Ethical Theory and Moral Practice (2023) 26:555–575
https://doi.org/10.1007/s10677-023-10409-w

A Republican Conception of Counterspeech

Suzanne Whitten[1]

Accepted: 22 July 2023 / Published online: 28 July 2023

Abstract

'Counterspeech' is often presented as a way in which individual citizens can respond to harmful speech while avoiding the potentially coercive and freedom-damaging effects of formal speech restrictions. But counterspeech itself can also undermine freedom by contributing to forms of social punishment that manipulate a speaker's choice set in uncontrolled ways. Specifically, and by adopting a republican perspective, this paper argues that certain kinds of counterspeech can *dominate* when they contribute to unchecked social norms that enable others to interfere arbitrarily with speakers. The presence of such domination can pose just as much a threat to freedom of speech as unchecked formal restrictions by threatening an individual's discursive status, revealing a problem for those who defend counterspeech as a freedom-protecting alternative. Rather than rejecting both counterspeech and legislation outright, however, this paper argues that the republican principle of *parsimony* ought to be exercised when deciding on appropriate harmful speech response. While the principle of parsimony allows for suitably-checked formal punishment for some of the most egregious forms of harmful speech, citizen-led counterspeech must be guided by a reliable set of norms against the use of social punishment where those who do engage in social punishment face certain costs. The presence of robust, widely-known, and reliable norms thus supports both formal and informal responses to harmful speech while maintaining a secure discursive status for all.

Keywords Counterspeech · Hate speech · Republicanism · Non-domination · Freedom of speech · Shaming

✉ Suzanne Whitten
suzanne.whitten@qub.ac.uk

[1] Queen's University Belfast, Belfast, UK

1 Introduction

Consider the story of Palestine-born US-citizen Majdi Wadi (Mouk 2020). Since arriving in Minneapolis in the early 1990s, Wadi had made a great success in founding and running ‘Holy Land’, a large Mediterranean-themed restaurant, supermarket, and catering company that eventually employed around 200 workers. In June 2020, Wadi’s 24-year-old daughter was exposed on social media as having posted a series of racist and anti-Semitic tweets as a teenager.[1] The posts were shared widely across a range of platforms, drawing widespread negative attention and criticism, mostly aimed at Wadi himself. Protestors, both online and in-person, publicly rejected the racist messaging of the tweets and urged the public and other companies to boycott the business. The furore surrounding the exposure resulted in significant cost to Wadi’s livelihood (Fadel 2020). Around $5 million worth of business contracts were lost, the landlord in charge of the largest of Holy Land’s premises chose to evict the business, and Wadi eventually had to move his family out of their home after threats were made to their safety.

In such instances of *counterspeech*, understood here as expressive challenges to harmful utterances, the aim of exposure,[2] which is often accompanied by criticism, is to amass public rejection of a harmful message and to challenge the authority of the speaker to enact a racist utterance (Radzik et al. 2020: 50). Counterspeakers in this case may have felt that they were exercising a duty to speak out and prevent the racist utterances from enacting norms that subordinate Black and Jewish members of society (Maitra 2012), to assure members of those groups that their equality was a matter of public importance (Lepoutre 2017: 864), and to deter would-be racist speakers from making similar utterances in future (He et al. 2021).

But the cumulative effect of the actions of individual counterspeakers in this case is illustrative of the way in which counterspeech can also authorise destructive forms of social punishment, very often far beyond the grasp of individuals who, in exercising duties to ‘call out’ oppressive expression, have contributed to the online cacophony of disapproval (Tosi and Warmke 2016; Teekah 2015; Aly and Simpson 2019). The outsized influence of social media in today’s world means that the kind of experience suffered by Wadi is not uncommon[3] (Thomason 2021; Klonick 2016).

What are we to make of such cases? For those who defend ‘more speech’ as a noncoercive, free speech-protecting alternative to extensive formal bans, everything appeared to have unfolded as expected: individual counterspeakers did not call upon the state (or the social media company) to impose costs upon Wadi or his family, and instead his daughter’s racist tweets were publicised and her authority to enact a racist norm challenged (Langton 2018a). The internet counterspeakers who called for the exposure, criticism, and boycott of Wadi’s business will thus have sent a strong message that racist speech is not acceptable and will likely have deterred (at least some) would-be hate-speakers who came across the story (Miškolci et al. 2018).

[1] Some of these tweets included: “Ghetto people are always naming their kids after stuff they can’t afford, Mercedes, Bentley, Pearl, Life Insurance” and “#IfIwasPresident I’d finish off what hitler started and rule the world” (Facebook 2020). Available from: https://www.facebook.com/photo?fbid=2708816462696509&set=pcb.2708816506029838. Accessed 03/04/23.

[2] For a discussion of how exposure works as a form of social punishment, see Nagel (1998).

[3] Some other high-profile examples of this form of social punishment include the case of Justine Sacco (Ronson 2015) and Juli Briskman (Graef 2018).

While the counterspeech in this case could be viewed as a success from an efficacy standpoint, there are good reasons to be concerned about such forms of counterspeech that- either intentionally or inadvertently- also authorise social punishment in this way. Counterspeech theorists themselves have taken seriously the potential risks posed by counterspeech practices, constructing their approaches in ways that minimise burdens on victims (Howard 2021) and that shield against the oppressive harms of amplification (McGowan 2018) and backlash (Saul 2021). In parallel, a burgeoning literature on the ethics of public shaming has urged caution when engaging in online discourse, due to the potential effects on targets' social reputation, material circumstances, freedom of speech, and mental health (Frye 2022a; Billingham and Parr 2019; Elford 2021; Fritz 2021).

This paper, which argues from a republican standpoint, maintains that we should also take seriously the ways in which counterspeech can contribute to *domination*. Just as arbitrarily enacted and applied formal regulations expose individuals to unchecked interference, counterspeech practices can, at the same time as they challenge harmful expression, *also* risk the enactment of social norms that authorise significant levels of interference. In environments where such interference is especially costly and uncontrolled, the presence of such a risk manipulates a speaker's choice set in a way which distorts the conditions necessary in order to enjoy freedom of speech effectively. So-called 'free speech infrastructures' of this kind will thus threaten the equal, non-dominated status of citizens by preventing them from taking part in the discursive contestation and construction of the shared rules and norms of society. Just as there are checks on state regulation of speech, then, there ought to also be checks on the social regulation of speech.

My argument will proceed as follows. In Sect. 2 I begin by critically assessing the claim that counterspeech is a reliably noncoercive, free speech-preserving substitute for formal speech regulations. Section 3 articulates and defends an alternative republican approach to harmful speech which provides a more substantive account of the wider formal and informal conditions required in order to enjoy a free, discursive status. One formal way in which one's discursive status can be protected- hate speech law- is then critically examined (Sect. 4) and shown to be unreliable in many cases. Section 5 advances a republican account of counterspeech as an informal method of securing a non-dominated, discursive status using speech. Certain forms of counterspeech are shown to interfere just as arbitrarily with freedom of speech as formal regulation, thereby posing a threat of domination. In response, those concerned with threats to non-dominated discursive status from both formal and informal sources must exercise *parsimony*, which in the case of counterspeech requires securing robust norms against the use of social punishment.

Before starting, a brief preliminary note. In this paper, I take a broad view of the kind of speech that counterspeech may be used in response to. In that sense, I will not confine my area of concern to 'hate' speech only, understood as expression which denigrates or undermines the equal standing of members of society because of their identity (McGowan 2012; Langton 2012), but will also consider speech that is oppressive in nature (McGowan 2019: 106). One of the positives of counterspeech is that it can form part of a coordinated response to all sorts of harmful speech, including speech which contributes to damaging stereotypes or associations (Lemeire 2021), propaganda (Stanley 2015), 'dog-whistles' (Saul 2018a), and 'microaggressions' (Rini 2020), as well as other generally disparaging ways in which certain groups are discussed and addressed in the public sphere. With this goal in mind, then, I use the term 'harmful' speech to describe all those forms of speech that contribute in some way to hierarchies of status that rank individuals according to their perceived membership of a particular identity group.

