Article

Terror, Hate and the Demands of Counter-Speech

Jeffrey W. Howard

Department of Political Science, University College London

E-mail: jeffrey.howard@ucl.ac.uk

(Received 23 July 2018, revised 18 March 2019, accepted 22 August 2019)

[BEGIN ABSTRACT]

It is a familiar mantra of American politics that the best response to dangerous speech that incites violence and spreads hate is 'more speech'. Yet the principle obscures at least three crucial questions. Who, in particular, is to undertake the counter-speech that the doctrine recommends? What, exactly, are they required to do? And why is it morally justified to demand that they do it? The author argues that counter-speech should to be relied on, at least in part, to defuse the dangers of dangerous expression, but that it is not enough to cheerlead its abstract importance and then sit back and hope for the best. Someone needs to do the work, and do it well. This article defends the thesis that all citizens have a moral duty to engage counter-speech against dangerous expression. Focusing on counter-speech against expression that implicitly or explicitly advocates wrongful criminal violence, the author argues that these duties can be derived from a much more basic normative source: the samaritan obligation, held by all moral agents, to rescue others from risks of harm. The specification of these duties' content, however, depends upon interdisciplinary work that integrates normative theory with social scientific research on human communication.

[END ABSTRACT]

Commented [KF1]: This is capitalized in the OED – before changing it throughout I wanted to check whether you are using lowercase for a particular reason?

Keywords: free speech; hate speech; dangerous speech; counter-speech; samaritanism; civic duties

[BEGIN TEXT]

The United States stands in marked contrast to most of the world's liberal democracies in extending legal protection to speech that implicitly or explicitly advocates violent criminal conduct. The First Amendment to the US Constitution protects white supremacists championing vigilante lynching just as it protects religious fanatics on YouTube advocating the murder of infidels. In the famous words of Judge Louis Brandeis, where circumstances allow, 'the remedy to be applied' to dangerous expression 'is more speech, not enforced silence' (*Whitney v. California*, 274 U.S. 357 (1927)).

This idea – that 'the best response to speech that government deems dangerous…is more speech' (Sunstein 2018, 248) – is among the most distinguishing features of American political thought and practice. Yet the principle obscures at least three crucial questions. *Who*, in particular, is to undertake the counter-speech that the doctrine recommends? *What*, exactly, are they required to do? And *why* is it morally justified to demand that they do it? If we are to rely on counter-speech, at least in part, to defuse the dangers of dangerous expression – and I will argue that we should – it is not enough to cheerlead its abstract importance and then sit back and hope for the best. Someone needs to do the work, and do it well.

Unfortunately, contemporary political theory does not have an adequate answer to these questions of *who*, *what* and *why*. As a result, the 'more speech' slogan, so prominent in popular debate, continues to lack a sound normative

foundation. This article contributes to its construction. I defend the thesis that all moral agents have the duty to engage counter-speech against dangerous expression. Focusing on counter-speech against expression that implicitly or explicitly advocates wrongful criminal violence, I argue that these duties can be derived from a much more basic normative source: the samaritan obligation to rescue others from risks of harm. The task of combatting extremist ideas that maliciously encourage violent attacks on individual rights, either directly or indirectly, falls on us all.

Crucially, my thesis is compatible with any number of positions one might take on whether dangerous speech should be banned. Brandeis's canonical 'more speech, not enforced silence' formulation implies that we must choose between criminalization *or* counter-speech. But this is a false dichotomy. It is perfectly coherent to advocate banning dangerous expression, such as hate speech or terrorist advocacy, while nevertheless enjoining counter-speech as a vital supplement. There are several plausible views one can take on the issue of criminalization, depending on perceptions of what is the proper scope of the legal doctrine of freedom of expression. I take no stand here on the ongoing debate as to which of these views is correct. My core thesis is compatible with them all.

The argument that all moral agents are obligated to confront dangerous speech through counter-speech invites the following objection. The empirical evidence shows that human beings are often resistant to evidence or arguments that contradict their beliefs; indeed, exposure to countervailing views may even backfire, leading people to grow more confident in their initial views. How could it be, then, that people have a duty to defuse dangerous ideas by arguing against them, if such efforts may face considerable difficulties?

Commented [KF2]: Sbould this be 'engage in'?

Commented [KF3]: Or simply "people"

The response to this concern is not to rebut it, but rather to assimilate its insights into my specification of the core argument. The samaritan basis of the view to be defended clarifies that the duty at stake is not a duty to engage in counter-speech for its own sake, or for the symbolic benefits of solidaristic expression. Its point is to rescue others from harm. Thus the duty cannot be simply to *engage* in counter-speech; it must be to engage in *effective* counter-speech that is designed to overcome the psychological obstacles that would otherwise inhibit its success.

This requirement has implications for the matter of *what*, exactly, counter-speech duties require of us. I argue that the specification of these duties' concrete demands is unavoidably the task of interdisciplinary work that integrates social science into normative theory. Unless we take seriously empirical findings about how human beings are and are not persuaded effectively, we risk encouraging counter-speech that is either ineffective or downright counterproductive. This is an important result: ham-fisted counter-speech that is likely to counterproductively provoke or inflame dangerous attitudes is seriously objectionable because it further endangers the very people it intends to assist. Just as effective counter-speech is morally required, I will argue, counterproductive counter-speech is morally forbidden.

The next two sections set the stage for the argument by reviewing some limits of the current scholarly treatment of counter-speech. Then, I provide my account of counter-speech duties and pinpoint their samaritan justification. I then present the aforementioned empirical objection and update my specification of the argument accordingly. I conclude by tracing some implications of recent empirical scholarship for the content of counter-speech duties.

[LEVEL A HEADING] BYPASSING THE IMPASSE

How should we respond to *dangerous speech*: speech that implicitly or explicitly encourages wrongful criminal violence, such as speech that advocates terrorism or incites racial hatred?[1] As mentioned briefly above, US Supreme Court Justice Louis Brandeis most famously articulated the idea that counter-speech is the appropriate response. His euphonious prose no doubt accounts for at least part of the idea's enduring appeal:

[BEGIN DISPLAYED QUOTE]

To courageous, self-reliant men…no danger flowing from speech can be deemed clear and present unless the incidence of the evil apprehended is so imminent that it may befall before there is opportunity for full discussion…[T]he remedy to be applied is more speech, not enforced silence (*Whitney v. California* 274 U.S. 357 (1927)).

[END DISPLAYED QUOTE]

This position eventually became the official view of the U.S. Supreme Court in *Brandenburg v. Ohio* 395 U.S. 444 (1969).

