

# How to talk back: hate speech, misinformation, and the limits of salience

Politics, Philosophy &amp; Economics

2023, Vol. 22(3) 315–335

© The Author(s) 2023



Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/1470594X231167593

[journals.sagepub.com/home/ppe](https://journals.sagepub.com/home/ppe)**Rachel Fraser** 

University of Oxford, UK

## Abstract

Hate speech and misinformation are rife. How to respond? Counterspeech proposals say: with *more* and *better* speech. This paper considers the treatment of counterspeech in Maxime Lepoutre's *Democratic Speech In Divided Times*. Lepoutre provides a nuanced defence of counterspeech. Some counterspeech, he grants, is flawed. But, he says: counterspeech can be debugged. Once we understand *why* counterspeech fails – when fail it does – we can engineer more effective counterspeech strategies. Lepoutre argues that the failures of counterspeech can be theorised using the ideology of salience. *Negative* counterspeech fails because it reinforces the salience of the very ideas or associations that it contests. His solution? *Positive* counterspeech – a form of counterspeech which avoids the salience trap. I argue that the salience paradigm is ill-suited to theorise the failures of counterspeech. I suggest some alternatives. Further, I show that these alternative paradigms make importantly different *practical* recommendations – recommendations concerning how we ought to engineer our counterspeech – from those issued by the salience paradigm.

## Keywords

hate speech, misinformation, counterspeech, deliberative democracy

## Introduction

Deliberative democrats place inclusive public speech at the heart of the democratic ideal (Lepoutre, 2021). The ideal of inclusive public speech is typically argued for twice-over, once epistemically, once politically. Lepoutre articulates both with a characteristic sparse elegance. The epistemic argument for inclusive political communication crosses

---

### Corresponding author:

Rachel Fraser, Exeter College, University of Oxford, Oxford, OX1 2JD, UK.

Email: [Rachel.fraser@philosophy.ox.ac.uk](mailto:Rachel.fraser@philosophy.ox.ac.uk)

standpoint epistemology with the Millian argument for free speech. It generally goes something like this:

The knowledge required to identify and address social problems tends to be dispersed across different social groups. Without inclusive political communication, there is no effective way to pool these dispersed epistemic resources; hence, inclusive political communication is required if we are to collectively address and identify social problems.

This is appealing. Black citizens are more likely than white to appreciate the problem of police brutality. Women are better placed than men to grasp the pervasiveness and moral seriousness of sexual harassment. The depth and character of rural poverty will be more salient to rural citizens than their urban counterparts. The patterning of politically relevant knowledge with social position is not accidental: the place one occupies in a social structure shapes one's physical and normative environment; differences in such environs give rise to differences of knowledge. And when knowledge is widely dispersed, communication offers the most promising route to widespread epistemic convergence.

The political argument for inclusive political communication is similarly appealing. Here it is in outline:

Domination occurs when those who wield power may do so *arbitrarily*, viz., without the existence of checks to ensure that their exercises of power track the interests of those over whom it is wielded. Access to public discourse is a valuable political tool with which to guard against domination: by speaking in the public sphere, citizens can help to ensure that political decisions track their interests.<sup>1</sup>

The trouble is obvious. The ideal of inclusive political communication might look good on paper. Off paper, its prospects appear bleak. When we look at the public sphere, we find them 'saturated with emotional appeals, including intensely negative emotions such as rage and resentment' (Lepoutre, 2021: 2).

[S]peakers routinely use their airtime to ridicule, demean, or vilify others. And where sincerity should reign, campaigns of misinformation instead proliferate unimpeded. (Lepoutre, 2021: 2)

Rather than pooling knowledge, then, public speech can disperse misinformation and entrench epistemic fragmentation. Call this the *informational challenge* to the inclusive ideal. And rather than guarding against domination, speech – in the guise of hate-speech – can oppress and exclude. Call this the *affective challenge* to the inclusive ideal.

Anger, vilification, and lies are nothing new. But contemporary pessimists about the value of inclusive public speech typically do more than gesture to the perennial unpleasantness of political rhetoric, as if no one had noticed its nastiness for the last three hundred years. (Granted, when reading Mill at his most wide-eyed, one might forgive them this.) Rather, they present their pessimism as fed by two relatively recent developments. First, the internet. It's a truism that the internet in general – and social media in particular – makes misinformation easier to spread; insofar as misinformation threatens to undermine the value of inclusive public speech, the internet threatens to undermine

the value of inclusive public speech. Second, a growing body of empirical work, held to suggest that ‘most citizens process political information in deeply biased, partisan, motivated ways rather than in dispassionate ways’ (Brennan, 2016: 37). So armed, contemporary pessimists can argue that the gap between the inclusive ideal and the often ugly realities of public speech is not the benign gap that must hold open between any contentful norm and its target. Rather, there are principled reasons, buried deep in our cognitive and social infrastructures, which block these ideals from applying to creatures like us.

The optimist, when confronted with such a challenge, cannot offer the glib rejoinder that they are interested in how things ought to be, rather than in how they are. After all, ‘even for an ideal, there is such a thing as being *too* distant from reality’ (Lepoutre, 2021: 2). As Lepoutre points out, ideals that are too distant from their targets may fail to guide real-world political action (Lepoutre, 2021: 2).

The challenge for the optimist, then, is deep and serious. Happily for the optimist, Lepoutre’s defence of optimism position is deep and serious, too. And its depth and seriousness are only two of its many virtues: the book is careful, sensitive, and nuanced. It is never pious, sneering, or doctrinaire. It’s also very nicely written, with an easy, unobtrusive grace. To defend the inclusive ideal against the informational and the affective challenge, Lepoutre argues that the problems posed by hate speech and misinformation can be tackled by mobilising the inclusive ideal rather than by departing from it. His proposal revolves around *counterspeech*: speech which aims to undo – or at the very least ameliorate – the harms of hate speech and misinformation.

In less capable hands, such a proposal might read as studiously ingenuous. Lepoutre, though, makes the proposal sound both generous and realistic. But whilst I am broadly sympathetic to Lepoutre’s emphasis on counterspeech – and his corresponding wariness of legal remedies – I have reservations about some of the details of his picture. I’ll start by discussing the material in his fourth chapter, where he considers the challenge posed by hate speech for the inclusive ideal. I’ll then move on to discuss the material in Lepoutre’s fifth chapter, where he considers the challenge posed by misinformation.

But before moving on to this more detailed consideration of Lepoutre’s position, let us take a step back. The problems of hate-speech and misinformation are problems that might be viewed through multiple lenses. One *might* approach them from the perspective of a value-neutral sociologist, as someone who takes a sanguine interest in the dynamics of political speech.<sup>2</sup> But this is not the only option. One might, alternatively, approach these phenomena as an engaged political actor, as someone who is struggling against oppression. From such a perspective, hate speech and misinformation will always be shot through with normative significance, and marbled with social detail, features elided by the cool eye of the sociologist.