2 Counterspeech as a Threat to Freedom?

The political theoretical literature on harmful speech is, for the most part, concerned with solving the following puzzle. That is, how should we respond to speech which undermines the equal dignity and perpetuates the oppression of some of society's most vulnerable members, without in turn undermining one of the fundamental components of democracy: freedom of speech?

Defenders of *counterspeech*, including those who argue in favour of legal bans and state-led 'expressive' measures in limited cases, suggest that we need not sacrifice freedom of speech in order to protect those targeted by such harms. Instead[4], we can look to the actions of individual citizens themselves, who, in exercising mechanisms carefully conceptualised by ongoing work in political theory and feminist philosophy of language, can intervene in (McGowan 2019: 106), challenge (Fumagalli 2021), silence (Saul 2018b), 'block' (Langton 2018b), or otherwise undermine harmful speech using particular context-sensitive discursive 'moves'.[5]

For sceptics of speech regulation, counterspeech practices provide a solution to harmful speech without posing a threat to freedom of speech. On this view, state bans on speech which expresses an *idea*- whether hateful[6] or not- cannot be enacted without at the same time limiting the conditions of democratic legitimacy and undermining the autonomy of individual citizens (Post 2011; Strossen 2012: 379). Democratic legitimacy requires that the state afford citizens a sizeable sphere of freedom from intervention, within which they can exercise their expressive freedoms among themselves, deliberate on matters of democratic importance, and negotiate the terms of political life without undue influence or threat from above (Meiklejohn 1961; Baker 2012: 63). This 'democratic background' essential to such an arrangement can only be achieved where all are able to exercise voice and make their thoughts and opinions known (Baker 1989: 59; Post 2005: 144-7). While harmful speech may very well contribute to ongoing discrimination and inequality, so this argument goes,

[4] Influential cases for this view have been put forward by Strossen (1990, 2018: 130), Heinze (2016), Dworkin (2009), Post (1990), and Baker (1996).

[5] Other notable publications on the topic of counterspeech include Lepoutre (2017, 2019), Tirrell (2018, 2019), Brettschneider (2016), Gelber (2002, 2012), Nielsen (2012), Strossen (2018), and Howard (2021).

[6] Critics of hate speech laws, for the most part, agree that harmful speech can have real, tangible effects on those it targets. Their scepticism lies instead with the state's involvement in adjudicating and punishing such speech in a predictable and proportionate way (Strossen 2018: 14). Compared to other forms of legally-restricted wrongs, so they argue, speech is a special case, owing partly to the 'slippery' nature of identifying speech harm and to the fact that speech plays such an integral role in the enjoyment of freedom more broadly (Dworkin 2009). In short, placing this adjudicating-power in the hands of the state risks the involvement of factional interests, which has potentially serious downstream effects on the maintenance of democratic legitimacy (Sunstein 1995). The prime importance of the right to freedom of speech thus means that, to many, speech restriction is only to be permitted in the most egregious of circumstances and must be guided by a set of clearly-defined categories of restriction designed to avoid vague and overbroad interpretation and application.

it nonetheless must not be intervened with by the state.[7] Once we exclude content-based speech interventions from our arsenal of possible solutions to harmful speech, then our only hope is that individual citizens take the initiative and confront such speech themselves (Howard 2021: 926). Here, and following Justice Brandeis's 1927 concurring opinion in *Whitney*: "The fitting remedy for evil counsels is good ones… If there be time to expose through discussion the falsehood and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence."[8]

Not all are so optimistic about the efficacy of 'more speech' for securing freedom, however (Delgado and Stefancic 1996; Nielsen 2012). Harmful speech, as has been extensively argued, also negatively impacts the equal standing of targets in ways that prevent them from taking part in democracy effectively (Shiffrin 2011; MacKinnon 1993; Waldron 2012: 155). Quite apart from the emotional and psychological toll that such speech takes on victims (Delgado 1982; Matsuda 1989), the status-undermining effect of pervasive harmful speech will mean that it is likely that their attempts at confronting those who use such speech will be both burdensome and ineffective (Cohen 1993: 256; Delgado and Yun 1994). This will be especially the case, of course, for members of target groups whose social position automatically places them at a discursive disadvantage (Langton 2012).

The well-established negative impact of harmful speech for victims thus provides an obligation for those of us who are at minimal risk of harm to support the counterspeech of others by challenging such speech in our day-to-day lives (Goldberg 2018; McGowan 2018). Apart from those cases where counterspeech proves to be particularly ineffective (Cepollaro et al. 2022), most argue that we have no good reason to stand by and permit harmful utterances from continuing to shape and diminish the social standing of our fellow citizens (Ayala and Vasilyeva 2016; Langton 2018b: 161). As a set of practices that are believed to pose a minimal threat to freedom of speech, then, counterspeech continues to be an essential tool in our battle against pervasively harmful speech.

As the Wadi case and similar demonstrate, however, it appears that some instances of citizen-led counterspeech can impose costs that present just as much of a threat to freedom of speech as certain kinds of formal regulation. To understand how counterspeech might behave in this way, we need to explore how social punishment works as a method of norm enforcement. Social norms, understood as "clusters of normative attitudes plus knowledge of those attitudes" (Brennan et al. 2013: 15) are key determinants of human behaviour. The maintenance of norms is in large part dependent on the (actual or predicted) reactive responses of others, where

> the spontaneously emerging attitudes of others towards us when we act in norm-governed contexts operate as a distinctive kind of intangible sanctioning mechanism, culminating in us externalizing and ultimately conforming with norms (ibid.: 225).

[7] Critics of hate speech laws also express concern that the presence of such laws will give rise to a 'chilling effect' (Strossen 2018: 99). In theory, the chilling effect threatens free speech when individuals self-censor or refrain from engaging in public discourse due to concerns that their speech will be 'caught in the net' of hate speech law (Hare 2012). While there is limited evidence demonstrating the existence of a chilling effect as a result of hate speech law specifically (Brown 2015: 267), we can at least predict that such an effect will be more likely where such laws are poorly written or applied, such that individuals will be unclear as to what they are permitted to say in public. Where the chilling effect arises, individual citizens will not be able to exercise speech freely, in turn limiting their ability to engage in matters of democratic importance.

[8] *Whitney v. California*, 274 U.S. 357 (1927).

Public shaming, including the online variety (Billingham and Parr 2019; Frye 2022a), is an especially powerful method of social norm enforcement that ensures that those norms perpetuate over time (Anderson 2000; Jacquet 2015). By holding wrongdoers (i.e., norm-breakers) to account in public settings, shamers make wrongdoers and audience members (Radzik 2016; Detel 2013) aware of the presence of a morally authoritative norm while simultaneously expressing the moral commitments of the community[9] (Billingham and Parr 2020).

In reality, however, norm enforcement of this kind will also give rise to harms that are very frequently *coercive* in nature (Aitchison & Meckled-Garcia 2021). According to Linda Radzik, social shaming inflicts harms on targets by playing on their "need for human contact and cooperation" and "desire for goodwill" from others (Radzik et al. 2020: 36). Shaming also impairs a target's self-esteem by encouraging forms of social disapprobation that aim at the root of who they are as persons (Smith 2010). The infliction, or threat of infliction, of these forms of social punishment are coercive when they manipulate the choice set available to the shamed individual in future by rendering some of their actions extremely costly. These costs are amplified where the shaming also involves authorisation of 'tangible' costs on the shamed. Revealing a target's personal information online (also known as 'doxing'), or issuing demands to their employer that they be disciplined or sacked are common examples which involve the imposition of high costs.

Where such threats are commonplace and where the costs are high, it seems that counterspeech can, just as formal responses to harmful speech, also pose a substantial threat to our freedom of speech. One such concern has been raised by Elford (2021), who, in drawing a parallel between state and social imposition, describes how the Internet has exposed individuals to a greater level of reputational vulnerability than ever before. This threat of social punishment online, with its downstream 'real-life' consequences, limits freedom of speech by attaching substantial costs to one's speech. Following Radzik's account of social coercion, then, we might say that, where such discursive environments are commonplace, whether individuals are directly engaged with them or not, then the speech choice sets of some will be unduly limited in a way that diminishes freedom of speech.

