, 109–128. New York: Routledge.

Tanesini, A.forthcoming. "Speech in Non-Ideal Conditions: On Silence and Being Silenced." In *Sbisà on Speech as Action*, edited by L Caponetto and P Labinaz. Cham: Palgrave MacMillan.

Tirrell, L. 2017. "Toxic Speech: Toward an Epidemiology of Discursive Harm." *Philosophical Topics* 45 (2): 139–161.
Tirrell, L. 2018. "Toxic Speech: Inoculations and Antidotes." *The Southern Journal of Philosophy* 56: 116–144.
Waldron, J. 2012. *The Harm in Hate Speech*. Cambridge, MA: Harvard University Press.