The approach to hate-speech taken by Lepoutre might be thought to steer unduly close to that of our value-neutral sociologist, and in ways that belie the intellectual heritage of which he is an inheritor. The feminist philosophical tradition he often draws on, for example, tends to be more pugnacious in its politics than he is. But rather than seeking to excavate this tension, I will, for the most part – a few cautionary notes will be sounded – stick fairly closely to the methodological tack taken by Lepoutre.<sup>3</sup> Let us explore. Let us see how far such a tack can take us.

## Defining hate speech

Lepoutre takes hate speech to be – roughly – ‘speech that communicates or otherwise expresses the inferiority of other members of society’ (Lepoutre, 2021: 106). One such an account, ‘hate speech’ becomes (happily, in my view) something of a misnomer: one can express the thought that a group is inferior without hating or resenting that group. Think of the benevolent sexist (Glick et al., 2000) or the unprejudiced but cynically populist politician. Less happily, it’s also an account of hate speech with a broadly Kantian flavour, on which the problem with hate speech is its denial of something like recognition respect: ‘hate speech emphatically rejects the basic standing of its targets as equals’ (Lepoutre, 2021: 86). Perhaps such an account captures the moral texture of some hate speech. But in other cases, it is at best highly artificial.

I have two arguments for this charge of artificiality. First, Lepoutre’s account of hate speech is unduly fixated on ‘inferiority’ as its key moral category. Second, Lepoutre’s account of hate speech is insufficiently *political*. Let’s take these arguments in turn.

First, consider three of Lepoutre’s key examples:

Hate speech might include newspaper articles falsely attributing essential dangerousness to a minority group (‘Muslims are terrorists’). Or it might take the form of leaflets portraying some groups as subhuman (for instance, depictions of black people as ape-like). Sometimes, hate speech explicitly expresses the exclusion of a minority group from the political community, as in Pegida’s infamous posters asserting ‘Rapefugees not welcome’ (Lepoutre, 2021: 86).

Only the second of these examples expresses the *inferiority* of other members of society. To regard someone as threatening by nature, or as a dangerous outsider need not be to regard them as one’s *inferior*. Seeing someone as an equal is, alas, no guarantee of fellow-feeling; there are some forms of loathing and cruelty which we reserve for our fellow humans. Consider misogynistic violence, often which is often framed by its perpetrators as *punishment* for its victims’ moral transgressions (Manne, 2016). In such cases, recognition of women’s humanity does not serve as a break on violence, but is rather a pre-condition for its intelligibility (Manne, 2016). Manne provides a helpful model for thinking about claims like ‘Muslims are terrorists.’ Such claims mark Muslims as moral transgressors. Ergo, they do not dehumanise. ‘Moral transgressor’ is not the sort of concept we apply to (non-human) animals. Rather, it is a category whose application presupposes the distinctively human capacity for moral agency.<sup>4</sup>

It might be helpful, then, to distinguish between two kinds of hate speech. Some hate speech marks its targets as sub-human, or presents them as something less than ‘full persons.’ Call this *inferiorising* hate-speech. But not all hate speech is like this. Some hate speech is *expulsory*. *Expulsory* hate speech does not deny its targets’ basic moral standing. Rather, it attempts to expel its targets from the speaker’s *political* community.<sup>5</sup> Where inferiorising hate speech says ‘You are my inferior’, expulsory hate speech says ‘You do not belong’. Claims like ‘Muslims are terrorists’ and ‘Rapefugees not welcome here’ are, I hold, most aptly conceptualised as expulsory hate speech. Serious wrongdoing is often linked with attenuated or degraded forms of political belonging.

(Consider the phenomenon of prisoner disenfranchisement.) Hence claims like ‘Muslims are terrorists’, by marking their targets as moral transgressors, suggest that their targets *are not or should not* get to be (fully) part of our political community.

Now for the worry: that Lepoutre’s characterisation is insufficiently political. Look again at Lepoutre’s three key examples. Not one is an example in which some random individual claims that someone else is his inferior. In all three, speech targets an already oppressed and vulnerable social group. In all three, speech is used to entrench an existing, brutal hierarchy. This is not a normatively incidental feature of the examples. It is crucial to their moral-cum-political charge. But Lepoutre’s account of hate speech makes no mention of hierarchy. His characterisation of hate speech, then, risks being un-moored from the concrete features of hate speech on whose account we *care* about the phenomenon in the first place.

This is not just nitpicking. They matter. Infelicities in our delineation of hate speech might bleed into infelicities in our normative and practical diagnoses. Broadly Kantian accounts of hate speech both invite and dovetail neatly with ‘dignitarian’ account of its harm. On dignitarian accounts, citizens, as a matter of justice, must not only have their dignity recognised and upheld by other citizens; they must also know and be assured of this recognition. Knowledge of this steady recognition is what hate speech attacks; without it, citizens cannot ‘pursue their aims and participate in civil and political life without fear or shame’ (Lepoutre, 2021: 88).

But dignitarian accounts of hate speech face a dilemma. Either they define dignity narrowly, or they define dignity broadly. Suppose we define dignity narrowly, for example, as a matter of recognition respect. To say that someone is entitled to recognition respect is to say that they are ‘entitled to have other persons take seriously and weigh appropriately the fact that they are persons in deliberating about what to do’ (Darwall, 1977). It is clear that inferiorising hate speech undermines its targets’ assurance of recognition respect. It is less clear that expulsive hate speech does the same. Suppose someone asserts that only white people belong in Britain. This is a revolting sentiment. But does it undermine its’ targets’ assurance that they will be treated with recognition respect? That’s less unclear. Not even the most fevered imperialist thinks only Brits are persons.

Suppose then that we define dignity broadly. Waldron, for example, understands dignity very broadly, as ‘a person’s basic entitlement to be regarded as a member of society in good standing’ (Waldron, 2012: 85). Understood so broadly, assurances of dignity clearly *are* undermined by claims like ‘only white people belong in Britain’. For such expulsive claims undermine non-white citizens’ assurances of belonging. It undermines their confidence that they are regarded as members of society in good standing because it undermines their confidence that they are regarded as really *being* members of society in the first place.

The trouble is that once we understand dignity so broadly, it starts sounding fantastic to keep thinking of it as something to which all persons have a basic entitlement. Most people do not have a ‘basic entitlement’ to be regarded as members of most societies. I am not, for example, entitled to be regarded by Danes as a member of Danish society; *a fortiori* I have no entitlement to be regarded as a member of Danish society *in good standing*.

The dilemma, then, is this. If dignity is narrowly defined, as something that all persons have, it will be hard to account for the harm of specifically expulsory hate speech in dignitarian terms.<sup>6</sup>

On the other hand, if dignity is defined more broadly, then we may be able to capture the harm of expulsory hate speech in dignitarian terms. But we will have given up on the idea that dignity is something all persons have (or to which all persons are entitled) in virtue of their personhood. Put differently, there is a deep tension here between the cosmopolitanism of the Kantian soil in which dignitarian accounts of hate speech have grown up, and the profoundly *uncosmopolitan* structure of the actually existing political communities whose boundaries hate speech is so often concerned to police.

