Prejudicial Speech: What's a Liberal to Do?

MARI MIKKOLA

Abstract
This paper discusses potential responses to harmful prejudicial speech. More specifically, it considers how different types of prejudicial speech merit different responses. The paper distinguishes hate speech, discriminatory speech, and toxic speech as different *types* of speech that are prejudicial or oppressive – they are not of the same kind diverging only in their severity and explicitness. As these sorts of problematic speech are categorially distinct, the paper holds, they also demand differential remedies. The task of this paper is to consider such remedies, their potential effectiveness, and compatibility with the liberal value of free speech.

1. Introduction

Prejudicial and intemperate speech is a thorny issue in liberal societies. Freedom of speech is a central liberal value. But, if left unchecked, it permits morally and socially undesirable expressions. This raises important questions about the limits of free speech: what (if anything) should be done about prejudicial speech?[1] For a start, it is worth remembering that no legal free speech principle *de facto* admits all forms of speech. Verbal threats are not defensible on free speech grounds and outlawing them raises no free speech concerns whatsoever (they constitute crimes). Defamation is legitimately actionable by tort law because it incurs serious harms to the defamed that outweigh significant and compelling free speech interests. Still, a large of bulk of speech (at least in the USA) is afforded protected status: it is judged to advance substantial free speech interests that justify non-restriction and non-interference, even if the speech incurs some harms and/or is offensive. The interests on which the value of free speech is typically taken to hinge in philosophical discussions include: the pursuit of truth and knowledge, ensuring democratic deliberation and functioning, and fostering personal autonomy and individual progress. But do these interests justify

[1] Note that 'speech' here is taken to denote not just spoken words and utterances, but also signs, public recordings, written words, non-verbal symbols, and other means of expression.

doi:10.1017/S1358246124000067

non-interference in and a laissez-faire attitude toward prejudicial speech? Must we simply accept the harms of intemperate speech within a liberal framework as the price we pay for broader freedoms? Should 'our' response simply be to tolerate the intolerable, and to fight speech with more speech in the marketplace of ideas? If not, what (kinds of) governmental or legal interventions on speech might still be unacceptably illiberal?

My view is that answers to these questions depend on the *type* of prejudicial speech in question. I will here focus on hate, discriminatory, and toxic speech. The position that I hold in appealing to different types of speech goes against prominent current views about (what I call) 'prejudicial speech' and (others call) 'hate speech'. In recent years, both in academic philosophy and in public discourse, conceptions of *hate speech* have proliferated, and the conceptual terrain has become hard to navigate. In two high-profile papers, Alexander Brown (2017a, 2017b) considers a staggering quantity of academic and public literature, and concludes that the concept of *hate speech* encompasses:

> a family of different purposes including but not limited to highlighting forms of harmful speech, flagging up socially divisive forms of speech, identifying forms of speech that can undermine people's sense of equality, articulating civility norms, and labelling forms of speech that undermine democracy; a family of types of speech including but not limited to insults, slurs, and epithets, words that express or articulate ideas relating to the moral inferiority, group defamation, and negative stereotypes or generics; a family of types of speech act including but not limited to insulting, disparaging, degrading, humiliating, misrecognising, disheartening, harassing, persecuting, threatening, provoking, inciting hatred, discrimination or violence, and justifying or glorifying discrimination or violence; and a family of characteristics including but not limited to race, ethnicity, nationality, citizenship, status, religion, sexual orientation, gender identity, and disability. (Brown, 2017b, p. 600)

Given this proliferation of what 'we' mean by *hate speech*, some philosophers (Brown included) have begun refining the definition of *hate speech* by expanding its scope to catch as many expressions as possible. These (in my terms) 'expanded' strategies are increasingly popular in philosophical work. Brown along with Katharine Gelber (2021) are two prominent recent advocates in print. My contention, however, is that we should eschew the expanded strategy and embrace a pluralist strategy (I argue for this view in more detail elsewhere, see Mikkola, forthcoming 2024). 'Prejudicial speech' – for me

– is an umbrella term that encompasses different types of intemperate speech ranging from: hate (narrowly understood) to discriminatory (more broadly conceived) to toxic (diffuse and amorphous) to some still other type(s) of speech. Importantly for my typology, this division isn't about the seriousness or harmfulness of speech, with hate speech being the most serious kind. Rather, it is about scope, palpability, and definiteness, which makes a difference to 'our' reactions to prejudicial speech. With the pluralist approach, I hold, we can forge more nuanced responses and avoid getting bogged down by an unhelpful false binary of do nothing or enforce draconian restrictions.

My task in this paper is to motivate the idea that there are different types of prejudicial speech that merit different responses within a liberal framework. There is no conception of prejudicial or hate speech supposedly carved in nature's joints – it is up to us to define them in a beneficial and fitting manner. Moreover, I hold that interventions, and even restrictions, on speech are compatible with liberal commitments to free speech. But the way this is done must be relative to the type of speech in question; as I see it, the prospect of a blanket response to prejudicial speech is extremely poor.