What is the putative philosophical justification of this view? The standard justification of this position is that counter-speech is to be preferred because it is the only morally available remedy. For most free speech theorists in the American tradition, criminal sanctions on dangerous speech would simply violate a properly specified legal right of freedom of expression (see Howard 2019 for a review of this debate, focusing on hate speech). Among philosophers, this position is typically grounded in deontological moral arguments about the autonomy interests of moral agents *qua* speakers (Baker 1978), listeners (Scanlon 1972; Strauss 1991), thinkers (Shiffrin

[1] In using the term dangerous speech, I follow Benesch (2012).

Commented [AQ4]: AQ: References [Scanlon (1972), Strauss (1991), MacKinnon (1987), Sanders (1997), Clayton and Stevens (2013), Myers (1975), Eagly, Wood, and Chaiken (1978), Haidt (2012), Rebasa et al. (2010) and Leydon and Cook (2008)] is cited in text but not provided in the list. Please provide complete publication details to insert in the list, else delete the citation from the text.

2014) or democratic citizens (Heinze 2016). Among lawyers, there also tends to be an instrumental appeal to the inherent dangers of granting the state the power to regulate speech it disfavours (for example, Stone 2004). For all these theorists, the endorsement of counter-speech traces back to the fact that it is one of the only remedies properly at our disposal for combatting dangerous speech, given the impermissibility of criminal sanction as a mode of response. Many – probably most – of counter-speech's explicit champions embrace this rationale (for example, Brettschneider 2012).

Yet one need not advocate such a strident view of the protective scope of free speech in order to endorse a role for counter-speech. For starters, even those who believe the legal right to free speech does not properly protect dangerous expression – who think such speech falls outside the right's protective ambit – may still favour counter-speech over criminalization if the former is more effective at combatting dangerous ideas.

Further, as mentioned at the outset, Brandeis's suggestion – that we face a choice of whether to criminalize dangerous speech, or else permit it and argue against it – presents us with a false dichotomy. There is no reason why counter-speech and forms of legal redress could not operate together (cf. Gelber 2012b, 214). Those who doubt that dangerous speech is properly protected by the legal right to freedom of expression could well recommend the criminalization of (some or all) dangerous speech alongside a continued effort to argue against it (in this spirit, see also McGowan 2018).

So it would be a mistake to assume that we cannot make progress on the topic of counter-speech before addressing the fundamental question of whether dangerous speech is properly protected by the legal right to free speech. The question of whether

to ban such speech, and the whether to argue back against it, are orthogonal.[2] This is obvious from the vantage point of an ordinary citizen wondering whether she ought to argue back against dangerous speech. While the ordinary citizen of course may wonder whether she should support campaigns for greater regulation, the immediate normative question when she encounters dangerous speech is not *should we ban this speech*?

Indeed, for US citizens the questions of criminalization and counter-speech are palpably orthogonal, as they are not even legally entitled to pursue the former matter, given the Supreme Court's prevailing interpretation of the First Amendment. Likewise, even in jurisdictions such as England and Wales, where inciting racial hatred and encouraging terrorism are crimes, citizens are not relieved of the question of whether to argue back against hatred in their midst. Much hateful speech will simply not (obviously) run afoul of any criminal statute. Consider, paradigmatically, cases in which someone is expressing sympathy for extremist ideas without offering a full-throated endorsement of them. Here, too, the normative questions of whether to engage in counter-speech – and if so, what sort, and why – arise. Therefore my argument seeks to bypass the present impasse over whether dangerous speech is properly criminalized.

Commented [KF5]: Or: still wonder/still debate?

A final preliminary remark. One pay-off of disentangling the analysis of counter-speech from the vexed question of criminalization is that it may help us to guard against a certain form of complacency that can afflict the defenders of counter-speech. The complacent assumption I have in mind is that, once criminalization is ruled out and dangerous speech is permitted, the requisite counter-speech will just

[2] I am grateful to two anonymous reviewers for pressing this point.

happen – we just have to sit back and wait for it. The traditional defenders of a 'marketplace of ideas' as the rationale for free speech often appear to make this sort of assumption:

[BEGIN DISPLAYED QUOTE]

> [I]f only truth is left free to combat error, in an open market-place of ideas, humanity is bound to become more enlightened and better off (Bay 1975, 391).

[END DISPLAYED QUOTE]

I return to this much-maligned notion of a marketplace of ideas below. For now, I simply want to stress an obvious but still under-acknowledged point: just as there is no guarantee that the truth *will* prevail in a marketplace of ideas, there is no guarantee that the requisite counter-speech *will just happen*. Someone will have to do the work, and do it well. Assessing *who* should do it, explaining *what* exactly they should do, and justifying *why* it can rightly be demanded of them is the substantially neglected task of a normative account of counter-speech, to which I now turn.

[LEVEL A HEADING] TOWARD A THEORY OF COUNTER-SPEECH DUTIES

By engaging in counter-speech, we may accomplish at least two goals. First, we can potentially prevent dangerous speakers from persuading listeners to adopt false beliefs about what they ought to do, as those listeners are exposed to compelling counter-arguments. Secondly, we can potentially enable a moral transformation of the speakers themselves, as they may become convinced that their initial views are mistaken. But is there a *duty* to engage in speech of this sort? If so, who holds it?

My thesis is that ordinary moral agents – regular citizens like you and me – have a duty to challenge dangerous ideas through counter-speech. This idea cannot possibly be an original one; the American celebration of counter-speech seems to presuppose it. It is, in this respect, a moral insight hiding in plain sight. But because so few, if any, have bothered to defend this idea explicitly, it is massively under-theorized. Crucially, the scholarly literature lacks an argument for the claim that ordinary citizens ought to engage in counter-speech against dangerous expression. In other words, we need a moral justification that explains why they bear the costs that the duty imposes on them.

Commented [KF6]: Or simply: I argue

Despite the absence of such an argument from the literature on freedom of expression, there is some related work that sets us down the right path. Consider, first, Corey Brettschneider's argument that *the state* ought to combat hateful expression through government speech that aims to persuade illiberal citizens 'to adopt the values of equal citizenship' (2012, 72). Brettschneider's position productively undermines the orthodox, binary view that the state must either ban hate speech or let it run rampant.

Still, his argument, which focuses on government speech, has limited applicability to my subject for two reasons. First, there are obvious strategic questions about whether the liberal state is an effective vehicle for persuading adult citizens who condemn liberal values to abandon their views. Secondly, and more importantly for the present purposes, Brettschneider's argument only justifies why *the state* should engage in such counter-speech; indeed, this is all it sets out to justify. His view is that, by permitting hateful speech, the liberal state risks leading its citizens to believe that it actively endorses or is otherwise untroubled by hateful views. Silence, Brettschneider thinks, would constitute *complicity* in their perpetuation (71). The state

thereby has a duty to speak in order to prevent such complicity. But this is merely an argument for what the state must do; it does not explain what ordinary citizens must do.[3] Brettschneider himself briefly affirms the idea that citizens also ought to engage in speech that reaffirms their commitment to liberal democracy (e.g., 79–80), but does not explain why.