The problem compounds for those inclined to think of hate speech primarily in terms of its social function – the maintenance of hierarchy – rather than as a communicative act which happens to express a particularly noxious attitude. To focus on hate speech, rather than on the social hierarchies it serves to consolidate, might well be read as ideological (in the pejorative sense): an attempt to ‘pin the blame’ on unpleasant individuals rather than grapple with unjust social structures.

But for now, let us bracket these worries and travel further down Lepoutre’s argumentative stream. Back to the inclusive ideal.

## Hate speech and the inclusive ideal

The challenge posed by hate speech for the inclusive ideal is obvious enough. Once we recognise the *harm* of hate speech, we come under pressure to ban hate speech. But once we ban hate speech, we seem to have abandoned, at least to some degree, the ideal of properly *inclusive* public speech.

One might try to deny that there is any genuine tension here: perhaps the ‘inclusion’ in inclusive public speech comes with a *sotto voce* break clause: ‘except for racists’. A more insidious option would be scepticism as to whether hate speech is really all that bad. (‘Snowflakes!’) Lepoutre, thankfully, takes neither route: he is not sanguine about hate speech any more than he is sanguine about the harms of banning it. Often, he points out, hate speech is produced by socio-economically disaffected citizens who feel, sometimes justifiably, that their voices have been ignored by mainstream parties; thus, coercively suppressing hate speech diminishes the freedom of many ‘relatively disadvantaged citizens’ by diminishing their ability to make political elites track their interests (Lepoutre, 2021: 91). Insofar, then, as one buys the political argument for inclusive public speech in the first place – which, recall, appeals to the value of freedom as non-domination – one is under pressure to accept at least a *prima facie* case against criminalising hate speech. The optimist about inclusion had better find a more congenial alternative. That alternative is *counterspeech*. Counterspeech strategies attempt to address harmful speech not by criminalising it, but by countering it with more and better speech.

Lepoutre’s case for counterspeech is a comparative one. It comes in two parts. Part one says that counterspeech comes with fewer moral costs than criminalisation. Part two says that counterspeech and criminalisation are at least on a par when it comes to preventing or ameliorating the harms of hate speech. It follows that we should prefer counterspeech to criminalisation.

Lepoutre largely takes the first part of the case – the claim that counterspeech does not come with the same moral costs as criminalisation – for granted. The major argumentative burden comes in showing that counterspeech is not a knife in a gunfight.

Counterspeech strategies face two major objections. First, that counterspeech is *ineffective*. It cannot prevent, disincentivise, or ameliorate the harms of hate speech. Second, that it is *worse* than ineffective. Counterspeech, the objection goes, risks exacerbating the harms of hate speech:

Suppose a public figure asserts ‘Xs are lazy parasites’ ...[V]erbally countering this assertion – for instance, by saying ‘That’s false! Xs are not lazy parasites’ – seems an ineffective way of reversing the initial utterance’s damaging conversational effects. Likewise, responding to a misogynist’s claim that women are submissive by saying ‘But women are *not* submissive’ seems misguided. The intuitive problem, in both cases, is that challenging the hateful view somehow ends up consolidating its place in the public conversation, and maintaining its harmful effects (Lepoutre, 2021: 98).<sup>7</sup>

Lepoutre follows Robert Simpson’s gloss of these dynamics, which appeals to the ideology of salience:

When a hateful speaker expresses a vilifying association (say, between being an X and being a parasite), they thereby make that association more salient to listeners. In this context, trying to repudiate the vilifying association risks amplifying its salience. That is, exclaiming ‘Xs are not lazy parasites!’ may challenge the claim that Xs are lazy parasites. But even as it does so, it risks magnifying the place of this stereotype within the public conversation (Lepoutre, 2021: 98).

Lepoutre’s key argumentative move is to point out that such worries only really apply to one kind of counterspeech, what he calls *negative* counterspeech. Negative counterspeech explicitly negates the content of hate speech. But, as Lepoutre points out, there are alternatives. According to the *positive* counterspeech proposal, when confronted with a hateful speaker who asserts that Xs are parasites, we should not say ‘You are wrong: Xs are not parasites!’ Instead, we should say (for example) ‘Xs help support the NHS’. Rather than reiterate, albeit by denying, the link between *being an X being and a parasite*, positive counterspeech works to seed a new, more positive association. Countering hate speech, on Lepoutre’s positive proposal, ‘is less about directly contesting a distorted vision of the world, and more about affirming a correct vision of the world’ (Lepoutre, 2021: 102).

This is a subtle contribution to an already mature dialectic. And I am sympathetic to the broad contours of Lepoutre’s attempt to rehabilitate and articulate refined versions of counterspeech. But I am hesitant about some of the details; in particular, the work Lepoutre wants the ideology of *salience* to do.

The first thing to note is that Lepoutre switches somewhat freely between two different kinds of salience-talk. Sometimes, he talks about an *association* being salient; sometimes about a *stereotype* being salient. One option, of course, would be to treat these as equivalent, that is, to take stereotypes to be associations. This, though, is not an especially

attractive view of stereotypes: stereotypes often feature in inferential processing; hence they must have something like propositional structure (Levy, 2015). Associations, by contrast, have no propositional structure:

[A]lthough associations connect explicitly represented concepts, the connections between concepts do not themselves represent anything...[T]here is nothing to the content of an association other than the content of the solitary concepts the associations relate... Thus an association between *salt* and *pepper* cannot represent a determinate proposition (Johnson, 2020: 1224).

The lack of propositional structure possessed by associations imposes strict limits on the kind of explanatory work they can do. That I strongly associate *salt* and *pepper* is just as likely to make me believe *salt is good with pepper* as *salt is bad with pepper*. That I strongly associate the concepts *cat* and *dog* does not dispose me to think that cats *are* dogs. Similarly, that I associate group *X* with laziness is no better an explanation of the belief *Xs are lazy* than it is of *X's are not lazy* (Johnson, 2020). This is because associations *per se* are 'semantically transparent': they do nothing to represent what sort of relationship obtains between the associated contents (Johnson, 2020). Accordingly, there is no way for associations to differentiate between a sentence saying '*s*' and a sentence saying '*s* is false'. Hence, a bare association on its own, however salient, cannot undermine the dignity or good standing of a group: it lacks the internal structure required to play that role.

Perhaps, then, we ought to concentrate on stereotypes, where stereotypes are construed as having propositional structure. When we make a stereotype salient, then, we do more than simply make a semantically blind association salient: we make (something like) a truth-apt claim salient. The problem, on Lepoutre's view, with counterspeech which merely negates such stereotypes is that it does nothing to undermine and risks *entrenching* the salience of the negated stereotype. Positive counterspeech, by contrast, avoids this danger.