To make my case, I will first provide some conceptual mapping of different kinds of prejudicial speech. I will then consider what is the harm of these kinds of speech. Finally, I will consider what can be done about prejudicial speech while respecting free speech rights.[2]

2. Prejudicial Speech: Initial Conceptual Mapping

In characterising (then) recent work on hate speech, Stanley Fish (2012) notes that although everyone agrees words used directly to incite violence against a person or a group count as hate speech, there is much disagreement over the remaining (in his view) hard cases. Fish's apparently uncontentious type of hate speech mirrors the well-known US case of *Chaplinsky v. New Hampshire* (1942) that took a particular narrow class of speech to be obviously punishable without raising any free speech concerns (albeit not a case that Fish explicitly cites). This included 'the lewd and the obscene, the profane, the libelous [sic], and the insulting or "fighting" words—those which by their very utterance inflict injury or tend to incite

[2] Given the topic of this paper, I will discuss some example cases of prejudicial speech though I won't be mentioning (let alone using) slurs or hateful epithets.

an immediate breach of the peace'.[3] In the spirit of *Chaplinsky v. New Hampshire,* David Brink develops an understanding of *hate speech* that is a narrow one. On this view, hate speech is about prohibited harassment by personal vilification, and an expression counts as such if and only if

(a) it employs fighting words or non-verbal symbols that insult or stigmatize persons on the basis of their gender, race, color, handicap, religion, sexual orientation, or national and ethnic origin;
(b) it is addressed to a captive audience [when the speech is relatively difficult to avoid];
(c) the insult or stigma would be experienced by a reasonable person in those circumstances; and
(d) it would be reasonable for the speaker to foresee that his words would have these effects on a reasonable person in those circumstances. (Brink, 2001, p. 135)

This type of speech comes apart from other types of prejudicial speech, which are less clear cut. Discriminatory speech mirrors or reflects group stereotypes and represents groups or their members as inferior by virtue of these stereotypes (p. 133). Nonetheless, it does so in putatively non-vilifying and non-invidious ways. Although discriminatory speech may be and is odious and offensive, it is said to deserve free speech protections: it may nonetheless serve a socially valuable function by contributing to public debates and by enhancing deliberative practices insofar as we can challenge the speaker and the underlying stereotypes being espoused. Examples of this type of speech might be:

- Politician in an interview endorsing bio-behavioural or biologically essentialist views to explain racialised patters of criminality.
- Professor holding that well-meaning but misguided affirmative action policies are bringing undertalented students from marginalised backgrounds to university, since such students just are 'by nature' unsuited for university education.
- Manager arguing in a meeting against promoting female candidates on account that they will be more interested in having babies than in having careers.

More recently still, there has been a prominent focus on toxic speech. Lynne Tirrell (2021) takes such speech to include derogations, epithets, and slurs. But it involves more: it is

[3] 315 U.S. 568 (1942), https://supreme.justia.com/cases/federal/us/315/568/. Accessed 23 August 2023.

> a broad and mercurial category, […] [also including] speech that acts more chronically by gaslighting, undermining, threatening, and more […]. Some toxic speech surreptitiously re-orients people away from their settled values and conceptions of the good. (Tirrell, 2021, p. 116)

Toxic speech is broadly understood and denotes a diffuse kind of speech that is amorphous and undermines what is good for individuals. Some examples I have previously used to illustrate this type of speech include (see Mikkola, 2021):

The Brexit-Bus:
bright red *Leave* campaign bus stating 'We send the EU 350 million pounds a week, let's fund our NHS (National Health Service) instead'.

Trump 'the Winner':
stating in an interview when facing electoral defeat in November 2020 that 'This is a fraud on the American public. This is an embarrassment to our country. We were getting ready to win this election. Frankly, we did win this election. So our goal now is to ensure the integrity – for the good of this nation, this is a very big moment – this is a major fraud on our nation'. (The same sentiments of winning and the election being fraudulently stolen have of course been expressed many times in the years that have followed.)

Corona = Agenda 21:
during a 1992 UN Conference on Environment and Development, 177 national leaders (including George Bush Sr.) signed a non-binding statement of intent aiming to take action in order to ensure sustainability given population growth. This agreement, known as Agenda 21, has been dubbed by alt-right and political extremists as a secret plot to impose a totalitarian world order in a nefarious effort to use environmentalism as a means to crush freedom. In late-2020, groups protesting against restrictions brought on by the COVID-19 pandemic carrying signs stating 'Corona = Agenda 21' were seen (at least) in Germany, Switzerland, and the Netherlands. Their message is, in short, that the corona pandemic is used by global elites to annihilate people's freedoms and to reduce the world's population to advance the elite's iniquitous ends.[4]

[4] Whether the protesters thought that the pandemic is a hoax used for these ends or whether the real pandemic was used for nefarious ends isn't entirely clear. But this does not make a substantive difference to my discussion of the example.

This typology then understands prejudicial speech to range from hate speech (narrowly understood) to discriminatory speech (more broadly conceived) to toxic speech (that is diffuse and amorphous).