Commented [KF7]: I think this is an important point that should be moved to the main text.

Outside the scholarship on free speech, there has been some related debate among scholars of Rawls's political liberalism about the proper way to respond to so-called unreasonable citizens, who embrace illiberal or inegalitarian ideas. Matthew Clayton and David Stevens, for example, have explored the idea that reasonable religious citizens could engage with unreasonable co-adherents to persuade them to adopt a better interpretation of their shared religion (2013, 75). Yet Clayton and Stevens merely defend the claim that reasonable citizens have reason to engage with unreasonable views (80); they explicitly refrain from taking a stand, as I shall, on whether citizens have a *duty* to do so (81n).[4]

Commented [KF8]: Perhaps be a bit more specific here – e.g. more inclusive?

[3] For a related argument, which contends that the state ought to support the activity of counter-speech through various policy measures, see Gelber (2012b), especially at 214. For a further (qualified) defense of state counter-speech, see Lepoutre (2017).

[4] Clayton and Stevens (2013) suggestively float, though do not pursue, the idea that *the natural duty of justice* (81) could be a possible justification. As one reviewer has suggested to me, building on this suggestion, our duty to support just institutions might well enjoin us to take measures to stop others from violating those institutions' demands – and counter-speech might be one of these measures. I view this possible argument for counter-speech duties as complementary to the one I develop here. I focus on the samaritan argument instead for two reasons. First, it relies on a less controversial premise: the idea that there are general duties of rescue is widely accepted, whereas the natural duty of justice is less so. Secondly, the samaritan argument has a wider scope since, for example, it explains why counter-speech could be required in a state of nature (where there are no just institutional demands to

Gabriele Badano and Alasia Nuti have gone further, arguing that, according to a Rawlsian view, reasonable citizens have a 'duty of pressure' (2017, 4) toward unreasonable citizens to try to change their views, focusing on the case of citizens who support 'right-wing populist' political parties. While they (like Clayton and Stevens) do not explicitly connect their claims to debates about free speech and counter-speech, I view their excellent proposal as allied with the overall research agenda that I am advocating in this article. However, their argument has limited applicability to my topic here. For starters, I worry whether their proffered justification succeeds even for the case in which it is meant to apply. They derive their 'duty of pressure' from Rawls's 'liberal principle of legitimacy, which requires that state decisions about the most fundamental political questions be settled on the basis of public reasons' (12) – reasons justifiable to all as free and equal. But since the state must decide questions of basic justice in accordance with public reason, it need not follow that discursive engagement with unreasonable co-citizens is morally required. As long as citizens who participate in politics are discharging *their* duty of civility – to 'explain to one another on…fundamental questions how the principles and policies they advocate and vote for can be supported by the political values of public reason' (Rawls 2005, 217) – they can plausibly insist that *they* have done their due diligence in respecting their fellow citizens.

But even if Badano and Nuti's justificatory argument succeeds, it cannot justify the kinds of counter-speech duties to combat dangerous speech of the sort I am proposing here. My concern here is not with speech that risks inspiring voters to enact

be violated). More speculatively, if the justification for an institutional rule banning murder is itself the moral imperative to rescue people from murder, it may be that the institutional argument is itself derived from the samaritan one (cf. Wellman 1996).

right-wing populist legislation; it is with speech that risks inspiring agents to engage in criminal violence. An argument focused squarely on the legitimacy of state policy does not straightforwardly justify the responsibility of ordinary citizens to argue back against this crucial category of dangerous speech.[5]

[LEVEL A HEADING] COUNTER-SPEECH DUTIES

[LEVEL B HEADING] *A Samaritan Justification*

How, then, are citizens' counter-speech duties justified?[6] I propose that we view such duties as derived principally from a much more basic moral obligation: *samaritan duties*, 'our duties to rescue others from peril when this assistance is not unreasonably costly' (Wellman 2013, 86).[7] Such duties are *natural*, in that they apply to all moral agents simply by virtue of being moral agents; as such, natural duties 'apply to us without regard to our voluntary acts' (Rawls 1999[1971], 98). These duties clearly apply between fellow citizens of a political community, but are not restricted to co-

[5] Another relevant body of work, which I do not discuss here, concerns 'reasoning from conjecture'. According to Rawls, this occurs when 'we argue from what we believe, or conjecture, are other people's basic doctrines, religious or secular, and try to show them that, despite what they might think, they can still endorse a reasonable political conception that can provide a basis for public reasons' (2005, 594). Scholars have defended the permissibility of this practice (see Schwartzman 2012), though no one has defended it as required. If this form of counter-speech proves empirically promising, my account here could recommend it. For a compelling criticism of reasoning from conjecture in the context of counter-terrorism policy, see Badano and Nuti (2019).

[6] I use the terms 'citizens' and 'agents' interchangeably.

[7] See Wellman (1996, 211–237) for a fuller defense of the idea of samaritan duties, which take their name from the parable Jesus tells in Luke 10:25–37; in its current usage, it need not have religious implications.

nationals; following Rawls, 'they obtain between all as equal moral persons' (99). These duties enjoin us to rescue our fellow moral agents from harm, whether this is in the form of rushing rivers, falling rocks, wild animals or indeed human beings. Such duties require whatever is necessary and proportionate, provided it is not unreasonably burdensome, to rescue our fellow human beings – to swim, jump, punch or even kill.[8] My novel proposal is that samaritan duties also sometimes enjoin us *to speak*.

Consider the following hypothetical to illustrate such a samaritan duty. Suppose that a speaker, Samantha, is attempting to persuade a listener, Lenny, that he has an obligation to kill an innocent victim, Victoria, at some later date (perhaps because of Victoria's race, sexual orientation or religious views). Bart, a bystander, can take action to stop this. Suppose the only available[9] course of action that Bart can take is to try to persuade Lenny not to engage in the attack against Victoria by pointing out the falsehood of the relevant beliefs. Suppose also that doing so is not unreasonably burdensome for Bart. I argue that the samaritan duty of rescue enjoins Bart to rescue Victoria through the medium of speech.

This proposal also applies to cases in which counter-speech directly operates to affect a dangerous party. Suppose that the speaker, Samantha, reveals that she is contemplating attacking Victoria. Again supposing that the only course of action that

[8] For a discussion of the moral duty to kill in order to rescue others, see Fabre (2007).