So far, then, we have found that both formal and informal harmful speech response potentially threaten freedom of speech under certain conditions. But, given the debilitating effects of harmful speech on target groups, a 'hands off' approach also does not appear to be sufficient. What should be done? One way of avoiding the worst excesses of necessary norm enforcement practices is by abiding by a set of *duties*. For some, duties to avoid imposing foreseeable risks of coercion on others (Elford 2021: 163) and to engage with the moral agency of speakers (ibid.: 168) preclude us from using methods that enact and authorise social punishment. Others, however, argue that duties of this sort can be compatible with social punishment so long as certain conditions are observed (Billingham and Parr 2020). Across the range of duty-focused measures is a stress on engaging appropriately with the discursive agency of wrongful speakers, where such engagement is considered both respectful and necessary for successfully persuading them out of the beliefs that motivated their utterance in the first place.

The idea that we might have duties of this kind seems plausible. However, it is not so clear that a call that counterspeakers exercise certain duties is sufficient to *reliably* secure

[9] Quite simply, racist speech violates a norm of 'anti-racism', sexist speech violates a norm of 'anti-sexism', and so on.

freedom of speech. Absent a set of checking mechanisms, it is not clear that such duties will be consistently exercised, especially by those who are already disposed to supporting harmful counterspeech practices. In such environments, then, individual speakers will live under threat of potential coercion and so will not enjoy free speech in the proper sense. Risks of interference will be amplified for those from marginalised groups. In figuring out what a reliable freedom of speech might require, then, what is needed is an account of how discursive environments more broadly secure or undermine the freedom of speech of individuals. This aim, I suggest, can be achieved by adopting an alternative republican framing of harmful speech response which takes seriously the *domination* posed by particular discursive arrangements. Here, a 'free speech infrastructure'- which takes free speech not as pre-institutional but as *dependent upon* well-designed and controlled formal laws, rules, and social norms of society- must be of a sufficiently freedom-promoting quality in order for individuals to enjoy a robust discursive status.

3 Republicanism and the Free Speech Infrastructure

The republican tradition[10] is at heart concerned with the freedom afforded to individuals when they are self-governing, where the constitutive political and social arrangements of society are decided 'on the people's terms' (Pettit 2012) and subject to ongoing critical scrutiny from an active citizenry committed to exercising 'eternal vigilance' (Pettit 1997: 250). To ensure that the laws and norms of the land are decided upon in such a way, individual citizens must enjoy a status in which they are free in a *non-dominated* sense. Freedom as non-domination, as articulated by Pettit (2012: 58), arises not simply when an individual experiences a lack of interference from others, but when they enjoy a lack of arbitrary (or "uncontrolled") interference. Along such lines, a republican will be concerned with the kinds of structural arrangements that need to be in place in order for an individual to enjoy a *robust* freedom as non-domination, such that they feel secure that they can go about their lives without having to submit to the unpredictable and arbitrary whims of others. Importantly, the republican ontological position on the necessary interdependency of human beings takes free status as something that *depends* upon, rather than is hindered by, a combination of the rule of law and strong social norms. As Skinner (2010: 97) describes, one who enjoys such a status as a 'freeman'

> is consequently someone to be reckoned with, someone who can look you in the eye, who can reason and negotiate on terms of equality without ever feeling the need to doff the cap or bend the knee. As a result, freemen are said to enjoy the respect- and the self-respect- that comes of being known to speak frankly and behave without fear or favour, acting solely as reason and conscience dictate.

[10] It is important to acknowledge, of course, the diverse and rich range of work which has emerged from the republican literature in recent decades. For the purposes of this paper, however, I focus on a strand of 'neo'-republicanism first developed by Pettit (1997, 2012) and adapted in various ways by scholars including Laborde (2008), Lovett (2022), and Skinner (2010), among many others. While there are important differences between their perspectives, the thrust of my argument is built upon core principles common to each.

The robustness of my status as a free person, importantly, must be such that everyone is aware of my status, including myself (Montesquieu 1977: 202). In that sense, individuals "should have an undominated status both in the objective and the subjective or inter-subjective sense of status" (Pettit 2012: 83).

The subjective component of non-domination, satisfied where the laws and social norms of society are such that they serve as a reminder that one's standing is secure, is important for the following reasons. For one, a robust awareness that the state will protect against arbitrary infringements on one's freedom provides individuals with the confidence to engage in the democratic process, which in turn ensures democratic legitimacy (Rostbøll 2015). Put differently, when the knowledge that I am respected as an equal co-deliberator in the decision-making process is enshrined in the law, I come to view myself as someone whose *voice* will be taken seriously (Laborde 2008: 16). Further, when the norms that make up our social relationships and our civic interactions *also* reflect my standing as an equal, I can thereby form bonds of trust with my fellow citizens (Pettit 1997: 261), which not only satisfies key features of social *inter*dependence but also allows me to exercise *in*dependence from others (Braithwaite and Pettit 1990: 67; Pettit 1997: 241). Individuals who enjoy non-domination, then, are free in both a political and social sense, in a society that respects their contributions as a co-creator of shared rules and norms and which allows them to enjoy social relations of a sufficiently high quality (Pettit 1997: 265; Laborde 2008: 17−8).

The presence of robust norms and formal rules for the enjoyment of non-domination also applies to the enjoyment of freedom of speech. In carving out a characteristically republican understanding of free speech, Philip Pettit marks a distinction between two different ways in which our speech among a range of options can be considered 'free' (Pettit 2018: 61). On the first (typically liberal) account, which Pettit calls 'unhindered' speech, I am free in my speech options to the degree that I am free to choose from among a range of options (ibid.: 62). The second (typically republican) account of 'protected' speech requires not only that I am free to choose from among a range of speech options, but that I am *securely protected* in my choices (ibid.: 64).

Importantly for this form of protected speech is that this security of discursive status is something that I can rely upon in my communicative interactions with others. A protected discursive status thus requires that there are costly burdens in place that deter would-be interferers from preventing me from speaking. To enjoy a secure discursive status, such that I feel confident to live alongside others without fear of interference in my speech, it is not enough that these costs reduce the likelihood that I will be interfered with. Rather, free speech on this account requires a *robustly reliable* set of formal and informal norms to be in place. The costs attached to norm-breakers will need to be

> severe enough to enable you to say your bit on any topic- and, figuratively, to look others in the eye as you do so- without the power of interference of others giving you good reason, by local criteria, for fear or deference (ibid.: 65).

The benefits of protected speech over unhindered speech are clear. Under unhindered speech, I am free to speak so long as others, no matter how much power they have over me, decide not to interfere with me. For example, I may work for an employer who has the power to discipline workers for complaining about working conditions with colleagues. So long as I am in my employer's good favour I would be free in an unhindered sense. In contrast, to be free in a *protected* sense would require that there are regulations and norms in place that *prevent* employers from disciplining workers for their grumblings. We can see that in the latter case, I enjoy a discursive status not afforded to me in the former. Where my free speech is protected in this way I do not have to continuously monitor my speech among colleagues for fear of future punishment, nor do I need to ingratiate myself to my employer by flattering them or bending to their whims. Importantly, now that I am free to discuss working conditions with colleagues, we will have the capacity to check and monitor the power of the employer. Protected speech offers us a richer form of free speech than unhindered speech alone. As Pettit notes, the distinction reveals how freedom of speech is not to be understood pre-institutionally, but rather

> [i]t is only by dint of law and regulation- and supportive social norms- that speech gets to be protected, and gets to count as free (ibid.: 67).

The requirement for a robust set of suitably-controlled laws *and* norms as a prerequisite for freedom of speech has notable implications for how we approach our responses to harmful speech. Taken together, the formal and informal aspects of a free speech infrastructure need to be such that one's discursive status is suitably protected. This framing thus demands a rather more substantive set of conditions for the enjoyment of freedom of speech in the face of harmful speech than those described in Sect. 2. Rather than working out how to minimise interference in the realm of speech, then, we must consider what both the formal and informal of speech conditions ought to look like in order for individuals to enjoy freedom of speech effectively, whether one is a potential victim of harmful speech or an ordinary speaker. In the following, and by employing the above republican framework, I first critically examine a common formal response in the form of hate speech law for its suitability for protecting discursive status. Laws of this type, I argue, struggle to satisfy the demands of non-arbitrariness, thus making them unreliable for securing non-domination except in a very narrow set of cases.