3. What's the Harm?

Legal free speech principles do not *de facto* typically admit all forms of speech despite liberal commitments to free expression. As Fish (1993) puts this elsewhere provocatively in a book title, *There's No Such Thing as Free Speech: And It's a Good Thing, Too*. Standardly, free speech interventions are justified, but only if the harms caused and/or constituted by some speech outweigh significant free speech interests. Hence, in thinking about the types of prejudicial speech outlined above and appropriate responses to them, the first step is to assess what those harms are and their seriousness. Initial pre-theoretical considerations might suggest the following analysis of the harms. Hate speech looks to be harmful in being offensive and insulting. Discriminatory speech, then again, is stereotyping and misleading, thereby generating harms to its recipients. And finally, toxic speech is downright false, even if non-offensive, but this suffices for it being harmful (enough). On standard liberal grounds following J.S. Mill, though, these harms can be mitigated and hence do not warrant intervention. As is well known, offensiveness is not a harm in the right kind of way according to Mill (just think of his discussion of corn dealers). Misleading and stereotyping statements can be valuable even when odious: by challenging the speaker, we can debunk the views expressed, and thus advance democratic deliberative practices. On Millian grounds, even false statements have social value in advancing truth and knowledge seeking.

Nonetheless, Mill holds that autonomous functioning and democratic self-governance demand the exercise of deliberative capacities that only develop in free societies: 'observation to see, reasoning and judgement to foresee, activity to gather materials for decision, discrimination to decide, and [...] firmness and self-control to hold his deliberate decision' (1974, p. 139). As I see it, the three types of speech under examination undermine precisely the development and/or exercise of such deliberative capacities. Importantly though, they do so in different ways; therefore, we need different sorts of responses to different types of prejudicial speech.

Let's start with hate speech. Brink takes hate speech in the narrow sense to be harmful in hampering deliberative practices: it limits its

recipients' participation in deliberative exchanges and prevents recipients from getting a fair hearing when they try to participate (2001, pp. 140–1). In this sense then, hate speech is akin to defamation, where free speech interests do not mitigate harms incurred – hate speech should not subsequently count as protected speech. Examples of these sorts of harms have been gleaned via interviews with targets of hate speech (Gelber and McNamara, 2016; see also Brink, 2001; West, 2012). They include: negative stereotyping, feelings of fear, existential pain, disempowerment, withdrawal from expressive opportunities, silencing, exclusion, dehumanisation, provocation to anger, restrictions on the ability to identify with one's (ethnic, national, racial) group. These harms seemingly conflict with typical free speech interests, and specifically with (a) ensuring democratic deliberation and functioning; and (b) fostering personal autonomy and individual progress. It is consequently not obvious that 'our' commitment to free speech on the whole mitigates the harms of hate speech.

Still, Brink holds that discriminatory speech is not harmful in the same manner as hate speech. The former supposedly leaves open the possibility of deliberation, open debate, and challenging of the speaker. Furthermore, Brink holds, even if 'merely discriminatory speech' has the effects of marginalising and silencing its recipients, narrowly understood hate speech is 'generally worse' and 'comparatively easy to identify' (2001, p. 148). I agree that hate speech in the narrow sense is easier to identify than more diffuse discriminatory speech. But I disagree that it is generally worse since (I hold) discriminatory speech is arguably more prominent and widespread. This being the case, I argue elsewhere (2019) that discriminatory speech is particularly detrimental in eroding self-trust. Self-trust is an attitude we take toward ourselves because we have confidence in our epistemic abilities. As Elizabeth Fricker puts it:

> Each one of us in one's everyday life relies on one's core package of cognitive faculties – perception, proprioception, memory, intellectual intuition and introspection – reliably to deliver one true beliefs […]. The core phenomenon of epistemic self-trust consists in one's ungrounded reliance on one's cognitive faculties reliably to yield one true beliefs. (Fricker, 2016, p. 154)

And for Karen Jones,

> intellectual self-trust is an attitude of optimism about one's cognitive competence in a domain. Self-trust manifests itself in feelings of confidence, in dispositions willingly to rely on the deliverances

> of one's methods and to assert what is believed on their basis, and in modulating self-reflection […]. Developmentally, our intellectual self-trust is created interactively [in relation to our parents, teachers, peers]. (Jones, 2012, p. 245)

With these ideas in mind, think back to Mill's deliberative capacities noted above that seemingly appeal to self-trust: 'observation to see, reasoning and judgement to foresee, activity to gather materials for decision, discrimination to decide, and […] firmness and self-control to hold his deliberate decision' (1974, p. 139). Moreover, recall free speech interests that are taken to justify non-interference. Democratic functioning requires free speech to facilitate deliberative exchanges, and to enable informed political decision-making. And exercising autonomy involves certain competencies, for example, the ability to act according to one's own interests. My contention is that discriminatory speech erodes the kind of self-trust needed to support these free speech interests. It is in *this* sense harmful (though probably in other ways too). Still, this harm looks to be sufficient to undermine justifications for the permissibility of discriminatory speech. Self-reporting suggests that those subject to discriminatory speech are also marginalised, put off, and silenced by the negative stereotypes expounded in the absence of hateful personal vilification. Members of underrepresented groups are excluded from deliberative practices and ignored when they try to participate without anyone employing insulting fighting words or slurs. Non-hateful stereotypical speech may then also close off further rational exchanges, prevent challenging of the views expressed, and hamper democratic cultures. By way of example, I will consider espousing stereotypes about women in philosophy. Many such examples can be found in (the now defunct) blog *What's it like to be a woman in philosophy?* (Archive of posts still available online):