This theorisation faces two challenges. The first is explanatory, the second extensional. Explanatory challenge first. If we think of stereotypes as semantically transparent associations, it's relatively clear why both an assertion that *Xs are not parasites* and an assertion that *Xs are parasites* might make the stereotype that *Xs are parasites* more salient. The stereotype, being semantically blind, 'doesn't care' about the presence or absence of the 'not'. But it turns out we shouldn't think of stereotypes this way: they have propositional structure. But this gives rise to a puzzle. In general, it does not seem that an assertion of 'not-*p*' makes the proposition that *p* a salient one, or vice versa. If you say 'Kampala is the largest city in Uganda', then – unless I am unusually suspicious of you – I don't immediately start entertaining the idea that Kampala is *not* the largest city in Uganda. But for Lepoutre's account to work, where *p* articulates a stereotype, that is precisely the dynamic we must observe. Once we fix on an account of stereotypes as propositionally structured, it becomes puzzling why negative counterspeech would make salient the stereotypes they contest.<sup>8</sup> That, in turn, suggests that salience is not the right tool for understanding why negative counterspeech seems, on an intuitive level, to consolidate the harmful effects of hate-speech.

Now for the extensional challenge. In pumping our intuitions about the worryingly consolidatory dynamics of (some) counterspeech, Lepoutre asks us to consider



someone responding to a sexist's claim that women are submissive by saying 'But women are not submissive'. That, he says, seems misguided: the response seems to consolidate the sexist view's place in public discourse. That all seems right to me. But now consider a variant of the case. Suppose Sam asserts that women are submissive. His colleague Bob responds with heavy sarcasm: 'You're right! Women are *so* submissive. My wife does whatever I tell her'. Bob's sarcastic response seems a good deal more promising than simply asserting 'Women are not submissive'. Rather than somehow entrenching the idea that women are submissive, it makes such a view seem ridiculous.

There is, then, an intuitive difference between negative counterspeech and what we might call sarcastic counterspeech. But the salience paradigm is too crude to capture the difference between the two. For surely Bob's response to Sam makes his sexist stereotype salient (if only in order to ridicule it). The problem with negative counterspeech can't be that it makes negative stereotypes salient, because *adequate* (or at least, more adequate) counterspeech sometimes makes negative stereotypes salient.

These challenges to the salience paradigm undermine Lepoutre's defence of the inclusive ideal. Lepoutre's defence of the inclusive ideal depends on our buying that positive counterspeech avoids the problems associated with negative counterspeech, because it challenges negative stereotypes without making them salient. But if we give up on using salience to theorise the shortcomings of negative counterspeech using the ideology of salience, it's unclear that we should be any more optimistic about positive counterspeech than we are about negative counterspeech.

However, I think the spirit, if not the letter, of Lepoutre's defence of the inclusive ideal can be preserved. Let's start is with the difference between negative and sarcastic counterspeech in response to the claim that women are submissive.

A promising way to model appeals to the ideology of *questions under discussion*. The basic idea here is that we can model conversational contributions as attempts to answer (an often implicit) *question*.<sup>9</sup> Take an assertion like 'Jane ate a cookie'. The valence of such an assertion will be different depending on whether the background question structuring the conversation is 'What did Jane eat?' or, 'Who ate the cookies?' One way to get a grip on the difference is to consider the different follow-up questions that these differences of question-under-discussion make appropriate. If the question under discussion is 'What did Jane eat?', it is natural to follow-up the assertion that Jane ate a cookie with a question like 'Was that all she ate?' By contrast, if the question under discussion is 'Who ate the cookies', such a follow-up is somewhat unnatural; far more natural would be a follow-up like 'Was she the only one to eat the cookies?' Keep in mind: these dynamics are complicated by the pervasive tendency for conversational score to update so as to make contributions count as 'correct play': asking 'Did anyone else eat the cookies' will often *change the question* under discussion from 'What did Jane eat?' to 'Who ate the cookies?' (Lewis, 1979).

One advantage of the questions under discussion framework is that it provides a neat way to distinguish at-issue from not-at-issue content. Consider the following examples:

1. Jane, who is vegetarian, ate all the cookies.
2. Sally, who ate some of the cookies, did not eat any of the salmon.

In both sentences, there is an intuitive difference between the status of the content expressed by the underlined constituents and that expressed by the unmarked constituents. The underlined content feels ‘sneaked in’, or like an aside. One way to capture this difference is to appeal to the analytic of questions under discussion: the underlined content acquires its incidental feel because it is naturally read as irrelevant to answering the question under discussion – which will be something like, ‘Who ate all the cookies?’ for (1), and, ‘Did Sally eat any of the salmon?’ for (2).

With this framework in place, we can propose a new, more nuanced account of the pragmatic effects of hate speech. Hate speech does not simply make certain claims or associations salient. Rather, when someone asserts ‘Xs are parasites’, the conversational score updates so as to make this contribution count as correct play. What this means, given the question-under-discussion framework, is that the question under discussion becomes something like ‘Are Xs parasites?’. This suggests both (i) a new way to capture how hate speech erodes assurances of dignity, (ii) why negative counterspeech is ineffective and (iii) why sarcastic counterspeech does better than negative counterspeech.

First of all, to be a member of society whose dignity is secure is precisely to be a member of groups whose status is not under discussion. For in general, knowing  $p$  is not (rationally) compatible with active inquiry into whether  $p$  (Friedman, 2019). Hence if we, as a society, are or seem to be collectively inquiring into whether Xs are parasites, we are not naturally read as knowing, collectively, that they are not. But Xs dignity is only secure if they know that we, collectively, know that they are not parasites. So if an assertion of ‘Xs are parasites?’ updates the conversational score so that ‘Are Xs parasites becomes a question under discussion’, it blocks Xs from knowing that there is collective knowledge that the answer is ‘no’.

Second, we can use the questions under discussion framework to better understand why negative counterspeech is not effective. Let’s start by returning to some of the innocuous examples above. Suppose you assert ‘Jane ate a cookie’. This assertion is naturally read as an attempt to answer a question like, ‘Did Jane eat a cookie?’. If I respond to my assertion by saying ‘Jane didn’t eat a cookie’, it is clear that you disagree with me as to how this question should be answered. But this overt disagreement risks masking a deeper level of agreement: both of us are tacitly treating the question of whether Jane ate a cookie as our question-under-discussion. In general, if you say that  $p$  and I say that not- $p$ , we agree that the question-under-discussion is, ‘Is it the case that  $p$ ?’. Applied to the case of negative counterspeech, then, we get the following distressing picture. If you say ‘Xs are parasites’, you erode Xs assurance of dignity by updating the conversational score so that ‘Are Xs parasites?’ is the question under discussion. When I respond by saying ‘Xs are not parasites’, this is a failure qua counterspeech not because it keeps a negative stereotype salient, but because it too treats ‘Are Xs parasites?’ as a question under discussion. And it is the very presence of this question as a question under discussion which erodes Xs assurance of their dignity. In short, if hate speech works in part by distorting the background informational structure of public discourse, negative counterspeech does not work because it leaves said informational structure intact.

Third, we can use the question-under-discussion framework to explain the difference between negative and sarcastic counterspeech. When Bob responds to Sam’s assertion

that women are submissive with sarcastic ridicule, his response changes the question under discussion from a question about whether women are submissive to a question like 'why would anyone be stupid enough to think that women are submissive'. Hence even if both negative and sarcastic counterspeech may make sexist (for example) stereotypes salient, their effects on the *background informational structure* of the discourse to which they contribute are very different. Bob's response changes the conversation from one in which the status of women is under discussion to a discussion in which the status of *sexists* is under discussion.