4 Formal Responses to Harmful Speech and the Problem of Arbitrariness

The republican understanding of freedom as conditional means that, when approaching the question of speech regulation, republicans will be concerned not with the demonstration of 'harm' but with determining whether the at-issue speech (whether libel, incitement, or hate speech) must be limited in order to secure freedom as non-domination (Pettit 1997: 36). Importantly, the proposed regulation in question must also satisfy the conditions of being both 'co-exercisable' and 'co-enjoyable', so as to avoid state domination towards a minority of citizens.[11] A free speech infrastructure that provides protection to some groups and not

[11] Pettit's stipulation that any proposed law must be 'co-exercisable' and 'co-enjoyable' is reflective of a republican commitment to limited government. To ensure that the republic remains a system determined by and led by the people, rather than some arbitrary power or group of rulers, the rule of law must be applicable to all equally and must be subject to continuous scrutiny. The laws themselves must also be known by all and be easily understood by those who are held by them. A law which is not constructed with appropriate input by the people, or which does not allow for scrutiny, will instead be imposing upon citizens a threat of arbitrary

others in ways which increase domination overall will thus not be one in which all enjoy the independent status required to take part in public life effectively.

There are three key ways in which hate speech laws potentially increase, rather than reduce, the domination of citizens. Firstly, it is not clear that many hate speech laws- in whatever form- can be written and applied in a way which satisfies a republican commitment to non-arbitrariness.[12] One level of arbitrariness arises in the way such laws describe the kind of speech they deem to be punishable. Take, for example, one of the most common categories of hate speech ban and that which is defended by Waldron (2012) as offering minimal threat to free speech: group defamation law. Such laws can be criticised for an inherent vagueness, which makes them difficult to implement in fair and balanced ways. Article 20 of the International Covenant on Civil and Political Rights (ICCPR), for instance, outlaws speech involving "[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence." A similar version can be found in Sect. 18(1) of the United Kingdom Public Order Act 1986, which targets those insulting words where the speaker "intends… to stir up racial hatred." In Canada, the Supreme Court has defined hate speech as that which "is likely to expose" individuals to hatred, defined as "unusually strong and deep-felt emotions of detestation, calumny and vilification" and "enmity and extreme ill-will" which "goes beyond mere disdain or dislike."

As commentators have pointed out, however, the broad and overly-vague wording of such laws makes it extremely difficult for the courts to apply them consistently (Heinze 2006). From a republican perspective, the greater the amount of discretion at the hands of those involved in the criminal justice system- in this case, the jury- the more prone a particular law or legal practice is to produce domination. As Strossen (2018: 76) argues in her reflections on the Canadian example:

> If you were a juror, would you be able to distinguish between speech that conveys "disdain," which is not punishable, and speech that conveys "detestation" or "vilification," which is? And if you were voicing your own strongly negative views about some person or group, how secure would you feel that officials would consider your words to communicated protected "disdain" rather than punishable "enmity" or "extreme ill-will"?"

Such uncertainty is intensified by the inherent heterogeneity of contemporary liberal democracies. Police, prosecution, and jury identification of 'vilification' and 'disdain' will depend strongly on how the individuals involved are socially-situated, meaning that implementing such laws evenly, even while taking great care to exercise caution, will be a considerable challenge in plural societies. Differences in culture, religious belief, class, gender, and race all have been shown to strongly influence public perceptions of the 'harm' in speech, thereby limiting the possibility of satisfying a standard 'reasonable person' test (Cox 2016; Moran 2003). Existing power inequalities in society between the majority and those with minority values thus "ultimately implements a majoritarian consensus that is prejudicial towards minorities or, more accurately, those holding minority views" (Kohl 2018: 114).

interference. Further, proposed laws must also be easily interpreted and non-arbitrarily applied by the policing and criminal justice system in order to satisfy the conditions of non-domination.

[12] I acknowledge that the term 'hate speech laws' covers an extremely diverse range of laws and regulations which, like any other category of legislation, suffer differing levels of potential arbitrariness (Brown 2017a).

A second kind of arbitrariness can be found in how hate speech law necessarily distinguishes between those groups which fall under their protection and those which do not. The problem, according to Eric Heinze, is that such formal distinction between protected and unprotected groups within the law will "inevitably create two tiers of citizens- those who are protected from offensive speech, and those left unprotected from equally offensive speech" (2006: 555). To illustrate, Heinze points to common forms of prejudice which, while not falling under formal definitions of hate speech, are nonetheless just as damaging to their victims, pointing out that

> many physical, mental or psychological conditions that provoke hurtful speech, such as low intelligence, ugliness, frailty, mental illness, physical debility or malformation, poor coordination, blindness or deafness may be neither self-induced nor easy to change, and have histories of stigmatism stretching further back than any current form of racism (ibid.: 566-7).

It could reasonably be argued that, while the above groups of people will suffer the harms of 'offensive' speech, such harms do not contribute to the *systemic* forms of discrimination experienced by those groups typically protected by hate speech laws. This argument can be supported with reference to the so-called 'symbolic' or 'expressive' argument for formal bans (Brown 2015: 240). According to this view, such laws can be justified not by the frequency of incidents but by their expressive or symbolic role in reminding systematically subordinated groups of their equal standing. In republican terms, both the existence of the law and its enforcement are believed to provide that intersubjective reminder of robust status necessary for freedom as non-domination[13]. By enacting a law designed to protect a specific group, then, the state is not expressing the idea that the group are indeed morally superior to other members of society (Galeotti 2002: 156). Rather, they are responding to the fact that there exists a background social hierarchy in which that minority group are *already* placed in a diminutive social position. By enacting a law of this type for symbolic purposes, then, the state could be said to be attempting to shift social understandings of that group such that they and their fellow citizens can 'walk tall and look others in the eye' in the way republicans describe (Pettit 2012: 3).

In practice, however, the demonstration of systemic discrimination has not taken such a central role in the decision over which groups to protect and nor, when it is invoked, does it tend to be very effective at securing equal standing in the way defenders of the symbolic argument claim (Lepoutre 2020)[14]. The UK Public Order Act (1986), for example, has been guided in their selection not by the relative social standing of groups but on predicted prosecution rates. It is because of this criterion for qualification that the Act has been criticised for excluding disability and gender-identity as protected characteristics (Brown 2017b: 300).

Further, the differing threshold evidential burdens required for prosecution across groups means that full protection of certain groups is not possible without posing a threat to freedom

[13] We can also see this claim mirrored in Jeremy Waldron's influential defence of hate speech laws as providing a sense of "assurance" (Waldron 2012: 4; Brown 2015: 240) that the rights and equal dignity of vulnerable groups are respected and will be upheld in the society in which they go about their lives.

[14] Pointing to the lack of empirical evidence to demonstrate that hate speech laws have been effective, one U.N. report produced by the High Commissioner for Human Rights concluded that "the criminal model is not an efficient tool when it comes to addressing the real causes of discrimination" (Bertoni 2011).

of speech. As it stands, a higher evidential burden is in place in cases involving religious hatred compared to those involving racist hatred. Together with an additional set of clauses, the higher standard required for cases of incitement to religious hatred aims to protect non-hateful criticism of religion and to reduce the risk of a 'chilling effect' (Barendt 2011) at the same time as it protects religious individuals from hate (Brown 2017b: 35). Arguably, however, some of the most common and influential anti-LGBT and anti-Muslim speech, for example, operate under the guise of 'criticism' (Mondon and Winter 2017; Kolhatkar 2014). That current hate speech law in the UK aims to protect non-hateful criticism is, of course, a positive from the perspective of free speech, but it still leaves us with a large proportion of harmful speech uncovered. If incitement laws do not cover some of the most common and consequential cases, then, their purpose comes under question (Clooney and Webb 2017).