> In 2000 I [a female] was interviewing for jobs for the first time […]. I was sitting at the head of the table looking out at all the men – there was one female graduate student there, that's it. I finished my talk and the questions began. The professor who I would have been replacing raised his hand and said 'So […] we haven't had a woman teach fulltime in the department for 40 years, why should we hire one now?' Absolute silence, no one said a word. Rather than saying something clever like, 'you clearly shouldn't as you are not ready' and leaving the interview, I stammered something about perhaps this would help their enrolment, as I would have liked to have had a female role model

> when I was an undergrad. To this he replied 'Well, if we want to recruit more female students why shouldn't we just hire some hot, young guy?' I was totally flummoxed by this point and just trying not to a) yell or b) cry as I knew either of these actions would reinforce his ideas about women […]. NO ONE at the table said a word.
>
> At one of the first [graduate] seminars I went to, I was the only girl. I raise an objection. I'm told that I have misunderstood the point. I hadn't – the professor in charge of the seminar pointed this out twenty minutes later once all the boys had finally got round to saying what I said initially. I try to speak again later. My point is completely ignored. Two minutes later, a male makes exactly the same point. The objection in his mouth is hailed as decisive. I worry that my being dismissed and ignored is not because of my gender but because I am foolish; I worry that I don't love philosophy because almost every seminar I go to leaves me second guessing my ownbilityies.
>
> [During my first year] I made a concerted effort to participate and make at least one good comment or question in every meeting of the pro-seminar. However, at the end of the semester when we each got a report on how we did from the two (male) professors, this is what they wrote: '[Name] was sometimes a bit quiet, and we wondered whether she was a bit disengaged.' All the other people who were in that class who I told about this agree that, on the basis of my actual participation, this was unfair.

As I see it, in these examples non-vilifying discriminatory expressions have arguably also had the effect of silencing women and rendering their contributions invisible in a manner that generates self-doubt – that is, in a manner that engenders typical hallmarks of eroded intellectual self-trust.

What about toxic speech then? Earlier examples I noted were misleading, outright false, and/or without compelling justification: they are toxic in polluting our democratic milieu sometimes in very material and concrete ways. Lynne Tirrell (2021) has recently characterised toxic speech (in the sense outlined above) as being akin to a poison or virus, where propagating it can lead to an outbreak. Recall that it is 'a broad and mercurial category' and can reorient 'people away from their settled values and conceptions of the good' (Tirrell, 2021, p. 116). The toxic and poisonous effects of this sort of speech hinge on the endorsement of the speech. Tirrell understands endorsement generally to be 'automatic, not something added; it takes special care

to restrict endorsement when we must' (p. 125). Although I agree with Tirrell that toxic speech undermines what is good for individuals in perfidious ways, I think that the harm of toxic speech is to be understood differently: its contamination and contagiousness is more active than those of poisons and viruses. (I discuss this at greater length in Mikkola, 2021.)

Thinking about the sorts of examples I noted above suggests that people are actively keen to endorse misleading, false, and unjustified claims, even when they are vehemently and openly challenged. This sort of endorsement may happen 'on the spot'. But it isn't spontaneous in the sense of wholly lacking control and being without conscious thought or attention. To put it in a slogan form: endorsement does not happen to people; people make it happen. Subsequently, I hold, toxic speech corrupts and perverts well-functioning epistemic agency. This is the primary harm of toxic speech, which has many real-world secondary and material harms when people for instance act on some piece of toxic speech. Still, it is 'our' default underlying cognitive situation that enables this sort of corrupting. Our cognitive architecture is like untreated steel: without prevention and if exposed to both oxygen and water, steel will rust. Hence, we treat and coat it. When the coating is damaged, we repair and recoat it to prevent rusting. In a similar fashion, it seems to me, without a protective coating our cognitive architecture will rust too. Furthermore, prevalent and infamous toxic speech examples suggest that a sort of proper cognitive 'rust prevention' is lacking. The willingness to endorse and to embrace half-truths, falsehoods, and unjustified claims is suggestive of an underlying problem at the core of epistemic agency: it has been left exposed to the elements without sufficient protections. This is what ultimately enables the corruption of well-functioning epistemic agency.