[9] It might be the only *legally* available course of action if the legal right to free speech in the jurisdiction rules outs coercively suppressing speech. Or it might be the only *empirically* available course of action; even in jurisdictions where such speech is criminalized, there are numerous cases in which coercion is nevertheless not a feasible option (e.g., because the speech, while dangerous, does not meet the relevant legal threshold).

Bart can take is to resort to speech, and that it is not unreasonably burdensome for him to do so, we should consider Bart to be under a duty to (try to) stop the attack.

The general idea that we have natural duties to defuse unjustified threats posed by others is not controversial. The literature on defensive harm analyzes such 'defense-of-other' samaritan duties (for example, Fabre 2007). These cases typically involve inflicting harm on an assailant who poses an unjust threat. The familiar idea is that when an agent culpably initiates an unjust attack, she is liable to suffer some harm – that is, she is not wronged by its imposition – deployed against her defensively (provided it is necessary and proportionate).[10] Note that the threat need not be certain to materialize; an evil sniper may simply be a bad shot. Yet the duty to attempt a rescue still applies. What matters is that an innocent person faces 'peril' (Wellman 2013, 86) – that is, a grave *risk* of harm – from which a bystander can rescue her.

Notice that counter-speech cases are distinct from traditional defense-of-other rescue cases, since they do not entail the application of force.[11] They involve *talking*, not *harming*. This makes counter-speech duties easier to justify than duties in traditional defense-of-other cases. Even if communication with others might qualify as bothersome or unwanted, and in sufficient volume become unwarranted harassment, the act of communicating with somebody does not ordinarily qualify as harmful. In these cases, then, the question of whether the discursive 'force' used to rescue the victim from attack is disproportionate simply will not arise. Even if being

[10] See McMahan (2005, 386–405) and Quong (2012, 45-77).

[11] In other work, I explore the claim that counter-speech should be preferred over coercion because the prospect of its success means that coercion is unnecessary, and thus involves imposing gratuitous harm. Nothing in my argument here, however, depends upon that claim.

persuaded to abandon bad ideas might be a psychologically stressful experience, this cannot constitute a setback to an agent's interests, and thereby cannot be considered to cause harm.

But counter-speech cases differ from traditional defense-of-cases in another respect, which may make counter-speech duties *more* difficult to justify. In the cases that concern us here, the target of one's counter-speech has not necessarily formed any intention to perpetrate a crime. Our focus is temporally upstream, concerning the normative beliefs that lead agents to form evil intentions. This fact may seem to make counter-speech duties more difficult to justify, since the threat is merely a potentiality – something the agent *may* end up doing, to victims whose identities are not yet known. This skepticism is motivated by an understandable concern for moral parsimony, for not overcrowding the domain of duties.

But we should reject it. It would be strange for morality to be silent on our duties to prevent wrongdoing until the very moment that a plan to perpetrate wrongdoing is formed and underway, at which point morality suddenly fires up its concern. By contrast, I suggest that the interests of innocent citizens that justify more familiar duties of rescue also justify duties to reduce the incidence of cases in which rescue is necessary. Consider the ordinary rescue case: whatever interests justify the duty to rescue people from drowning surely also justify putting up signs warning people of dangerous currents, telling swimmers approaching the beach of the dangers they might face, and so on. Likewise, given that a subset of agents is sufficiently likely to form an intention to do wrong on the basis of dangerous ideas – and given that they pose substantial dangers, and that steering them away from evil ideas is likely to be more successful if it occurs before they have committed to a criminal plot – it would be puzzling for morality to ignore all of that. This point is especially apt in

Commented [KF9]: I recommend specifying what you mean here – this skepticism?

the context of dangerous hate speech, in which the mechanism of inspiring violence is diffuse: those who engage in hateful violence may well have been inspired by multiple sources over a period of time. As Bhikhu Parekh explains, 'A vicious and widespread hatred of a group does not spring up overnight.' Rather, '[i]t builds up slowly through isolated utterances and actions…all cumulatively capable of coarsening the community's sensibility, poisoning the minds of the young … and creating a situation in which it becomes a common practice to…show hostility to the target group' (2006).

The upshot is that we cannot simply wait for cases in which violence is imminent to speak up. No doubt we are required to speak up in these cases, where speech is particularly dangerous. But samaritanism is more demanding than that: it requires earlier intervention when it is not unreasonably costly. In the standard case, for example, it requires us to caution a child not to jump into a fast-moving current, to prevent the dire emergency from materializing. It may even require the cultivation of a certain samaritan disposition, to be on the lookout for cases in which help is needed. Below I explain that such a disposition need not be all-consuming. But we would think poorly of someone who was so inwardly focused that she did not notice the drowning child nearby. Indeed, insofar as one knows that there are many ponds in the area, with a continuing problem of children drowning in them, we would think it incompatible with a samaritan disposition to avoid such ponds. Likewise, insofar as our communities are filled with dangerous speech, it is incumbent upon us all to be on the lookout for opportunities to rebut it as we move through our daily lives.

[LEVEL A HEADING] SPECIFYING THE BURDENS

I have argued that counter-speech duties are easier to justify than other samaritan obligations because they do not involve the imposition of harm. Still, even if *receiving* communication does not qualify as especially costly, *engaging* in communication can be. An account of counter-speech duties, then, must address the burdens to be borne by those ordinary citizens saddled with them. All natural duties are subject to the condition that they are not unreasonably burdensome (Rawls 1999[1971], 294). If a putatively mandatory action will require one to sacrifice one's interests to an unreasonable extent, the action is not actually required. The crucial question concerns what 'unreasonable extent' entails. So just how onerous *are* counter-speech duties?

Because counter-speech duties are simply a type of samaritan duty, the answer to this question will not be *sui generis*. The right theoretical account of what qualifies as 'unreasonably burdensome' for samaritan duties generally will apply to counter-speech duties in particular. I submit that we do not need to settle that vexed debate[12] to accept that counter-speech duties exist; our general argument is consistent with any number of different views one might adopt about how demanding samaritan duties might be.

[12] Suppose I am rescuing someone from death. Just how burdensome does the rescue need to be before I am no longer required to do it? And how many rescues am I required to attempt? One answer, defended by Cécile Fabre, holds that a given rescue is unreasonably demanding if it 'would prevent the rescuer from leading a flourishing life…that is, being able to frame, revise and pursue a meaningful conception of the good' (2007, 365). While this is an attractive approach, or at least the basis of one, nothing I say here depends on this particular view.