These problems will pose a worry to republicans, not least because of the high risks already attached to state intervention in speech in the first place. Were hate speech laws proven to be ineffective at protecting the vulnerability of target groups, then their implicit imposition on citizen freedom can no longer be justified. Republicans recognise the chief importance of the law in securing citizen freedom (Pettit 1997: 36), but stress that the decision to attach state-imposed costs on behaviour can only be justified where the strengthening effect on *dominion* (i.e., non-domination) when regulation is in place far outweighs the significant weakening effects on dominion attached to regulation. The decision to invoke the criminal justice system as a response to harmful speech, then, must be guided by the principle of *parsimony*, on which Braithwaite and Pettit (1990: 79) state:

> Every act of punishment has the certain cost of diminishing someone's dominion and if we are concerned about dominion then every act of punishment will need positive justification. The rest-response will be mercy, the response that needs vindication punishment. More generally, the rest-response will be non-intervention, the response that requires justification will be intrusion.

What we can take from this is that the more difficult it is to distinguish between non-hateful public discourse and the incitement or stirring up of hatred (intended or foreseen), the less comfortable we should be with lowering the threshold for prosecution. Those groups that are not currently included under most hate speech legislation thus will tend to be those where the distinction between offensive speech and hate speech is more difficult to discern. To extend current incitement laws to include them would either require maintaining a high threshold on prosecution (and thereby to punish very few cases), or to unjustifiably infringe upon freedom of speech in a way which limits public conversation about important issues. In either case, the discursive status required for the enjoyment of non-domination remains unsecured. We must, then, look to other key aspects of our discursive environment for a response to harmful speech that is both reliable and freedom-protecting.

5 Counterspeech: A Republican Defence

Given republicanism's commitment to a joint system of laws and norms as necessary features of the equal standing of all, it appears that we have good reasons to count counterspeech as one of the necessary practices of securing such a system.[15] Counterspeech might thus be described as an exercise in enacting and expressing core civic values through the exercise of *vigilance* (Laborde 2008: 152). In the context of harmful speech, then, this not only means that a republican framing recognises that harmful speech is a widespread problem for freedom (Bonotti 2017), but that protecting the psychological, subjective (and intersubjective) aspect of equal standing requires that individual citizens take part in the contestation of that speech. In short, then, the apparent necessity of norms of citizen-led counterspeech in the maintenance of a free and equal society means that counterspeech is not just a form of desirable behaviour but is one of the central components of our free speech infrastructure.

As discussed in Sect. 2, however, counterspeech that uses forms of social punishment can also manipulate a speaker's choice set in uncontrolled or arbitrary ways. In most cases, no individual contribution to collective counterspeech on its own will threaten freedom of speech. Instead, online 'pile-on' scenarios will involve a range of uncoordinated individuals, described by Frank Lovett as a 'team' (2022: 32), who are drawn together temporarily by a "contingent convergence of purposes" (ibid.) to engage in the joint intentional action of norm enforcement. Importantly, the emergence of the team, and their success at manipulating a speaker's choice set, depends upon certain background conditions supported by social norms. These conditions provide the team with the *capacity* to exercise power over a speaker and distinguishes 'latent' teams that pose a domination threat from all of the other potential teams that might have the power to interfere (ibid.: 45).

Certain structural features of social media platforms make it possible for a set of previously-uncoordinated counterspeakers to contribute to a joint intentional action which aims towards holding a speaker to account for using harmful speech (Saul 2021; Coe et al. 2014). These features- which includes 'sharing', 'liking', and 're-posting' (Marsili 2021)- allow individual contributions to form a collective mass that can have major influence over the choice sets of others. For this form of collective counterspeech to dominate, however, we need to determine that the collective: (i) Has the *power* to interfere; and (ii) Has the *capacity* to interfere. The kinds of interference involved include those I described in Sect. 2, such as the social cost of public shaming, shunning, and stigmatisation as well as the authorisation of more 'tangible' material costs of loss of employment and physical threat to safety. The collective of counterspeakers in this case will indeed have the power to impose or authorise those social forms of punishment, the cost of which likely will be determined by the size and social standing of the punishing audience.

But does a newly-coordinated team of counterspeakers have the *capacity* to interfere with speakers? Here, we need to look to the background conditions in place at the time of

[15] Republicans also stress the *practical* limits of the law in securing non-domination in certain cases. According to this view, dominating behaviours which take place away from the glare of the public, such as domestic violence and female genital mutilation, cannot be tackled without first changing the norms of society. Quite simply, without strong social norms against a particular behaviour, it is unlikely that complaints will be made to the authorities and/or be taken seriously (Watkins 2015). While hate speech often takes place in the public sphere, its low detection and prosecution rate suggests that a more 'bottom-up', norm-first approach is thus necessary.

the coordination. Social media, in the last several years, has provided unprecedented opportunities for large numbers of people to congregate and communicate from all over the world (Papacharissi 2009). Twitter, for example, continues to be highly influential when it comes to influencing public debate in certain sections of society (boyd 2011). The opportunity to connect with a wide audience, in addition to the ongoing public influence of the platform, suggests that public shaming events that take place on social media have the potential to have far-reaching consequences. We also have on social media a communicative culture in which 'calling out' norm-breakers is common (Bouvier and Machin 2021; Bouvier 2020). The option for anonymity of users, plus the relative low-stakes nature of engaging in online environments in comparison to those in-person, makes contributing to online forms of punishment typically less costly than in-person forms of norm enforcement (Anderson et al. 2014; Santana 2014). This, we could say, has led in some part to the emergence of a norm that encourages users not to stay silent when they witness harmful speech, and to take part in counterspeech against it (Saul 2021:140-3).[16]

For the power and capacity conditions for domination to both be satisfied, then, counterspeech needs to take place in an environment in which the norms in place authorise, impose, and encourage harmful forms of social punishment for speaking (Frye 2022b). Unfortunately, the Internet is one such place where norms of this kind are commonplace (Basak et al. 2016; Stroud 2016; Goldman 2015). Where such norms are in place, counterspeakers will not censure one another for engaging in harmfully coercive counterspeech practices and may even join in themselves (Gervais 2015; Neubaum 2018). When audience members perceive that large numbers of people support these counterspeech practices then such methods come to be authoritative and therefore seen as appropriate and proportionate. Where such norms become widespread, employers who perceive a consensus regarding the actions of a particular speaker may feel that they can discipline or fire them without consequence (Broderick and Grinberg 2013).

How, then, might the republican framework of free speech approach informal methods of counterspeech of this kind? The central role that social norms play in the maintenance of a free and equal republic means that republicans will be comfortable with at least some sort of social consequence for norm-breakers (Pettit 1997: 254). In their work on the 'economy of esteem', for instance, Brennan and Pettit (2004) build on a centuries-long tradition in political theory on the central role that the 'intangible hand' of social approval and opprobrium can play in securing 'good' social norms, especially where the use of the law is either ineffective or inappropriate. Here, a suitably-harnessed desire for esteem offers an underutilised force with which we can affect human behaviour for the good of society.

There are two key limits to what the economy of esteem can do for counterspeech without worsening overall freedom. The first, which can be thought of as an *individualised* danger, is seen in those cases where public opprobrium results in a permanent or severe loss of reputation, such as when an individual becomes a source of stigma (ibid.: 318).[17] Secondly, and on a *societal* level, a culture in which systems of esteem and disesteem lose their sensitivity to behavioural changes will also no longer be capable of influencing behaviour in

[16] Note that for norms to be in place there need not be large numbers of people publicly supporting or enforcing the norm. All that is required is that people *perceive* that the norm exists and that they will be subject to negative consequences if they break it (see, for example Hardin 2009). The amplification of shaming events on social media thus sends a strong message to the audience that a norm exists against the shamed behaviour.

[17] Here, they refer specifically to J. S. Mill's (1969) concern about the tyranny of social opinion.

socially-desirable ways (Braithwaite 1989). Republicans will thus be wary of those methods that hinder the cause of securing norms that support non-domination, either by engaging in practices that authorise the imposition of severe costs on speaking or by distorting the economy of esteem such that it ceases to influence behaviour in a positive direction, for

> if the infamous are stuck with their notoriety, and the famous with their celebrity, then they are not going to be subject in the ordinary way to the discipline of an intangible hand. There will be little prospect for the infamous of redeeming their reputation, and little prospect for the famous- at least if they play it safe- of losing their reputation. *And so neither will be susceptible in the manner of ordinary folk to the desire for esteem or the aversion to disesteem* (emphasis added) (Brennan and Pettit 2004: 318).