4. What's a Liberal To Do?

To recap, I take it that hate speech is defined as directly vilifying personal harassment (for example, being subject to racist or sexist slurs). It is harmful in limiting and hampering democratic participation by often quite literally silencing and marginalising its recipients. Then again, discriminatory speech is more covert and non-vilifying but diminishing speech based on stereotyping. It is harmful in hampering the development and/or exercise of autonomy and individual progress by undercutting self-trust. Finally, toxic speech is a diffuse kind of problematic speech that is amorphous and undermines what is good for individuals. It is harmful in corrupting and

perverting well-functioning epistemic agency in a covert manner given 'our' default disposition to be corruptible given our cognitive architecture.

How then might we respond to these types of speech? One initial kneejerk reaction may be: *Do nothing!* If the harms of prejudicial speech are due to our cognitive architecture or inability to withstand offensive speech, there is nothing legally to be done – we are dealing with individual problems. On the other end of the spectrum, however, the reaction may be a call to: *impose draconian speech restrictions!* Since we are so prone to intemperate speech, speech should be heavily restricted and forcefully intervened in. Perhaps we can think of something less crude though. I will next consider some such interventions and even restrictions.

4.1 Hate Speech

To begin with, it is important to note that restrictions on 'our' freedoms are not *eo ipso* illiberal. There are restrictions on the purchase, distribution, and consumption of alcohol and cigarettes; there are plenty of restrictions on driving; we have employment laws to restrict our working lives (e.g., a compulsory retirement age in some jurisdictions, restrictions on how small or large employment contracts can be irrespective of individual wishes), just to name a few. So, there is no blanket liberal prohibition on legally intervening in 'our' freedoms. But, one might say, freedom of speech is different and a special case in being a basic right. Freedom of speech and expression is conceivably more important than 'our' freedom to purchase alcohol and cigarettes. Nonetheless, free speech interests do not justify a hands-off policy regarding hate speech in the narrow sense. First, the right to free speech isn't an absolute right to say whatever one wants. There is a difference between speech in the *ordinary* sense (whatever we utter, what comes out of our mouths) and speech in the *legal* sense that falls under a legal free speech principle. Hate speech is not obviously speech in the legal sense that deserves protection. The substantive question is whether it falls under the legal sense of *speech*; and understood as harassment by personal vilification, it arguably does not. Second, the right to free speech is framed in terms of advancing democratic functioning and fostering individual autonomy and progress. But evidence suggest that these are *not* advanced by hate speech; they are rather seriously stifled and undercut. There are then compelling grounds to think that harms caused are grave enough to warrant intervention. This of course still leaves

open what sorts of legal interventions to advance, and whether such interventions should fall under criminal or tort law (be considered a civil wrong). Let me say something about these options.

Criminal offenses need to satisfy the *mens rea* (guilty mind) and *actus reus* (guilty act) requirements. The precise formulation of these differs from one jurisdiction to the next, but to satisfy *mens rea* intentionality is key. For instance, one cannot accidentally rob a bank or commit murder. Unless intentionality can be sufficiently established, it is hard (even impossible) to convict someone of alleged criminal offenses. One way to establish intentionality relative to hate speech is to tie the definition of *hate speech* to the notion of *incitement*. By way of example: there is no formal definition of *hate speech* in International Human Rights Law and, therefore, most [UN] instruments refer to 'incitement to discrimination, hostility or violence' […], 'direct and public incitement to genocide'; and 'advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence' are strictly prohibited under International Law, as they are considered the 'severest forms of hate speech'.[5]

If hate speech turns on incitement, though, some apparently hard cases of hate speech raise no free speech concerns whatsoever. Inciting someone to commit a crime is in all jurisdictions that I know of itself a crime and appealing to free speech rights is not an adequate defence. For instance, inciting others to commit racist hate crimes with one's speech isn't defensible on free speech grounds or a matter of free speech. Moreover, there is nothing illiberal about criminalising incitement to commit hate crimes with one's speech. This again demonstrates that freedom of speech isn't about the freedom to utter whatever comes to mind. Having said that, I think it is a mistake to equate *hate speech* and *incitement*. There is (and should be) a substantive difference between *saying* something that counts as hate speech and *inciting* others to commit hate crimes with one's speech. The case for criminalising the latter is straightforward, but this is not so for the former. Be that as it may, my point is to highlight that there already are legal interventions on 'our' speech – so, a view that takes speech to be sacrosanct is misplaced from the start.

Perhaps, though, we should treat hate speech like defamation under tort law: as a written or oral statement that results in damage to a person's reputation. If a co-worker distributes lies about me

[5] https://www.un.org/en/hate-speech/united-nations-and-hate-speech/international-human-rights-law. Accessed 16 January 2024.

that harm my reputation due to which I fail to be promoted, I can seek legal recourse and damages from that co-worker for the harms incurred. Richard Delgado (1993) proposes an independent tort action for racial insults akin to defamation, where the notion of *racial insult* he employs is in line with Brink's definition of *hate speech*. One example of a US case that seemingly would fall under this tort that Delgado discusses is *Contreras v. Crown Zellerbach, Inc* (1977).[6] The Mexican-American plaintiff alleged that his fellow employees had subjected him to a campaign of racial abuse. Hence, he further alleged that he had suffered 'humiliation and embarrassment by reason of racial jokes, slurs and comments' (Delgado, 1993, p. 89). The plaintiff was further wrongfully accused of stealing the employer's property, which allegedly damaged his employment prospects and held him up to public ridicule. This case looks like a paradigm case of defamation, where racial identity and subsequent racial insult experienced are not merely additional and peripheral to the case, but are in fact at the centre of the harms generated and experienced.