What does all this mean for Lepoutre's recommendation that we go in for positive, rather than negative counterspeech? It's not good news. Lepoutre recommends positive counterspeech in the place of negative counterspeech. This prescription is motivated by the thought that we should decrease the salience of negative stereotypes and increase the salience of positive ones. But now consider the following (yes, highly idealised) exchange using the lens of questions under discussion, rather than salience:

Racist. 'Xs are parasites'.

Anti-Racist. 'Xs support NHS!'

Now, it is true that our anti-racist's positive counterspeech will do better than merely negative counterspeech. It is not really possible to read the anti-racist's assertions as attempts to answer the question 'Are Xs parasites?' Hence the effect of this conversational contribution will, if things go well, be to update the score so that 'Are Xs parasites?' is no longer a question under discussion. However. It is not all good news for our Anti-Racist.

It is, after all, extremely natural to interpret the anti-racist's counterspeech as an attempt to answer questions like 'What are Xs like?' and 'Do Xs contribute to society?' Now, to have these as questions under discussion is, to be sure, an improvement on having questions like 'Are Xs parasites?' as the questions covertly structuring public discourse. But to be a member of society whose dignity is fully and stably secure is to be a member of society whose groups are not the subject of such questions. In other words, because our anti-racist's counterspeech treats claims about the social character of Xs as at-issue content, the positive counterspeech re-inscribes much of the exclusionary force of the racist's initial hate-speech. It shifts the question under discussion a little, but not enough to fully undo the insidious pragmatic effects of the racist's contribution: it concedes that the status of Xs is, as it were 'up for debate', and so colludes in undermining Xs knowledge of their good standing. To *know* that one is in good standing, after all, is to simply take this good standing for granted; for the question of the good standing simply not to arise.

This is not to suggest is that Lepoutre's proposal is hopeless. But it does need some fine-tuning. The lesson is this: positive counterspeech will more effectively undo the harms of hate speech if it occurs as not-at-issue content, rather than as at-issue content.<sup>10</sup> Recall the not-at-issue underlined contents in (1) and (2). The underlined contents felt 'smuggled' in: they bypass the questions structuring the discourse. Positive counterspeech must treat its anti-racist content as not at issue content: that way, it gets imported into the conversation without embedding demeaning questions in the background informational structure. To illustrate the difference: when we assert 'Xs help support the NHS', the claim about Xs is at issue content. By contrast, when we say

things like ‘The NHS, which heavily depends on Xs, is a treasured institution’, it is not. The former risks embedding questions about Xs as ‘under discussion’ in a way that the latter does not.

Hence there is a significant difference between the recommendations issued in by a salience based diagnosis of hate speech and one which appeals to the questions under discussion framework. If making positive associations salient is the goal, the first contribution looks preferable to the second. But if I am right, a preference for the first contribution is misguided.<sup>11</sup>

## Misinformation

Misinformation, for Lepoutre, is speech that disseminates or promotes falsehoods about political matters. As with his definition of hate speech, then, Lepoutre adopts an ‘externalist’ understanding of misinformation, on which the inner-life of the speaker is not relevant. A misinformer, for Lepoutre, need not be a liar, or insincere. This is an especially crucial feature given how ‘depersonalised’ our informational environment has become: when it comes to bots who tweet harvested conspiracy theories or search engine rankings which make vaccine-myths highly clickable, there simply is no speaker after whose beliefs we might inquire. Nonetheless, we have misinformation.

The pervasiveness of misinformation looks to undermine both the epistemic and the political argument for inclusive public speech. The way in which it undermines the epistemic argument is straightforward and obvious. The epistemic argument for inclusive political communication relies on the thought that inclusive communication has epistemic benefits: it widely circulates knowledge that would otherwise not leave certain social groups. Pervasive misinformation makes clear that inclusive communication also comes with costs: falsehoods, as well as truths will circulate; hence the net epistemic pay-off of such communication may well be negative.

One thing Lepoutre does not remark upon is that *quite* how devastating the prevalence of misinformation is to the epistemic argument depends a little on one’s background epistemic commitments. For someone who takes the epistemic benefits of inclusive communication to be the wide circulation of truths, misinformation will alter the net score of inclusive policies by adding to the negatives, rather than by erasing the positives. But for those who take the epistemic benefits of inclusive communication to involve the circulation of knowledge, pervasive misinformation may have still graver effects. Suppose, for example, that knowledge requires true belief that is safe from error. If pervasive enough misinformation can make it unsafe to form beliefs based on public speech, it might alter the epistemic pay-off of inclusive speech by generating costs as well as benefits, but by destroying the benefits as well. Either way, of course, things are bad.

That pervasive misinformation undermines the political argument, as well as the epistemic argument for inclusive communication is less obvious. But Lepoutre’s case that it does is persuasive:

Take Obama’s claim that the Affordable Healthcare Act would allow voters who liked their current healthcare plan to keep it. Assuming the ACA was a good policy, this use of misinformation might get people to accept a better policy than they otherwise would have. Yet it

remains ...problematic. By inducing false beliefs in voters, this piece of misinformation erodes their control over political decisions. Insofar as they are misinformed about what the ACA consists in, they are not in a position to ensure that it tracks their concerns (Lepoutre, 2021: 109).

As with hate speech, attempting to counter misinformation with more speech can seem an unpromising strategy. Misinformation is ‘sticky’: it often continues to affect listeners even after the misinformation has been corrected:

Social psychologists have widely reported a ‘continued influence effect’, whereby attempting verbally to correct falsehoods is ineffective or worse. In one study, for example, Center for Disease Control flyers distinguishing myths from facts about vaccines have been shown to backfire: people who read the flyer end up being more likely to misidentify myths as facts – and to oppose vaccination – than people who do not. Similarly, Adam Berinsky finds that although rehearsing and correcting the ACA death panel rumours ...initially makes people somewhat more likely to reject those rumours, this effect largely disappears after a few weeks (Lepoutre, 2021: 116).

Lepoutre argues that when counterspeech is appropriately delivered, it can overcome the problem of stickiness. Lepoutre argues that counterspeech should be both *positive* and *diachronic*. The idea of positive counterspeech should by now be familiar from the above discussion of hate speech. Positive counterspeech does not simply negate or deny false claims. Where negative counterspeech would respond to a leaflet that proclaims ‘vaccines cause autism!’ with the claim that vaccines do not cause autism, positive counterspeech would positively affirm the safety of vaccines. Positive counterspeech would respond by saying, ‘vaccines are safe’, rather than by re-articulating the link between vaccines and autism. As Lepoutre points out, negative counterspeech is at present the dominant paradigm of counterspeech. If alternative paradigms offer more prospect of success, it is crucial that our modes of counterspeech shift away from the negative paradigm.