Putting the practical effectiveness of forms of social punishment to one side, severe or unpredictable social punishment in the realm of speech will also negatively impact the robustness of a citizen's discursive status. As with vague and overly-broad speech laws, a discursive environment in which counterspeech is guided by norms endorsing coercive forms of social punishment will also be one in which one's discursive status is insecure (at the very least) or potentially non-existent (at the very worst). Such an environment would have serious downstream effects for both individuals and for society as a whole by undermining citizens' subjective reminder of discursive status and by deterring them from speaking on matters of social or democratic importance. This will, of course, have an even more worrying impact on members of historically silenced or marginalised groups, who already experience higher levels of scrutiny and enjoy a relatively low degree of authority when engaging in public discourse.

If social punishment, especially in today's online world and as it pertains to forms of counterspeech, is as much of a threat to freedom as I have suggested, does this mean that counterspeakers must simply avoid social punishment completely? Building on the discussion of the principle of parsimony (Sect. 2), it seems that we must *also* invoke parsimony when deciding how to best to use counterspeech when responding to instances of harmful speech. As with the constraints imposed on the criminal law by parsimony, counterspeech measures must also be guided by norms which reduce the likelihood of posing a greater threat to dominion than the proposed cure. Where the conditions required for efficacious norm change via social sanction is shown to be absent, then there should be a set of robust and reliable norms in place against such sanction.

What should these norms look like, and how can we ensure consistent compliance? While I do not have space to fully explicate a positive account here, I will briefly outline a few thoughts on how online environments in particular might foster robust norms which aim to protect discursive status while also tackling the effects of harmful speech. In online environments, it is necessary to secure norms against methods of counterspeech which authorise the imposition of substantial costs on speakers. There should thus be norms in place against the use of harmful social punishment and against the imposition of tangible harms on speakers, such as contacting a wrongful speaker's place of employment or revealing their personal details online (Robards and Graf 2022; Douglas 2016; Han and Brazeal 2015). Importantly, such norms should be suitably costly, such that potential punishers are deterred from following through with their punishment (Brennan et al. 2013: 38).[18]

To ensure that norms of freedom-protecting counterspeech are reliable and that they continue over time, however, we must overcome two fundamental hurdles specific to the online realm. Both concern the difficulty of holding wrongful punishers accountable given the structure of online discursive environments. The first, is that 'teams' of online counterspeakers are large, unstable, and made up of agents who each hold varying degrees of authority (Ahrne and Brunsson 2011). A second accountability-limiting feature is that counterspeakers online have the option of engaging anonymously, making them less sensitive to disapproval. On the first issue, it might be possible to stabilise counterspeech behaviours by encouraging the growth of widely-recognised practices across online groups. Once those practices have become established, and guided by their own set of norms, then those who engage in counterspeech will be moved by social pressure to follow those practices over others. One example of such norm stabilisation can be seen in the online counterspeech movement #iamhere. The movement, which originated on Swedish Facebook in 2016, practices a form of organised counterspeech in which members scan online articles and comments sections for harmful speech before, by use of the hashtag, signalling to other group members to join in to counter or correct the speech. The movement, which now comprises over 150,000 members worldwide, has been shown to be extremely effective at combating hate by mobilising a collective while maintaining an ethos of respectful and open dialogue (Friess et al. 2020; Buerger 2021). Importantly, the norms of the movement require that counterspeakers do not harm others while engaging in counterspeech. The continued and well-established practices used by #iamhere reliably prevent those involved from using harmful methods by maintaining a group identity[19] guided by values of respect (Sunstein 1996/1997). Long-term and large-scale use of standardised counterspeech practices online can secure robust norms which can be relied upon in ways that protect discursive status while undermining the effects of harmful speech.

The second issue of anonymity can also be dealt with by establishing the form of collective group identity found in the above #iamhere example. Here, it is not necessary that individuals engaged in counterspeech use their real identity when taking part in counterspeech. Instead, the creation of an activist group identity can go a long way towards securing compliance with a set of norms of behaviour while counterspeakers carry out their work (McCarthy & Zald 2001). The creation of a large-scale common (yet flexible) coun-

[18] While each case must be judged on its own merits, it is very likely that the exercise of parsimony when applied to counterspeech will strongly endorse norms against online forms of exposure and shaming. Here, large audience size and the unforgiving nature of the online world mean that such campaigns are likely to involve a net negative impact for non-domination. For a similar argument, see Norlock (2017). In the case of large companies and governments, however, reputational damage inflicted via social media campaigns is more than likely going to satisfy a test of parsimony due to existing power inequalities.

[19] Such an imperative underpins those proposed methods of counterspeech which avoid overt confrontation in 'real-world' scenarios, or which involve commitment to long-term attitudinal change. In such cases, I might, for instance, politely correct a relative who uses an outdated term to describe a member of a minority group. I might also engage in forms of counterspeech that direct attention away from the harmful content of the utterance by moving the conversation swiftly onwards or by deliberately subverting the meaning of the speaker's utterance (Caponetto and Cepollaro 2022; Fumagalli 2021). Norms which guide counterspeech in effective and respectful ways will simply be easier to foster in real-world environments where one's concern for social disapproval, especially from those one already shares a relationship with, is motivating one's behaviour.

terspeaker identity, with its own set of norms, allows others to observe and imitate 'good' practice which, over time, solidifies those norms as the commonly-accepted standard (Gerbaudo and Treré 2015). Public figures who hold substantial sway over the attitudes of the community can also have a great deal of influence over whether or not the group, and its associated norms, become established over time (Chen and Liebler 2022).

6 Conclusion

Can we rely on counterspeech to tackle the harmful effects of hateful and oppressive speech without posing the same threats to free speech typically associated with formal regulation? As this article has demonstrated, we must challenge the assumption that counterspeech is a reliably freedom-protecting alternative to formal bans. The growing influence of social media allows previously-uncoordinated individual counterspeakers to impose a risk of significant costs on others. In this way, and when those costs are uncontrolled, both formal and informal forms of harmful speech response can contribute to the unfreedom of citizens by increasing domination.

In Sect. 2, I explored, and ultimately rejected, a common assumption that freedom of speech and formal regulation are necessarily in tension with one another. Such a framing suggests that formal responses to harmful speech are, despite their potential benefits, ultimately freedom-limiting, whereas counterspeech is freedom-protecting. Where social responses *do* cause harm, it is suggested that we rely on citizen duties to refrain from using counterspeech in harmful ways. Sceptical that such duties can *reliably* protect freedom, and by adopting an alternative republican framing (Sect. 3), I argue instead that the discursive status necessary for the enjoyment of freedom of speech is not pre-institutional but is instead *conditional* on certain formal and informal conditions. Formal bans are critically assessed for their suitability in fulfilling this task (Sect. 4) and shown to potentially undermine freedom when arbitrarily enacted and applied. Counterspeech as a freedom-promoting method of responding to harmful speech is discussed in Sect. 5 and is also shown to impose risks to freedom, especially in certain discursive environments. When approaching potential harmful speech response it is thus necessary to be guided by the principle of *parsimony*, which requires taking account of the conditions (i.e., formal *and* informal norms) provided by the entire free speech infrastructure and which provide the foundation of a secure discursive status for all.

Acknowledgements I would like to thank the editorial team and the anonymous reviewers for comments on earlier drafts of this paper. Special thanks also to Keith Breen for his helpful input.

Author's Contribution Not applicable.

Funding Not applicable.

Availability of Data and Material Not applicable.

Declarations

Conflict of Interest None.

Ethical Approval Not applicable.

Informed Consent Not applicable.

Statement Regarding Research Involving Human Participants and/or Animals Not applicable.

Competing Interests Not applicable.