An immediate objection to these sorts of torts, however, is that the tort of racial insult is illiberal since on classical liberal views, offensiveness is not a harm in the right kind of way. The insult may be odious, but something that we must apparently tolerate in a liberal society. However, this objection isn't compelling to me for reasons that are self-evident to people advancing restrictions on hate speech. Simply put: the tort isn't about offensiveness – the point is that hate speech is damaging in a very material sense by (for instance) depriving people of employment opportunities. 'Sticks and stones may break my bones. But words shall never hurt me' just is not true (see also Matsuda *et al.*, 1993; Maitra and McGowan, 2012). Of course, practically speaking, it may be difficult to measure the damages incurred. But the same is true of more 'straightforward' defamation cases, which makes them notoriously difficult to litigate. Still, legal scholars and practitioners do not seemingly hold that *therefore* the tort of defamation should be abolished. My underlying theoretical point here is that there is nothing *per se* illiberal about this sort of legal redress as we already have various similar torts. And again, it is important to stress the non-absolute nature of free speech rights: they cannot be appealed to successfully as defences in many tort actions.

[6] https://law.justia.com/cases/washington/supreme-court/1977/44623-1.html. Accessed 1 July 2023.

However, one might object further that the tort for racial insults is too draconian in being a form of punitive legal redress. Hence, it is not compatible with a liberal commitment to free speech – after all, punitive measures typically involve incarceration. I find this objection uncompelling as well, though. Advocating for punitive responses is not equivalent to advocating for imprisonment. There are many ways in which we punish offenders without imprisoning them (fines or community service, for instance). Perhaps even more radically, transformative justice models might be beneficial: this involves a process where the individuals involved are given the opportunity to address and repair the harm caused. Those affected recount how an act has affected them and what can be done to repair the harm. The perpetrator is then held accountable to the individual(s) affected by way of restitution. As with any form of punishment, there are no guarantees that the outcome will be successful. But my point is to highlight that much can be done about hate speech *without* invoking the bugbears of speech bans or censorship.

4.2 Toxic Speech

Given what I take to constitute the harm of toxic speech, it might look like a good idea to limit the spread of toxic misinformation to prevent it from corrupting and perverting epistemic agency. An example of this might be sensitive media policies that some social media platforms already have. This response does speak for some straightforward restrictions on what we find in the public domain – and frankly I have no objections to such restrictions. Still *de facto* this response will have limited scope due to the nature of the internet. It will also be difficult to decide what to restrict and how, given the diffuseness and indefiniteness of toxic speech. More plausibly, then, one might think that some educational efforts or cognitive therapy are needed to better protect epistemic agency from being perverted. The example of government healthy eating campaigns is akin to what I have in mind. Given our evolutionary past, it is 'our' default to consume calorific foods. When food was scarce and work involved physical strain, this 'toxicity' was dormant. But now, if left to our own devices without public health measures, many of us would be living deeply unhealthy lives. In a similar sense, to undercut the toxicity of some speech maybe we need to advance educational programs to cultivate epistemically virtuous agency.

This response requires that we identify which sorts of epistemic virtues ought to be cultivated – which, then again, hinges on the

kinds of vices involved. Epistemic vices make us bad thinkers and the corrupting influence of toxic speech is seemingly such a vice. Heather Battaly distinguishes three types of epistemic vice. First, vices may produce false beliefs, thus involving 'effects-vice'. Second, certain cognitive character traits (like closed-mindedness or intellectual arrogance) are exemplary of 'responsibilist-vice': it is a vice over which agents have cognitive control and if the agent does not work towards exercising such control, they will be blameworthy. Third, epistemic vice may involve bad epistemic motives like 'motives to believe whatever is easiest, or whatever preserves the status quo, or whatever makes one feel good, instead of motives for truth, knowledge, and understanding' (Battaly, 2017, p. 224) – thus involving 'personalist-vice'.

To undercut these vices, we need to cultivate and promote the requisite virtues. To undercut effect-vice, it seems that virtue-reliabilism is the answer. This is the view that epistemic virtues are reliable epistemic dispositions that produce fewer false beliefs and more true ones. Although this would undercut effect-vice, the concerns I have discussed in this paper raise doubts about whether we have such epistemic virtues at our disposal to begin with. After all, we look to be very susceptible to personalist-vices given how corruptible our 'hard-drives' appear to be. What we need instead is some form of virtue-responsibilism, whereby epistemic agents work to shape their cognitive traits in ways that render those traits more reliable (Montmarquet, 1992). Epistemically autonomous agency, then, involves taking responsibility and working towards undermining the influence of toxic speech (and requisite attitudes) that corrupt and pervert good epistemic functioning. I find this idea attractive, but have serious doubts about its efficacy. Those who are willing and able to take responsibility for their epistemic lives are already on board, so to speak, and already concerned about the state of their epistemic well-being. The difficult question is how can we persuade the non-believers (bluntly put) to get on board. I am increasingly pessimistic about effective interventions at a later stage after individual epistemic agency has formed. Early age interventions would be needed and, I expect, continuing educational efforts from early on can play a huge role. Of course, this suggestion presents several further challenges about the appropriate method, mode, and location for such educational efforts. Educational interventions in schools guided by national curricula may be and are seen by some as illiberal: it is not the role of the state to decide and dictate what kids get taught in schools, one might hold. Also, without a prior normative theory, it becomes difficult to distinguish 'good' from 'bad' curricula in a