Whatever the level of maximal demandingness might be, what is uncontroversial is that the demandingness of any given samaritan duty will partly be a function of, among other variables, (a) the gravity of the harms one is endeavoring to prevent and (b) the risk that they will eventuate. Both variables bear on our determination of the dangerousness of a given instance of dangerous speech (among others, such as, potentially, imminence). Speech that defends the moral imperative of mass terrorist attacks, to an audience susceptible to that message, is more dangerous than speech that defends lesser crimes to audiences less disposed to be persuaded. Because it is plausible to demand greater efforts from citizens in response to more dangerous speech, it is important to identify when, exactly, speech is dangerous.

Clearly empirical analysis is indispensable here (for example, Leader and Benesch 2016) – a point to which I return in the next section. Even so, we lack a sufficiently deep understanding of human behavior to make confident statistical predictions in this area. Thus individual agents must determine the dangerousness of speech on a case-by-case basis. It is therefore fitting to regard counter-speech duties as *imperfect* duties – duties that are exercised based on the discretion of individual agents.[13] The degree of empirical uncertainty makes it enormously difficult to specify *ex ante* what, exactly, each individual is required to do in specific circumstances; any attempt to render counter-speech duties 'perfect' would face this presently insurmountable epistemic hurdle.

Still, regardless of one's view of the demandingness of samaritan duties, it is obvious that counter-speech will be more burdensome for some people than it will be

[13] For a discussion of imperfect duties, see Hope (2014). See also Badano and Nuti (2017, 20), who describe the duty to pressure right-wing populists as imperfect.

for others. There is an important difference between a discursive encounter that moderately burdens the counter-speaker and one that exposes her to a considerable risk of death – for example, by requiring that she engage with dangerous people or incur risks of violent retaliation. This is particularly true of counter-speakers who are, at the same time, the prospective targets of dangerous speech. So far I have suggested that the job of counter-speech largely falls on bystanders who are not themselves endangered by the speech in question (cf. Gelber 2012b), but nothing in the argument restricts the duty to them alone. So the question arises: are targets of dangerous speech – paradigmatically, members of vulnerable groups smeared by hate speech – obligated to speak out against it?

The worry is that these individuals are already subjected to the intimidation, harassment and endangerment that exposure to such speech already involves. To insist that they are then supposed to confront such speech – rather than, say, walk away from or avoid it – thereby compounds these very harms or risks of harm. As Ishani Maitra and Mary Kate McGowan note, this seems patently unfair (2012, 9; see also Schauer 1992). Does my view nevertheless require it?

Note that if a target of dangerous speech is required to argue back, it is not because she owes it to herself to resist her own oppression, as some theorists argue (for example, Hay 2011; Silvermint 2013). The samaritan duty I have specified is other-regarding, not self-regarding. So the argument would need to be that targets of dangerous speech are required to argue back in order to help protect their *fellow* oppressed citizens from dangerous ideas. In this spirit, Ashwini Vasanthakumar argues that victims of injustice have a duty to assist their fellow victims, which she argues they can discharge through their testimony about their experience with injustice. Explicitly addressing the objection of over-demandingness,

Vasanthakumar's response is that so long as victims are not asked to sacrifice a basic interest (for example, risk their life), they have a duty to act (2018, 475).

I believe a reply along these lines is unavoidable so long as we view victims of injustice, as of course we should, as moral agents (this is also a central theme of Shelby 2018). Still, the seeming coarseness of this reply can be tempered by a few observations. For instance, there is the question of whether the targets of dangerous speech are the most effective speakers to defuse the speech's danger (cf. Gelber 2012a, 52). Later on, I will argue that when counter-speech is foreseeably ineffective, there is no requirement to engage in it. If targets of dangerous speech are poorly placed to persuade those verbally attacking them to revise their convictions, they are released from their duty to argue back. Further, even if victims of injustice can sometimes be expected to engage in counter-speech, they are likely to hit the maximum threshold of unreasonable burdensomeness – after which counter-speech becomes supererogatory – much more quickly than privileged bystanders. In other words, even if victims of injustice are still subject to counter-speech duties, it is plausible that these duties simply demand less of them, since in their case discharging these duties involves distinctive psychological and physical risks.

Commented [KF10]: I recommend this change as there is no clear first/secondly/thirdly framework here.

A final thought on burdensomeness, which applies to all prospective bearers of counter-speech duties (privileged and oppressed alike): just how much one ought to do on any given occasion may depend on whether there are other people or institutions working to defuse the threat of the dangerous speech in question. *Ideally*, of course, everyone should be undertaking his or her fair share of the samaritan workload. To return to the more familiar rescue context: if there are 100 children drowning and 100 swimmers on the beach, it is fair that each rescues one. But if some swimmers are refusing to do their share, it is not obvious that the morally motivated

are still only expected to rescue one each and then call it a day. Intuitively, there is a moral duty to pick up at least *some* of the slack – to rescue some more children – unfair though it is.[14] The maximum amount of work one can, in principle, be expected to do is thus not plausibly a function of how many others are helping; there is a certain number of rescues one is expected to do before the rescues become supererogatory, and that number does not depend on how many others are (or are not) helping in any given case. But when others *do* help, then in these cases one's moral workload can sometimes fall below the maximum threshold, as the work is shared out fairly.

How, exactly, this works in the context of counter-speech is clearly complicated. Simple cases of swimmers and drowning children make for good heuristic examples, but the real world – including for the direct intended site of application of those examples (that is, global poverty) – is much more complex (Wenar 2011). Still, if few voices are speaking out against hate in one's community, it seems plausible that one should feel a greater responsibility to do so. Likewise, if many voices are speaking out against hate, then although one should also do so, it is conceivable that one is not required to do as much. I say *conceivable* because it all depends on just how pervasive the hate is, and on the level of toxicity in the political culture. In some cases everyone will need to do the maximum that could be required of them (the counter-speech equivalents of cases in which there are 100 swimmers but millions of drowning children). In other cases, the overall workload will be manageably sized, such that an effective division of labor will leave each counter-speaker with less than her maximum possible allocation.

Commented [KF11]: Perhaps clarify what you mean here.

[14] This view is convincingly defended in Tadros (2016).

This point illuminates why my argument may have subtly different implications for different legal regimes. In jurisdictions such as the UK and Australia, counter-speakers have an important form of assistance in combatting dangerous messages – criminal law. It is conceivable that, in these jurisdictions, the demands of counter-speech are less stringent, simply because counter-speech need not play as much of a role, relative to the world in which there is no criminal assistance.[15] By contrast, in the United States non-legal modes of response are the only line of defense against dangerous speech.

Commented [KF12]: Countries that have not outlawed hate speech?