References

Ahrne G, Brunsson N (2011) Organization outside organizations: the significance of partial organization. Organization 18(1):83–104

Aly W, Simpson RM (2019) Political correctness gone viral. In: Fox C, Saunders J (eds) Media ethics, free speech, and the requirements of democracy. Routledge, London, pp 125–143

Anderson E (2000) Beyond homo economicus: new developments in theories of social norms. Philos Public Affairs 29(2):170–200

Anderson AA, Brossard D, Scheufele DA, Xenos MA, Ladwig P (2014) The "nasty effect:" online incivility and risk perceptions of emerging technologies. J Computer-Mediated Communication 19:373–387

Ayala S, Vasilyeva N (2016) Responsibility for silence. J Soc Philos 47(3):256–272

Baker CE (1989) Human liberty and freedom of speech. Oxford University Press, New York

Baker CE (1996) Harm, liberty, and free speech. South Calif Law Rev 70:979

Baker CE (2012) Hate speech. In: Herz M, Molnar P (eds) The content and context of hate speech: rethinking regulation and responses. Cambridge University Press, Cambridge, pp 57–80

Barendt E (2011) Religious hatred laws: protecting groups or belief? Res Publica 17:41–53

Basak R, Ganguly N, Sural S, Ghosh SK (2016) Look before you shame: A study on shaming activities on Twitter. In: Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion. ACM Press, Montréal, Québec, Canada, 11–12

Bertoni E (2011) A study on the prohibition of incitement to hatred in the Americas, OHCHR 12, 21. Available at: https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/ICCPR/NGOs2011/JBISantiagoWorkshop.pdf. Accessed 04/09/22

Billingham P, Parr T (2019) Online public shaming: virtues and vices. J Soc Philos 51(3):371–390

Billingham P, Parr T (2020) Enforcing social norms: the morality of public shaming. Eur J Philos 28(4):997–1016

Bonotti M (2017) Religion, hate speech, and non-domination. Ethnicities 17(2):259–274

Bouvier G (2020) Racist call-outs and cancel culture on Twitter: the limitations of the platform's ability to define issues of social justice. Discourse Context & Media 38:100431

Bouvier G, Machin D (2021) What gets lost in Twitter 'cancel culture' hashtags? Calling out racists reveals some limitations of social justice campaigns. Discourse & Society 32(3):307–327

Braithwaite J (1989) Crime, shame and reintegration. Cambridge University Press, Cambridge

Braithwaite J, Pettit P (1990) Not just deserts: a republican theory of criminal justice. Oxford University Press, Oxford

Brennan G, Pettit P (2004) The economy of esteem. Oxford University Press, Oxford

Brennan G, Eriksson L, Goodin RE, Southwood N (2013) Explaining norms. Oxford University Press, Oxford

Brettschneider C (2016) When the state speaks, what should it say?: how democracies can protect expression and promote equality. Princeton University Press, NJ

Broderick R, Grinberg E (2013) 10 people who learned social media can get you fired. *CNN*, 6 June. Available at: http://www.cnn.com/2013/06/06/living/buzzfeed-social-media-fired/index.html (accessed 6 May 2023)

Brown A (2015) Hate speech law: a philosophical examination. Routledge

Brown A (2017a) Hate speech laws, legitimacy, and precaution: reply to James Weinstein. Const Commentary 32(3):599–618

Brown A (2017b) The "who?" Question in the hate speech debate: part 2: functional and democratic approaches. Can J Law Jurisprud 30(1):23–55

Buerger C (2021) #iamhere: Collective counterspeech and the quest to improve online discourse 1–17

Caponetto L, Cepollaro B (2022) Bending as counterspeech. Ethical Theory & Moral Practice 1–17

Cepollaro B, Lepoutre M, Simpson RM (2022) Counterspeech Philos Compass 18(1):e12890

Chen L, Liebler CM (2022) #MeToo on Twitter: the migration of celebrity capital and social capital in online celebrity advocacy. New Media & Society *0*(0)

Clooney A, Webb PM (2017) The right to insult in international law. Columbia Hum Rights Law Rev 48(2)

Coe K, Kenski K, Rains SA (2014) Online and uncivil? Patterns and determinants of incivility in newspaper website comments. J Communication 64(4):658–679

Cohen J (1993) Freedom of expression. Philos Public Affairs 22:207–263

Cox N (2016) The freedom to publish "irreligious cartoons'. Hum Rights Law Rev 16:195

de Montesquieu S, C (1977) The spirit of laws. University of California Press, California

Delgado R (1982) Words that wound: a tort action for racial insults, epithets, and name-calling. Harv Civil Rights- Civil Liberties Law Rev 17:133–181

Delgado R, Stefancic J (1996) Ten arguments against hate-speech regulation: how valid? North Ky Law Rev 23:475–490

Delgado R, Yun DH (1994) Pressure valves and bloodied chickens: an analysis of paternalistic objections to hate speech regulation. Calif Law Rev 82:871

Detel H (2013) Disclosure and public shaming in the new age of visibility. In: Petley J (ed) Media and public shaming: drawing the boundaries of disclosure. I. B. Tauris, New York, pp 77–96

Douglas DM (2016) Doxing: a conceptual analysis. Ethics & Information Technology 18(3):199–210

Dworkin R (2009) Foreword to Hare, I., Weinstein, J. (eds.) Extreme speech and democracy. v-viii

Elford G (2021) Freedom of expression and social coercion. Leg Theory 27(2):149–175

Fadel L (2020) 'After being called out for racism, what comes next?' NPR [Online]. Available at: https://www.npr.org/sections/codeswitch/2020/07/28/891829285/after-being-called-out-for-racism-what-comes-next?t=1652091260177. Accessed 01/02/22

Friess D, Ziegele M, Heinbach D (2020) Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. Political Communication 38(5):624–646

Fritz J (2021) Online shaming and the ethics of public disapproval. J Appl Philos 38(4):686–701

Frye H (2022a) The problem of public shaming. J Political Philos 30(2):188–208

Frye H (2022b) The technology of public shaming. Soc Philos Policy 38(2):128–145

Fumagalli C (2021) Counterspeech and ordinary citizens: how? when? Political Theory 49(6):1021–1047

Galeotti AE (2002) Toleration as recognition. Cambridge University Press, Cambridge

Gelber K (2002) Speaking back: the free speech versus hate speech debate, vol 1. John Benjamins Publishing

Gelber K (2012) Reconceptualising counterspeech in hate Speech policy (with a focus on Australia). In: Herz M, Molnar P (eds) The content and context of hate speech: rethinking regulation and responses. Cambridge University Press, Cambridge, pp 198–216

Gerbaudo P, Treré E (2015) In search of the 'we' of social media activism: introduction to the special issue on social media and protest identities, information. Communication & Society 18(8):865–871

Gervais BT (2015) Incivility online: affective and behavioral reactions to uncivil political posts in a web-based experiment. J Inform Technol Politics 12:167–185

Goldberg SB (2018) Free expression on campus: mitigating the costs of contentious speakers. Harv J Law Public Policy 41:163

Goldman LM (2015) Trending now: the use of social media websites in public shaming punishments. Am Criminal Law Rev 52:415–451

Graef A 5 (2018) th, Women who claims she was fired for flipping off Trump motorcade sues former employer. CNN Politics [Online]. Available from: https://edition.cnn.com/2018/04/04/politics/woman-flipped-off-trump-sues/index.html. Accessed 05/05/23

Han S, Brazeal L (2015) Playing nice: modeling civility in online political discussions. Communication Res Rep 32:20–28

Hardin R (2009) How do you know? The economics of ordinary knowledge. Princeton University Press, NJ

Hare IM (2012) The harms of hate speech legislation. Available at: http://freespeechdebate.com/en/discuss/the-harms-of-hate-speech-legislation/. Accessed 05/12/21

He B, Ziems C, Soni S, Ramakrishnan N, Yang D, Kumar S (2021) Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp.90–94

Heinze E (2006) Viewpoint absolutism and hate speech. Mod Law Rev 69(4):543–582

Heinze E (2016) Hate speech and democratic citizenship. Oxford University Press, Oxford

Howard JW (2021) Terror, hate and the demands of counter-speech. Br J Polit Sci 51(3):924–939

Jacquet J (2015) Is shame necessary? New uses for an old tool. Pantheon, New York

Klonick K (2016) Re-shaming the debate: social norms, shame, and regulation in an internet age. Md Law Rev 75(4):1029–1065