The idea of diachronic counterspeech needs a bit more introduction. At a high level of abstraction, the idea of diachronic counterspeech is an attempt to shift away from a paradigm on which we think of counterspeech as an antidote, and towards a paradigm on which we treat it more like an ongoing form of inoculation. Lepoutre makes two main suggestions as to how we might affect such a shift. First, we should prioritise the wide circulation of politically important facts, making it more difficult for false claims to later ‘gain a foothold’ in the conversation. (No one, surely, could object to this.) The second suggestion is that we should:

warn the public about untrustworthy sources. This might involve identifying and exposing particular sources that are known to be untrustworthy. Or, alternatively, it might involve informing the public that many sources are untrustworthy, and giving them information that helps them distinguish trustworthy sources from untrustworthy ones (Lepoutre, 2021: 123).

Whilst I acknowledge the attraction of such ‘credentialising’ suggestions, they need to be handled very carefully. If implemented in certain ways, they might rob counterspeech

of its moral advantage over criminalisation. One important moral advantage counterspeech is supposed to have over criminalisation strategies is that the former does not impair citizens' ability to use speech to resist domination. But if inclusive public speech is to serve as a tool to guard against domination, then it is crucial that small, grass-roots news organisations be able to flourish. Such news organisations play a crucial role in articulating problems and concerns that may be invisible to those with power and influence within traditional news organisations. (For example, the problems facing renters are likely to be highly salient to young journalists, but far less salient to those with more power within journalism, who are likely home owners.) Credentialising approaches to managing credibility which stress the authority of established, traditional news media risk robbing more makeshift news organisations of the credibility they need if they are to play their crucial democratic function. If speech is to guard against domination, citizens need more than to be able to produce certain mere 'locutionary' acts (Austin, 1975; Langton, 1993). They need authority. Lepoutre could, here, find himself caught in a pincer movement. If the counterspeech strategies he recommends are to be robust enough to avoid the charge of ineffectiveness, they risk sanctioning state management of credibility to such a degree that counterspeech risks losing at least some of its moral advantage over strategies of criminalisation.

But let's bracket such worries of implementation. As in the case of hate-speech, I'm highly sympathetic to the overall approach. Lepoutre's insistence on developing and examining the most sophisticated and empirically informed counterspeech strategies, rather relying on a caricature of its possibilities, is admirable. I think he is right that, if imaginatively developed and carefully deployed, counterspeech *may* be able to salvage the inclusive ideal from the threats posed to it by misinformation. But as in the case of hate speech, I have some hesitations about the details of the picture he proposes.

To explain the stickiness of misinformation, Lepoutre once again appeals to the ideology of salience. This time round, though, the ideology of salience is supplemented with an appeal to the fluency heuristic. Fluently processed information is more likely to be accepted as true than information that is not fluently processed. A statement printed in high colour contrast, or presented in rhyming form, or a familiar accent, is more likely to be accepted, respectively, than one printed in low colour contrast, presented in a non-rhyming form, or an unfamiliar accent (Lewandowsky et al., 2012). Lepoutre posits that salience and the fluency heuristic lock together so as to make misinformation sticky. The resulting picture goes something like this:

Suppose that lots of people are saying 'vaccines cause autism'. If I try to counter this misinformation by saying 'it is not the case that vaccines cause autism', then my counterspeech, like the initial speech I am trying to counter, makes salient the association of the concept *vaccines* with the concept *autism*. But when such an association is salient, the claim that vaccines cause autism becomes easier to process: the salience of the association increases the fluency with which we process the claim. But then the fluency heuristic will kick in: the more fluently information is processed, the more likely we are to accept that information as true (Lewandowsky et al., 2012). Hence upping the salience of the vaccine-autism association risks making audiences more vulnerable to misinformation in the long run, by increasing the fluency with which information is processed.

Positive counterspeech, by contrast – in this context, something like ‘vaccines are safe’ – would not make misinformation easier to process (unless, of course, the dastardly misinformants switch messages, and start asserting ‘vaccines are *not safe*’).

As in the case of hate speech, Lepoutre’s recommendation – that we switch to positive counterspeech – depends on a particular diagnosis of why negative counterspeech fails. The normative recommendation is underwritten by an empirical model. But as in the case of hate speech, it’s not clear that Lepoutre’s empirical model is the right one. No empirical work is cited in support of the posited connection between salience and fluency. And alternative empirical models are available. Consider, for example, the ‘loose label’ model Lewandowsky et al., 2012. On the loose label, claims like ‘vaccines don’t cause autism’ gets encoded something like this: there is a positive claim – that vaccines cause autism – and this positive claim gets cognitively ‘tagged’ with a negation. But just as luggage may lose its labels in transit, so cognitively encoded claims may lose their negation-tags as time passes Lewandowsky et al., 2012.

The loose label model makes recommendations similar to those made by the salience model. If, rather than saying ‘vaccines do not cause autism’, we say ‘vaccines are safe’, we utter a sentence which is likely to be encoded without a negation tag, and so we utter a sentence whose full content is more likely to be recalled. Hence, despite positing very different underlying explanatory mechanisms, the differences between these mechanisms largely washes out at the practical level.

There’s another point of overlap between the two models. Both the salience and the loose label model are *irrationalist* models. They fit into what Kevin Dorst calls ‘the irrationalist narrative’ – a systematic picture of the world which aims to theorise political dysfunction in terms of rational dysfunction (Dorst, 2020). Politics goes wrong, says the irrationalist narrative, because we suffer from a host of cognitive biases, process information in partial and partisan ways, and transform what should be dispassionate cognition into a vehicle for self-protection.<sup>12</sup> But, as Dorst argues, the irrationalist narrative often relies on a highly oversimplified model of rational cognition.

Lepoutre’s book has an ambivalent relationship to the irrationalist narrative. Whilst his chapters on hate speech and misinformation – my focus here – borrow many of its tropes, elsewhere, the spirit of the book is very different. In his chapter on group cognition, for example, Lepoutre takes up a phenomenon which is often placed right at the centre of the irrationalist’s narrative: the fact that group membership influences how group members think about political issues. Whilst this seems to many like an obvious case of irrationality, Lepoutre argues that it is not so. Once we move away from crude models of rational belief, he says, we can see that group cognition may be an *exercise of* rationality, rather than a deviation from it. One might then wonder: can a ‘rationalist’ paradigm explain the failures of negative counterspeech?

I think we can. Let’s start by looking carefully at the evidence for the so-called ‘continued influence effect’. In one oft-deployed experimental paradigm, participants are presented with a report. The report contains a target piece of information. For some readers, this target information is subsequently retracted. For readers in a control condition, no correction occurs. Participants’ understanding of the event is then assessed (Lewandowsky et al., 2012). Lewandowsky et al. note that ‘research using this paradigm has consistently found that retractions rarely, if ever, have the intended effect of eliminating reliance on

misinformation, even when people believe, understand, and later remember the retraction' (Lewandowsky et al., 2012). A commonly used stimulus narrative involves a warehouse fire initially thought to have been caused by gas cylinders and oil paints that were negligently stored in a closet. The information that gas cylinders and oil paints that were negligently stored in a closet is subsequently retracted (Lewandowsky et al., 2012). Suppose that Sheila finds herself confronted with this sort of informational pattern:

Warehouse Fire. Sheila is told that a warehouse burnt down. Initially, on Tuesday, she is told, by someone she considers reliable that the factory burnt down because an employee left flammable materials in the smoking area. Then on Wednesday, she is told by someone else she considers reliable, that in fact no employee left out flammable materials. Initially, Sheila seems to accept this testimony. But then when she is asked on Friday how the warehouse burnt down, Sheila says 'an employee left flammable materials in the smoking area'.