We should be hesitant to draw any blanket conclusions. The claim that citizens in jurisdictions with stronger speech restrictions have less stringent counter-speech duties is conditional on the speech restrictions' success in reducing the overall amount of danger faced. Further, even if restrictions are successful, in truly toxic political cultures the supplementary work required of counter-speakers may still be significant. So speakers should not overestimate the ramifications of existing legislation for the stringency of their counter-speech duties. While it is possible that legislation could render these duties less stringent, it is far from straightforward.

But at least one ramification is clear: if there is no such legislation – if we choose to forego criminalization on the assumption that counter-speech can instead do the requisite job – we cannot rest on our laurels. We must ensure that the job is done, and done well.

[15] I assume without argument that it is manifestly implausible that the criminal law would be *sufficient* in mitigating the dangers of dangerous speech, rendering counter-speech unnecessary.

[LEVEL A HEADING] IMPUGNING THE EFFICACY OF COUNTER-SPEECH

The argument that we should reduce others' susceptibility to harm through counter-speech – thereby rescuing them from the dangers that would otherwise befall them – presupposes that counter-speech will work. Yet political theory has long questioned whether speech is an effective way to free people from the grip of mistaken convictions. To vindicate the argument, we need to confront that skepticism head on.

J.S. Mill inspired the view that, through the collision of ideas, we can persuade those who are wrong. He famously denounces 'the peculiar evil of silencing the expression of an opinion' since, by doing so, those who hold the opinion 'are deprived of the opportunity of exchanging error for truth' (Mill 1991[1859], 21). Crucially, a common way to interpret Mill's point is that it highlights the likely effects of unfettered expression (for example, Bay 1975, 391; though cf. Gordon 1997).

No general argument in favor of freedom of speech has received such sweeping condemnation as the idea that it promotes the acquisition of truth through the free market of ideas. [16] There are plenty reasons to doubt that, in unrestricted debate, the truth is simply fated to prevail. Even if we are confident that the truth is likely to emerge in contexts in which there is equal power and standing between all those involved in the deliberation, in real-world deliberations, there are considerable asymmetries of power, status and credibility (MacKinnon 1987, 155; Sanders 1997). Those who advance good arguments may not be taken seriously in light of hearers' prejudices, themselves the result of significant social injustice (Fricker 2007).

[16] For a good summary of longstanding criticisms of the marketplace metaphor, see Brietzke (1997).

But perhaps the most vexed worry points not to the disfiguring effects of injustice, which in principle can be mitigated over time, but to the cognitive fallibility of humanity itself. Human beings suffer from a variety of cognitive biases that undermine the likelihood that they will form justified, true beliefs. For example, people are disposed to favor and notice evidence that corroborates what they already believe – so-called *confirmation bias* (Nickerson 1998, Oswald and Grosjean 2004). Thus they are resistant to evidence or arguments that contradict a belief they already hold; indeed, exposure to countervailing views may sometimes backfire by strengthening initial beliefs (for example, Nyhan and Reifler 2010). Further, people actively prefer only those *sources* of argument and evidence that are likely to support their settled viewpoints (Iyengar and Hahn 2009). That, in turn, fuels another pathology: *group polarization*.

Commented [AQ13]: AQ: Please confirm the change of 'Oswald et al. 2004' to 'Oswald and Grosjean 2004' in the text citation as per the reference list.

When people are exposed to a homogenous set of views that accords with their own, an echo chamber is created that has the effect of polarizing participants' beliefs. When people with a similar position deliberate that position, they become more extreme in their commitment to it (Moscovici 1980; Myers 1975; Sunstein 2009; Sunstein 2018, 73). Startlingly, this mechanism applies to both sophisticated and unsophisticated citizens; recent work even suggests that the more informed people are, the more likely they are to become entrenched in their existing views – deploying their advanced numeracy, for example, to construct more elaborate defenses of what they already think (Kahan et al. 2017).

A full discussion of the breadth of cognitive limitations is beyond the scope of this article.[17] However, the phenomena discussed here establish that enough humans

[17] See also Bambauer (2006, 673–694). For the definitive treatment, see Kahneman (2011).

reason poorly enough of the time to call into question the empirical claim that the truth is likely to prevail in an atmosphere of open debate. That explains why so many have rejected the notion of a marketplace of ideas as a general justification of freedom of speech. Yet these phenomena also seem to impugn the more specific idea that we can rely on counter-speech to defuse the risks of dangerous speech.

[LEVEL A HEADING] EFFECTIVE COUNTER-SPEECH

[LEVEL B HEADING] *Toward an Interdisciplinary Research Agenda*

One response to this objection is deflationary; we should simply concede counter-speech's ineffectiveness. Indeed, if mere ineffectiveness were the only possible downside, we might endorse the norm: 'Try your best, even if it might not succeed.' If we fail to be persuasive, so be it; at least we tried. Counter-speech, according to this revised argument, would no longer be seriously oriented to reducing danger. This proposal relieves agents of the requirement to reflect on fraught empirical questions about the prospects of success of various courses of action. Given the significant degree of uncertainty involved in judging whether a particular attempt at counter-speech will be effective, the proposal is undeniably attractive.

But we should reject it. If counter-speech is overwhelmingly likely to be ineffective, what is the justification for obliging agents to engage in it? Of course, it is *permissible* to engage in impotent counter-speech. Yet if such testimony is bound to change nothing, agents can reasonably ask why they should bear the costs of counter-speech if there is no anticipated moral pay-off.

The far more important reason to reject the proposal is that *counter-speech can backfire;* it can exacerbate the danger to the very people one is trying to protect. And when an instance of counter-speech is likely to make matters worse, I argue that

it *is morally wrong*. Doing so violates a negative natural duty, held by all agents, to refrain from conduct (including speech) that foreseeably endangers others without sufficient justification. This duty follows straightforwardly from the same morally significant interests that justify the positive samaritan duty of rescue. The same interests in life that justify the positive duty to rescue a drowning child also justify a negative duty not to actively try to drown that child, or assist others trying to do so. The same logic applies to counter-speech.

The scholarship tends to overlook this point. Brettschneider (2012), who persuasively defended the idea that the state should actively condemn hateful viewpoints through its own expressive capacities, does not address the possibility that such efforts may backfire.[18] Clayton and Stevens, while not framing their discussion in terms of counter-speech, do a bit better. When defending the importance of engaging with those who adhere to unreasonable religious doctrines, they admirably recognize that both liberal philosophers and politicians are bound to alienate such citizens through liberal proselytism (2014, 81). Accordingly, they propose that unreasonable adherents can be better engaged by their more liberal co-religionists, who can try to persuade them to adopt an alternative interpretation of their shared faith. In a brief footnote, they concede implicitly that this may be unsuccessful if the liberal co-religionists are 'viewed as heretics' (82n) – a wise concession, given that members of rival sects may be the *worst*-placed people to persuade each other of anything (for example, a liberal Shi'ite engaging with a fundamentalist Sunni, or a reform Jew with an orthodox Jew). But this concession must be more than just a passing observation. It is crucial, for in cases when counter-speech is likely to be

Commented [AQ14]: AQ: Please confirm the change of year from 'Clayton and Stevens 2013' to 'Clayton and Stevens 2014' in the text citation as per the reference list.