non-ad hoc manner. Perhaps the government-dictated curriculum contains the right ingredients. But there is little or nothing to stop the curriculum from being directed at teaching something one may find undesirable. There is much dispute that would have to be settled about what schools can and should teach, how much influence parents have and should have, and whether curriculum design should be left to the state at all (just to name a few). Still, though we may disagree about the specifics, I see no reason to think that basic educational efforts to foster virtue-responsibilism and undermine epistemic vices are *per se* illiberal.

We should also bear in mind that toxic speech often functions like subliminal messaging akin to some forms of advertisement. (In fact, the success of advertising and marketing further demonstrates precisely how corruptible we are!) On Thomas Scanlon's prominent liberal view, it is legitimate to restrict subliminal advertising messages. Hence, one might think that it is legitimate and permissible to restrict toxic speech as well without compromising our commitment to free expression. Scanlon holds that subliminal messages interfere with audience autonomy in producing beliefs and desires that the audience has no control over. Subsequently, there is an interest 'in having a good environment for the formation of one's beliefs and desires' (Scanlon, 1978–9, p. 527). In other words, audiences have a positive autonomy interest in being free from manipulation. This interest (being free from manipulation) also *prima facie* justifies educational efforts to develop the sorts of capacities that enable us to exercise autonomy. I cannot say anything detailed here about the contents of such education, which of course raises complex practical questions. Still, yet again, there is nothing *per se* illiberal about educational efforts to engender conditions that enable us to develop and exercise autonomy-capacities and Millian deliberative capacities. In a similar fashion, there is nothing *per se* illiberal about state authorities providing nutritious school meals to undercut unhealthy eating habits. One might in fact think that this is precisely what they ought to do to enable good functioning of future citizens.

4.3 Discriminatory Speech

Let me now turn to discriminatory speech. Given that it is non-vilifying, there does not seem to be a good practicable way to legally proscribe or regulate such speech. My hope would be that the sort of cognitive therapy noted above can help undercut the prevalence and influence of discriminatory speech too. But apart from that,

must we simply fight it with counter-speech in the 'marketplace of ideas'? I take this proposal to be too simplistic and naïve given how speech seemingly affects us. In fact, the eroding effects of discriminatory speech undermine the idea of a self-regulating 'marketplace of ideas' wholesale. Still, since discriminatory speech turns on stereotypical attitudes and ascriptions, we can act against their influence on further attitudes and subsequent behaviour. Some social psychological research suggests that we can control stereotype activation with certain egalitarian goals (Kawakami *et al.*, 2000; Moskowitz and Li, 2011). With practice and over time, it is possible to develop an associative link between the goal to be egalitarian, and a specific target group. This is just *an* example, of course, but my general point is this: research on stereotype activation suggests that we can act against the influence of stereotypes and stereotyping. One way in which we can do so is by implementing better structural and organisational arrangements. Think about hiring and teaching practices that make use of anonymous CVs and grading to undercut stereotype activation. Or ensuring that there is organisational awareness for those occupying certain roles: for instance, that by occupying certain positions (like being a professor) one has special duties toward one's students and younger colleagues to foster inclusion. Again, there is nothing *per se* illiberal about organisational and institutional (re)structuring with the aim to promote egalitarian goals, inclusion, integration, and fairness. Of course, this may be badly executed – but that is another matter.

These interventions are more indirect by targeting the grounds of discriminatory speech, and hence substantially different from the interventions discussed above regarding hate and toxic speech. This demonstrates just how complex an issue we are dealing with. But importantly (I hold), it once more highlights that the dichotomy of *do nothing* or *impose censorship* is false. There is much that can be done on structural and institutional levels too.