Is Sheila being irrational? Proponents of the irrationalist narrative would say: 'yes'. It seems like Sheila has irrationally reverted to an earlier picture of the world, one she endorsed before she had all the evidence (viz., before she heard the retraction). And that seems bad.

But this is too quick. It relies on a myopic view as to what Sheila's total evidence consists in. It focuses only on her *testimonial* evidence, and ignores her *abductive* evidence. The irrationalist is right that Sheila's testimonial evidence looks to require her to suspend judgement as to whether an employee left flammable materials in the smoking area. But Sheila's evidence is not exhausted by her testimonial evidence. She also has abductive evidence. After all, Sheila is pretty sure that the warehouse burnt down. And she is aware of just one substantive hypothesis for why this happened: that the warehouse burnt down because an employee left inflammable materials in the break room. Let's call that hypothesis '*H*'. The hypothesis that an employee left flammable materials in the break room is a pretty good explanation for why the warehouse burnt down. The alternative – that an employee did no such thing – is a terrible explanation for why the warehouse burnt down.<sup>13</sup> That provides Sheila with some evidence – abductive evidence – for *H*. Now suppose, further, that Sheila can't think of any hypothesis *other than H* which does a decent job of explaining why the warehouse burnt down. Then the only way for her to make sense of the fact that the warehouse burnt down is to endorse *H*. But under those circumstances, it starts to look rationally permissible for Sheila to endorse *H even given the retraction*.<sup>14</sup> We can bring this out most clearly by comparing **Warehouse Fire** with a second example:

Mouse Droppings. Carys sees what she is fairly certain are mouse droppings in her kitchen one morning. She has two housemates, Bill and Ted. Carys regards both Bill and Ted as generally trustworthy and competent to roughly the same degree. Bill tells her that there are mice in the house. Ted tells her that there are definitely no mice in the house.

Under such circumstances, Carys should be pretty confident that there are mice in the house. And Carys' situation with respect to the claim that there are mice in the house is



rather like Sheila's situation with respect to the claim that an employee left flammable materials in the smoking area. Like Sheila, Carys has conflicting testimonial evidence. But, like Sheila, Carys' testimonial evidence is not all the evidence she has. Carys also has abductive evidence – she can see what she is fairly sure are mouse droppings on her kitchen floor. If there are no mice in the house, this is hard to explain. If there are mice in the house, on the other hand, it is easy to explain why she can see the apparent mouse droppings. If Carys engages in inference to the best explanation – in general, a perfectly respectable way to form beliefs – she will probably end up believing that there are mice in her house. If Ted wants to rationally persuade Carys that there are no mice in the house, he had better provide Carys with some alternative, non-mousey explanation for why there are apparent mouse droppings on the floor. (Perhaps 'Those are actually rat droppings, which are easily mistaken for mouse droppings. We have rats, not mice', or 'Those are fake mouse droppings that I bought in a joke shop; I put them there to fool you and Bill'.)

More picturesquely, whether it's rational for *A* to accept some claim *p* often depends on more than *A*'s testimonial evidence. It will often depend on the role *p* plays in *A*'s broader cognitive eco-system. If someone has a false belief that *p* that plays an important role in their cognitive eco-system, you can't always make it rational to give up on that belief just by telling them *p* is false, however authoritative a speaker you are. You will have to figure out what role *p* is playing in their overall cognitive eco-system, and provide them with some alternative that can play that role. When counterspeech threatens to disrupt a cognitive eco-system, it needs, if it is to be effective, to provide its audiences with the material to *repair* those disrupted cognitive eco-systems.

With respect to Sheila, the upshot is this. If Sheila is broadly rational, then to get her to stop believing *H*, we need to provide her with an alternative hypothesis – a hypothesis *H'* which does at least as good a job as *H* when it comes to explaining why the warehouse burnt down. (Perhaps 'The fire was deliberately started by a malicious former employee'.)

Is my rationalist model just wishful thinking? It seems not. Compared to 'bare' retractions, retractions supplemented with an 'alternative story' reduce hearers' reliance on the retracted content (Lewandowsky et al., 2012). That's because our beliefs are enmeshed in complex explanatory webs. Whether it's rational to give up on a belief often depends on how its removal would affect the rest of that web.

This suggests an account of why negative counterspeech fails that does not appeal to the ideology of salience. Negative counterspeech fails because it asks hearers to excise certain beliefs, without giving them any tools with which to repair their ruptured doxastic webs. Unlike the salience-based diagnosis, such a diagnosis does not motivate a pivot to positive counterspeech. Positive counterspeech provides no more reparative tools than negative counterspeech. To persuade rational listeners we need *reparative* counterspeech. Reparative counterspeech attempts not only to offer its addressees truths in the place of falsehoods, but also (i) explains why they were presented with the falsehoods in the first place, and (ii) models how to repair their webs of belief once they excise the falsehoods from them.

What would this mean in a real world case? Consider the situation of Polly. Polly is embedded in an anti-vax socio-epistemic network. The people she trusts have all begun to

tell her that vaccines dangerous. Negative counterspeech would mean telling Polly that vaccines are not dangerous. Positive counterspeech would mean telling Polly that vaccines are safe. Both positive and negative counterspeech will leave Polly with a puzzle. The people she trusts have been telling her that vaccines are dangerous. Coming to believe that vaccines are safe would mean coming to believe that the people she trusts have been telling her falsehoods. But why would the people she trusts be telling her false things? That's going to be hard for Polly to make sense of. Reparative counterspeech will help Polly to make sense of it. Reparative counterspeech will not only affirm the safety of vaccines but will also offer Polly an *explanation* for why her network has been feeding her false information. (Perhaps 'Misinformation about vaccines has become widespread. Many well meaning people have been deceived'.)

Thus, unlike the loose label model, my rationalist model offers different practical recommendations to those issued by Lepoutre's preferred salience model. When it comes to the stickiness of misinformation, then, a range of different mechanisms that might be responsible. These competing models offer different, and sometimes competing, prescriptions for those of us who want to make counterspeech as effective as possible.