[18] For a similar point about the 'overly optimistic' character of Brettschneider's proposal, see Schauer (2014, 501).

counterproductive, exacerbating rather than mitigating danger, *citizens act wrongly by engaging in it*, however noble their intentions.[19]

Thus the samaritan character of my argument should dispose us to care deeply about whether the counter-speech will be effective, merely ineffective or downright counterproductive. After all, the point is to rescue people from harm. The response to the empirical concerns raised in the last section, accordingly, is not to dismiss them, but to incorporate them into the specification of the argument. If counter-speech duties include an obligation to speak effectively, their content is necessarily informed by an empirical specification of what sorts of communicative strategies are effective and which are not. Only by attending to such empirical research can we avoid the temptation to rely on armchair conjectures about how to persuade others.

I close by tracing some important lessons of recent empirical scholarship for the content of counter-speech duties, of which there have been few. One upshot of my argument is that we have a moral obligation to push for greater empirical work on counter-speech that is sensitive to the kinds of concerns flagged up by a normative theory of the kind I have developed here. Still, there is much to learn from prior empirical studies.

A review of the empirical literature shows that we must not assume that what is considered effective counter-speech in an academic seminar also works in a non-academic setting.[20] For instance, consider one intuitively promising form of counter-speech: the correction of mistaken empirical beliefs on the part of one's interlocutors. Because false empirical claims often form crucial elements of arguments for

[19] Cf. Badano and Nuti (2017), who are commendably aware of the effectiveness issue.

[20] See the related discussion in Badano and Nuti (2019), who fault those who 'over-intellectualize' the task of de-radicalizing terrorists.

dangerous normative proposals (Leader and Benesch 2016), this may seem an especially suitable mode of engagement. Yet this commonsensical armchair suggestion ignores a crucial finding in behavioral political science, which is that *efforts to correct often backfire*. In their widely discussed work on this topic, Nyhan and Reifler (2010) argue that those in the grip of a false belief often become *more* committed to that belief after being confronted with evidence that it is false.[21]

This has an obvious implication for counter-speech that attempts to defuse the risk of dangerous expression. For example, suppose a dangerous speaker stirs up hatred by persuading audience members that a particular racial group is inherently violent, citing false statistics about their propensity to commit crime, or false statistics about their success in evading punishment (thereby motivating the desire for vigilante retaliation) – all as part of a calculated attempt to foment hatred. In such circumstances, honestly correcting the false statistics may backfire. Such counter-speech would therefore be morally impermissible.

The lesson from this empirical finding is not that one should never correct dangerous, mistaken beliefs.[22] One crucial pay-off of carefully attending to the empirical scholarship in this area is that, just as it can illuminate where counter-

[21] The backfire effect has also been questioned, for example by Wood and Porter (2019). But this simply reinforces my point that we need greater empirical research in order to fully understand the content of counter-speech duties. If Wood and Porter are right and counter-speech backfires less often than Nyhan and Reifler think, then all the better for counter-speech. Notice, however, that the scholarly debate has largely focused on correcting interlocutors' *factual* mistakes, not their *moral* mistakes.

[22] In circumstances in which some audience members stand to be persuaded by counter-speech whereas for others it may backfire, a speaker should try to judge what course of action would result in less danger overall.

speech goes wrong, it can also show where it goes right. In their attempt to model the reasoning errors that generate the backfire effect, Edward Glaesar and Cass Sunstein postulate that corrections are far *less* likely to backfire when the speaker is what they term 'a surprising validator'. Their proposal is that corrections 'need to come from sources that are seen as credible to the relevant audience and not as likely to be lying' (2014, 67). If correct, this gives the morally conscientious citizen a criterion for deciding how to discharge her imperfect duty to engage in counter-speech: she should focus on audiences that are likely to perceive her as credible with respect to the topic under discussion.

In line with this view, a person may be perceived as credible when the audience is surprised that she holds the view in question – for example, because she agrees with them on many other issues, or because the view in question is against her self-interest. For example, Glaesar and Sunstein give the example of how a pro-environmental message is more persuasive when delivered by someone who is pro-business (see also Eagly, Wood and Chaiken 1978), or a gun control advocate who previously supported the National Rifle Association (Glaeser and Sunstein 2014, 91). In the context of dangerous speech (in the sense used here), we might think of a devout Muslim citizen who relentlessly criticizes the West and holds seriously conservative beliefs, but nevertheless argues that terrorism by groups such as Al-Qaeda and Islamic State are impermissible. Such a person might be particularly effective in a position of leadership, in virtue of their reach – think of the conservative but anti-terrorist American cleric Yasir Qadhi (Elliott 2011).[23]

[23] See Clayton and Stevens (2014), 80ff, for related discussion.

Consider now a second intuitively promising form of counter-speech. A common assumption is that the right response to those in the grip of bad ideas is to supply them with rational arguments that rebut those ideas. Counter-*speech*, according to this view, best proceeds through counter-*argument*. However, while this may be the right *ideal* for academic deliberation, the empirical scholarship suggests that merely supplying counter-arguments is unlikely to succeed. Jonathan Haidt argues that human beings tend to arrive at their moral judgments intuitively; only later, when tasked with trying to defend their views to others, do they muster rational arguments to defend those judgments. In this way, 'moral reasoning is rarely the direct cause of moral judgment' (Haidt 2001, 814, and 2012). Instead, 'moral intuitions (including moral emotions) come first' (2001, 814.).

This psychological account has significant normative implications for how citizens should undertake the task of counter-speech. If Haidt's social-intuitionist account of moral judgment is correct, 'reasoned persuasion works not be providing logically compelling arguments, but by triggering new affectively valenced intuitions in the listener' (2001, 819; see also Edwards and von Hippel (1995), and Shavitt (1990)). One implication of this finding is that a speaker ought to persuade by trying to get her listener to experience a new intuitive or emotional response.[24] Consider the debate over same-sex marriage. It is no surprise for social intuitionists that receptivity to same-sex marriage increased as more people came to discover that their own friends and family members were gay (Pew 2013). From this perspective, it is

[24] This lesson has long been understood by political theorists; see Krause (2008, 125), who argues that 'our minds are changed when our hearts are engaged'. And see the related work on rhetoric, e.g., Chambers (2009), Garsten (2009), Garsten (2011). This has implications for the use of emotion in counter-speech itself; for a compelling defence of anger in counter-speech, see Lepoutre (2018).

unlikely that deductive syllogisms were what changed people's attitudes. Rather, the change was plausibly caused, in part, by the generation of new 'affectively valenced intuitions' (Haidt 2001, 819).