5. Philosophy as Vaccine?

In an editorial of *The Scotsman* in January 2021, philosophy was compared to a vaccine against examples of toxic speech like Trump 'the Winner':

> we are heading towards a new world in which philosophy and the ability to think logically become increasingly important. To use a current metaphor, we need to vaccinate ourselves against the

> virulent lies of people like Trump and the best way to do that is to teach the wisdom of Socrates and co to our children.[7]

The idea that there is an antidote to toxic speech through philosophy might look immediately appealing – at least to many philosophers. I too used to think of philosophy and the ability to think critically as being akin to a sort of vaccine (or rather, a bullshit filter) that inoculates us against the influence of the speech examples I have focused on here. I no longer share my earlier optimism. Given how easily our cognitive abilities and faculties can be perverted – or given how rustable they are – different remedies are needed. But, as I have suggested in this paper, there are various remedies at our disposal when dealing with hate, discriminatory, and toxic speech, where these remedies are consistent with a liberal commitment to free speech. There is much that can be done with a more nuanced and less knee-jerky understanding of this commitment. Still, I hold, for liberal interventions to be effective, we must give up the idea of 'the marketplace of ideas', where we can debate as equals and where the truth will triumph. In thinking about the different sorts of prejudicial speech within a liberal framework, we should not consider our limitations, flaws, and non-ideal speech situations as distortions and perversions of the ideal marketplace that (somehow supposedly) came first. Rather, they are 'our' default modes of being and should be the starting point.

References

H. Battaly, 'Testimonial Injustice, Epistemic Vice, and Virtue Epistemology', in I.J. Kidd, J. Medina, and G. Pohlhaus Jr. (eds), *The Routledge Handbook of Epistemic Injustice* (NY: Routledge, 2017), 223–32.

D. Brink, 'Millian Principles, Freedom of Expression, and Hate Speech', *Legal Theory*, 7 (2001), 119–57.

A. Brown, 'What is Hate Speech? Part 1: The Myth of Hate', *Law and Philosophy*, 36 (2017a), 419–68.

A. Brown, 'What is Hate Speech? Part 2: Family Resemblances', *Law and Philosophy*, 36 (2017b), 561–613.

[7] https://www.scotsman.com/news/opinion/columnists/donald-trump-philosophy-antidote-dangerous-liars-us-president-scotsman-comment-3090608. Accessed 16 January 2024.

R. Delgado, 'Words that Wound: A Tort Action for Racial Insults, Epithets, and Name Calling', in M. Matsuda, C.R. Lawrence III, R. Delgado, and K. Crenshaw (eds), *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (Boulder, CO: Westview Press, 1993), 89–110.
S. Fish, *There's No Such Thing As Free Speech: And It's a Good Thing, Too* (NY: Oxford University Press, 1993).
S. Fish, 'Going in Circles with Hate Speech' (2012), *New York Times*, https://opinionator.blogs.nytimes.com/2012/11/12/going-in-circles-with-hate-speech/. Accessed 19 June 2021.
E. Fricker, 'Doing (Better) What Comes Naturally: Zagzebski on Rationality and Epistemic Self-Trust', *Episteme*, 13 (2016), 151–66.
K. Gelber, 'Differentiating Hate Speech: A Systemic Discrimination Approach', *Critical Review of International Social and Political Philosophy*, 24 (2021), 393–414.
K. Gelber and L. McNamara, 'Evidencing the Harms of Hate Speech', *Social Identities*, 22 (2016), 324–41.
K. Jones, 'The Politics of Intellectual Self-Trust', *Social Epistemology*, 26 (2012), 237–51.
K. Kawakami, J. Moll, S. Hermsen, J. Dovidio, and A. Russin, 'Just Say No (to Stereotyping): Effects of Training in the Negation of Stereotypic Associations on Stereotype Activation', *Journal of Personality and Social Psychology*, 78 (2000), 871–88.
I. Maitra and M.K. McGowan (eds), *Speech and Harm: Controversies over Free Speech* (Oxford: Oxford University Press, 2012).
M. Matsuda, C.R. Lawrence III, R. Delgado, and K. Crenshaw (eds), *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (Boulder, CO: Westview Press, 1993).
M. Mikkola, 'Self-Trust and Discriminatory Speech', in K. Dormandy (ed.), *Epistemology of Trust* (NY: Routledge, 2019), 265–90.
M. Mikkola, 'A Distortion or "Our" Default?', *Aristotelian Society Supplementary Volume*, 95 (2021), 143–62.
M. Mikkola, 'Discriminatory vs. Hate Speech: Wherein Lies the Difference?', in M. Popa (ed.), *Oppressive Speech and Society: Philosophical Perspectives* (NY: Routledge, forthcoming 2024).
J.S. Mill, *On Liberty* (London: Penguin, 1974).
J. Montmarquet, 'Epistemic Virtue and Doxastic Responsibility', *American Philosophical Quarterly*, 29 (1992), 331–41.
G.B. Moskowitz, and P. Li, 'Egalitarian Goals Trigger Stereotype Inhibition', *Journal of Experimental Social Psychology*, 47 (2011), 103–16.

T. Scanlon, 'Freedom of Expression and Categories of Expression', *University of Pittsburgh Law Review*, 40 (1978–9), 519–27.
L. Tirrell, 'Discursive Epidemiology: Two Models', *Aristotelian Society Supplementary Volume*, 95 (2021), 115–42.
C. West, 'Words That Silence? Freedom of Expression and Racist Hate Speech', in I. Maitra and M.K. McGowan (eds), *Speech and Harm: Controversies over Free Speech* (Oxford: Oxford University Press, 2012), 222–48.