This might sound broadly optimistic: we just need enough data to figure out which of the models is best, and work out a template for counterspeech based on that data. Such optimism, whilst tempting, might be too quick. The plurality of available models raises something of a meta-level problem for counterspeech strategies. It seems unlikely that any one of the models sketched above is completely accurate, and all the others completely wrong. Rather, it seems likely that the best explanation of phenomena like stickiness will be multi-valent, and shift in emphasis from context to context. It follows that the most appropriate form for counterspeech to take will also likely vary from context to context: there will be no fixed template for counterspeech to follow. Perhaps there is nonetheless some sort of best-fit template for counterspeech that will work in a sufficiently wide range of contexts to count as 'good enough'. Perhaps we might, with enough data and imagination, fasten on it. But it may also turn out that the cognitive and epistemic mechanisms responsible for making misinformation sticky are too unruly to inform any kind of operationalisable policy, with the right blend of power and scalability, to meaningfully protect against misinformation. And so pessimism about counterspeech may be forced on us after all.

## Conclusion

For better or worse, we cannot avoid talking to each other. The inclusive ideal transforms this necessity into a virtue. Its powerful vision of public life has come under increasing pressure as philosophers grapple with the nature of hate speech, theorise the disintegration of traditional media, and digest a literature that suggests citizens are incorrigibly biased. Lepoutre's defence of the inclusive ideal seeks to show not just that the ideal can absorb these challenges, but that the practice of inclusive public speech has tools with which such challenges may be addressed. I hope he's right.

Lepoutre theorises both hate speech and misinformation in terms of a particular psychological paradigm: salience. This underpins both his diagnosis of why negative counterspeech is often ineffective and his recommendations as to how counterspeech might be

made more effective. Whilst I agree with Lepoutre that opponents of counterspeech have often been too quick to reject the strategy, I am sceptical that the salience paradigm offers a promising account of why counterspeech fails or how to improve its prospects. The salience paradigm is too coarse a tool with which to adequately theorise the subtle dynamics of public speech. Once we move away from the salience paradigm, though, and towards alternative frameworks – such as, for example, the questions under discussion framework – we get rather different prescriptions from those Lepoutre favours. If we are lucky, then careful enough thinking may deliver us a model of counterspeech which can be easily exported and applied across contexts. If we are unlucky, though, we may not be able to tell any unified storey about the dynamics of either hate speech or misinformation. The phenomena may turn out to be too deeply shifty and fragmented for us to develop any template for successful counterspeech.

### **Acknowledgements**

Thanks to Maxime Lepoutre, Neil Levy, Mona Simion, Andrew Williams, and two anonymous reviewers for helpful comments. Thanks also to audiences in Berlin and Oxford.


### **Declaration of conflicting interests**

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### **Funding**

The author received no financial support for the research, authorship, and/or publication of this article.

### **ORCID iD**

Rachel Fraser  <https://orcid.org/0000-0003-2535-0608>

### **Notes**

1. This is a slightly more circumspect formulation of the ‘anti-domination’ argument than Lepoutre sometimes gives; occasionally he writes as though proponents of such arguments take inclusive political communication to be a sufficient condition for the absence of domination, rather than one important tool in the democrat’s toolbox.
2. Thanks to an anonymous reviewer for pressing me on this point; the figure of the ‘value-neutral sociologist’ is their suggestion.
3. Thanks to an anonymous reviewer for pressing me to discuss this possible tension.
4. I will use ‘human’ and ‘person’ interchangeably here, ignoring the complications posed by non-human persons.
5. Many actual instances of hate speech will mix together inferiorising and expulsive strains. At an empirical level, the phenomena are thoroughly mixed. They are nonetheless analytically separable.

6. One might try something like: 'Dignity requires that an individual not be disadvantaged with respect to their membership within a political community on grounds of some un-chosen social identity'. But that doesn't seem right. Suppose some country ends up with a skewed gender ratio: there are four times as many women as there are men between the ages of 18 and 45. It seems legitimate – and not offensive to the dignity of female would-be migrants – for such a country to prioritise male migrants.
7. Lepoutre here articulates a line of argument made most forcefully by Mary-Kate McGowan (McGowan, 2009, 2019).
8. Maybe that is how stereotypes work! Whether they do or not is not the sort of thing that can be settled *a priori*, after all. I am articulating an explanatory challenge for, not saying that his view is obviously false.
9. See Benz and Jasinskaja, 2017 for an overview.
10. One might think: why go in for positive counterspeech at all, rather than sticking to sarcastic counterspeech? Sarcastic counterspeech is really only a viable strategy for speakers with the option to make direct retorts to those engaging in hate speech; it's hard to see how the strategy could be exploited in more distributed conversational environments in which participants are seldom directly responding to each other.
11. One option which Lepoutre does not consider is what we might call *undermining* counter-speech: speech which undermines the credentials of the hate speaker; for example, 'You've clearly no idea what you're talking about'.
12. A complication: one might argue that it is rational to rely on the fluency heuristic, insofar as fluency is a good proxy for familiarity, and familiarity a decent proxy for truth. See Levy, 2021 for arguments with this flavour.
13. In the sense that it does not raise the probability of the explanandum.
14. For Sheila's high confidence in *H* to be high, she must be fairly confident that *if there were some other adequately explanatory hypothesis, she would be able to think of it*. Maybe Sheila should not be confident in this regard.

## References

- Austin JL (1975) *How to Do Things with Words*. Oxford: Oxford University Press.
- Benz A and Jasinskaja K (2017) Questions under discussion: From sentence to discourse, 2017.
- Brennan J (2016) *Against Democracy*. Princeton: Princeton University Press.
- Darwall SL (1977) Two kinds of respect. *Ethics* 88(1): 36–49.
- Dorst K (2020) The rational question. *The Oxonian Review*.
- Friedman J (2019) Inquiry and belief. *Noûs* 53(2): 296–315.
- Glick P, Fiske ST, Mladinic A, et al. (2000) Beyond prejudice as simple antipathy: hostile and benevolent sexism across cultures. *Journal of Personality and Social Psychology* 79(5): 763.
- Johnson GM (2020) The structure of bias. *Mind; A Quarterly Review of Psychology and Philosophy* 129(516): 1193–1236.
- Langton R (1993) Speech acts and unspeakable acts. *Philosophy & Public Affairs* 22(4): 293–330.
- Lepoutre M (2021) *Democratic Speech in Divided Times*. Oxford: Oxford University Press.
- Levy N (2015) Neither fish nor fowl: implicit attitudes as patchy endorsements. *Noûs* 49(4): 800–823.
- Levy N (2021) *Bad Beliefs: Why they Happen to Good People*. Oxford: Oxford University Press.

- Lewandowsky S, Ecker UK, Seifert CM, et al. (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. In: *Semantics from different points of view*. Springer, pp. 172–187.
- Manne K (2016) Humanism: a critique. *Social Theory and Practice* 42(2): 389–415.
- McGowan MK (2009) Oppressive speech. *Australasian Journal of Philosophy* 87(3): 389–407.
- McGowan MK (2019) *Just Words: on Speech and Hidden Harm*. Oxford: Oxford University Press.
- Waldron J (2012) *The Harm in Hate Speech*. Cambridge, Massachusetts: Harvard University Press.

### Author biography

**Rachel Fraser** is an Associate Professor of Philosophy at the University of Oxford, and a Tutorial Fellow in Philosophy at Exeter College.