Kolhatkar S 10 (2014) th, The rise of e new, liberal Islamophobia. Available from: https://www.commondreams.org/views/2014/10/10/rise-new-liberal-islamophobia. Accessed 03/02/23

Laborde C (2008) Critical republicanism: the hijab controversy and political philosophy. Oxford University Press, Oxford

Langton R (2012) Beyond belief: pragmatics in hate speech and pornography. In: Maitra I, McGowan MK (eds) Speech and harm: controversies over free speech. Oxford University Press, Oxford, pp 72–93

Langton R (2018a) The authority of hate speech. In: Gardner J, Green L, Leiter B (eds) Oxford studies in philosophy of law, vol 3. Oxford University press, Oxford, pp 123–152

Langton R (2018b) Blocking as counterspeech. In: Fogal D, Harris D, Moss M (eds) New work on speech acts. Oxford University Press, Oxford, pp 144–162

Lemeire O (2021) Falsifying generic stereotypes. Philos Stud 178(7):2293–2312

Lepoutre M (2017) Hate speech in public discourse: a pessimistic defence of counterspeech. Soc Theory Pract 43(4):866–867

Lepoutre M (2019) Can "more speech" counter ignorant speech? J Ethics Social Philos 16(3):155–191

Lepoutre M (2020) Hate speech laws: expressive power is not the answer. Leg Theory 25(4):272–296

Lovett F (2022) The well-ordered republic. Oxford University Press, Oxford

MacKinnon CA (1993) Only words. Harvard University Press, MA

Maitra I (2012) Subordinating speech. In: Maitra I, McGowan MK (eds) Speech and harm: controversies over free speech. Oxford University Press, Oxford, pp 94–120

Marsili M (2021) Retweeting: its linguistic and epistemic value. Synthese 198(11):10457–10483

Matsuda M (1989) Public response to racist speech: considering the victim's story. Mich Law Rev 87:2320–2381

McCarthy JD, Zald MN (2001) The enduring vitality of the resource mobilization theory of social movements. In: Turner BS (ed) Handbook of Sociological Theory. Kluwer Press, New York, pp 533–565

McGowan MK (2012) On 'whites only' signs and racist hate speech. In: Maitra I, McGowan MK (eds) Speech and harm: controversies over free speech. Oxford University Press, Oxford, pp 121–147

McGowan MK (2018) Responding to harmful speech. In: Johnson CR (ed) Voicing dissent. Routledge, New York, pp 182–200

McGowan MK (2019) Just words: on speech and hidden harm. Oxford University Press, Oxford

Meiklejohn A (1961) The First Amendment is an absolute. The Supreme Court Review 245–266

Mill JS (1969) On Liberty. In: Robson JM (ed) Collected works of John Stuart Mill, vol 10. University of Toronto Press, Toronto

Miškolci J, Kováčová L, Rigová E (2018) Countering hate speech on Facebook: the case of the Roma minority in Slovakia. Social Sci Comput Rev 38(2):128–146

Mondon A, Winter A (2017) Articulations of islamophobia: from the extreme to the mainstream? Ethnic & Racial Studies Review 40(13):2151–2179

Moran M (2003) Rethinking the reasonable person: an egalitarian reconstruction of the objective standard. Oxford University Press, Oxford

Mouk Y (2020) Stop firing the innocent. The Atlantic [Online]. Available at: https://www.theatlantic.com/ideas/archive/2020/06/stop-firing-innocent/613615/. Accessed 03/02/22

Nagel T (1998) Concealment and exposure. Philos Public Affairs 27(1):3–30

Neubaum G (2018) United in the name of justice: how conformity processes in social media may influence online vigilantism. Psychol Popular Media Cult 7(2):185–199

Nielsen LB (2012) Power in public: reactions, responses, and resistance to offensive public speech'. In: Maitra I, McGowan MK (eds) Speech and harm: controversies over free speech. Oxford University Press, Oxford, pp 148–173

Norlock KJ (2017) Online shaming. Social Philos Today 33:187–197

Papacharissi Z (2009) The virtual sphere 2.0: the internet, the public sphere, and beyond. In: Chadwick A, Howard PN (eds) Routledge handbook of internet politics. Routledge, London, pp 230–245

Pettit P (1997) Republicanism: a theory of freedom and government. Oxford University Press, Oxford

Pettit P (2012) On the people's terms: a republican theory and model of democracy. Cambridge University Press, Cambridge

Pettit P (2018) Two concepts of free speech. In: Lackey J (ed) Academic freedom. Oxford University Press, Oxford, pp 61–81

Post R (1990) Racist speech, democracy, and the First Amendment. William and Mary Law Review 32:267–328

Post R (2005) Democracy and equality. Law Cult Humanit 1:142–153

Post R (2011) Participatory democracy and free speech. Va Law Rev 97:477–489

Radzik R (2016) Gossip and social punishment. Res Philosophica 93(1):185–204

Radzik, R., Bennett C, Pettigrove G, Sher G (2020) The ethics of social punishment: the enforcement of morality in everyday life. Cambridge University Press, Cambridge

Rini R (2020) The ethics of microaggression. Routledge, Oxford

Robards B, Graf D (2022) "How a Facebook update can cost you your job": News coverage of employment terminations following social media disclosures, from racist cops to queer teachers. Social Media & Society 1–22

Ronson J (2015) So you've been publicly shamed. Riverhead, New York

Rostbøll CF (2015) Non-domination and democratic legitimacy. Crit Rev Int Social Political Philos 18(4):424–439

Santana AD (2014) Virtuous or vitriolic: the effect of anonymity on civility in online newspaper reader comment boards. Journalism Pract 8:18–33

Saul JM (2018a) Dogwhistles, political manipulation, and philosophy of language. In: Fogal D, Harris D, Moss M (eds) New work on speech acts. Oxford University Press, Oxford, pp 360–383

Saul JM (2018b) Beyond just silencing: a call for complexity in discussions of academic free speech. In: Lackey J (ed) Academic freedom. Oxford University Press, Oxford, pp 119–134

Saul JM (2021) Someone is wrong on the internet: is there an obligation to correct false and oppressive speech on social media? In: MacKenzie A, Rose J, Bhatt I (eds) The epistemology of deceit in a postdigital age: Dupery by design. Springer, Switzerland, pp 139–157

Shiffrin S (2011) Freedom of speech and two types of autonomy. Const Commentary 27:337–345

Skinner Q (2010) On the slogans of republican political theory. Eur J Political Theory 9(1):95–102

Smith A (2010) A theory of moral sentiments. Penguin, New York

Stanley J (2015) How propaganda works. Princeton University Press, NJ

Strossen N (1990) Regulating racist speech on campus: a modest proposal. Duke law J 484

Strossen N (2012) Interview. In: Herz M, Molnar P (eds) The content and context of hate speech: rethinking regulation and responses. Cambridge University Press, Cambridge, pp 378–398

Strossen N (2018) Hate: why we should resist it with free speech, not censorship. Oxford University Press, Oxford

Stroud SR (2016) Be a bully to beat a bully: Twitter ethics, online identity, and the culture of quick revenge. In: Davisson A, Booth P (eds) Controversies in digital ethics. Bloomsbury Press, London, pp 264–278

Sunstein C (1995) Democracy and the problem of free speech. Publishing Res Q 11(4):58–72

Sunstein C (1996/1997) Social norms and social roles. Free markets and social justice. Oxford University Press, Oxford, pp 32–69

Teekah A (2015) Lessons from SlutWalk: how call-out culture hurts our movement. Herizons 29(2):16–21

Thomason K (2021) The moral risks of online shaming. In: Veliz C (ed) The Oxford handbook of digital ethics. Oxford University Press, Oxford

Tirrell L (2018) Toxic speech: inoculations and antidotes. South J Philos 56(1):136

Tirrell L (2019) Toxic misogyny and the limits of counterspeech. Fordham Law Rev 87(6):2433–2452

Tosi J, Warmke B (2016) Moral grandstanding. Philos Public Affairs 44(3):197–217

Waldron J (2012) The harm in hate speech. Harvard University Press, MA

Watkins D (2015) Institutionalizing freedom as non-domination: democracy and the role of the state. Polity 47(4):508–534

Whitney v. California, 274 U.S. 357 (1927)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.