The social-intuitionist model delivers an important lesson for specifying the content of counter-speech duties: namely, that we should adopt an expansive understanding of what 'counter-speech' might involve. Just as the US Supreme Court views 'expression' as a significant category, including books, movies, poetry, comedy and more, counter-speech can come in a variety of forms. The aim of a particular intervention into another's life may simply be to supply her with an experience that loosens the grip of some dangerous prejudice she holds, or better inoculate her against one to which she would otherwise be especially susceptible.[25] That may involve introducing her to a new person, recommending a particular film or simply telling her a provocative story that powerfully conveys an alternative narrative.

The purpose of these encounters need not even be considered indirect attempts to alter moral convictions. The aim may simply be to help the person see that their own good involves freeing their life of certain hateful attitudes. One of the more important lessons of Rawls's political theory is that citizens will only reliably be moved to do what justice requires if they view it as congruent with their own good (1999, 436). There is no reason, then, why counter-speech should not take the form of prudential appeals to what a good life involves.[26] Indeed, speech of the sort that Clayton and Stevens mention earlier – in which members of reasonable religious sects

[25] The importance of experiences is also stressed in Sunstein (2018, 40).

[26] A reviewer helpfully asked whether speech that reminds interlocutors that their expression is illegal – e.g., in jurisdictions that ban the speech – qualifies as a legitimate form of counter-speech. The answer is that it simply depends on whether speech of this sort is effective.

attempt to convince illiberal co-religionists to embrace more liberal versions of their shared conceptions of the good – fits into this category.

These lessons are limited. Part of the point of articulating the normative argument for counter-speech duties is that it issues a moral imperative for empirical research on this topic.[27] Outside of academia, there is much to learn from practitioners. For example, the United States Holocaust Museum's manual on how to counteract dangerous speech offers a wide array of important suggestions on how to engage in counter-speech (Brown 2016; see also Benesch et al. 2016). Part of the work to be done involves engaging with those involved in the practice of counter-extremism (as documented, for example, in Bjørgo and Horgan 2009), either with Islamist groups (for example, Rebasa et al. 2010; Stern 2010) or white supremacists (for example, Leydon and Cook 2008). Scholars – empirical and normative alike – need to assist with this task. Only a fully interdisciplinary research program – that is attuned to both normative and empirical concerns – can offer the kind of specification of counter-speech duties that, I have argued, we require.

[END TEXT]

Commented [AQ15]: AQ: Please confirm the change of '2016' to 'Brown 2016' in the text citation as per the reference list.

References

[27] More empirical work is not all that is needed. Philosophical work – namely, in the philosophy of language – also plays a crucial role in helping us determine what kind of counter-speech is likely to be effective or ineffective. For an instructive example, deploying the insights of Austinian speech-act theory, see Langton (2018).

Badano G and Nuti A (2017) Under pressure: political Liberalism, the rise of unreasonableness, and the complexity of containment. *Journal of Political Philosophy* **26**(2), 145–168.

Badano G and Nuti A (2019) The limits of conjecture: political liberalism, counter-radicalisation and unreasonable religious views. *Ethnicities*. doi:10.1177/1468796819866356.

Baker CE (1978) Scope of the first amendment freedom of speech. *UCLA Law Review* **25**, 964–1040.

Bambauer DE (2006) Shopping badly: cognitive biases, communications, and the fallacy of the marketplace of ideas. *University of Colorado Law Review* **77**, 649–710.

Bay C (1975) Access to political knowledge as a human right. *Human Context* **7**, 388–398.

Benesch S (2012) Dangerous speech: A proposal to prevent group violence. *World Policy Institute*, 12 January. Available from https://worldpolicy.org/wp-content/uploads/2016/01/Dangerous-Speech-Guidelines-Benesch-January-2012.pdf.

Benesch S et al. (2016) Considerations for successful counterspeech. Dangerous Speech Project. Available from https://dangerousspeech.org/considerations-for-successful-counterspeech.

Bjørgo T and Horgan J (eds.) (2009) *Leaving Terrorism Behind: Individual and Collective Disengagement.* New York: Routledge.

Blasi V (2004) Holmes and the marketplace of ideas. *The Supreme Court Review*, 1–46.

Commented [Techset16]: AQ: Please provide volume number in Blasi (2004).

Commented [AQ17]: AQ: Blasi (2004) and Singer (1972) are listed in the reference list but not cited in the text. Please cite in the text, else delete from the list.

Brettschneider C (2012) When the State Speaks, What Should It Say? Princeton, NJ: Princeton University Press.

Brietzke PH (1997) How and why the marketplace of ideas fails. *Valparaiso University Law Review* **31**, 951–969.

Brown RH (2016) Defusing hate: a strategic communication guide to counteract dangerous speech. United States Holocaust Memorial Museum.

Chambers S (2009) Rhetoric and the public sphere: has deliberative democracy abandoned mass democracy? *Political Theory* **37**, 323–350.

Clayton M and Stevens D (2014) When God commands disobedience: political liberalism and unreasonable religions. *Res Publica (Liverpool, England)* **20**, 65–84.

Edwards K and Von Hippel W (1995) Hearts and minds: the priority of affective versus cognitive factors in person perception. *Personality and Social Psychology Bulletin* **21**(10), 996–1011.

Elliott A (2011) Why Yasir Qadhi wants to talk about jihad. *The New York Times Magazine*, 17 March.

Fabre C (2007) Mandatory rescue killings. The Journal of Political Philosophy **15**(4), 363–384.

Fricker M (2007) *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Garsten B (2009) *Saving Persuasion: A Defense of Rhetoric and Judgment*. Cambridge, MA: Harvard University Press.

Garsten B (2011) The rhetoric revival in political theory. *Annual Review of Political Science* **14**, 164–174.

Gelber K (2012a) Speaking back: the likely fate of hate speech policy in the United States and Australia. In Maitra I and McGowan MK (eds), *Speech and Harm: Controversies Over Free Speech*. Oxford: Oxford University Press.

Gelber K (2012b) Reconceptualizing counterspeech in hate speech policy (with a focus on Australia). In Herz M and Molnar P (eds), *The Content and Context of Hate Speech: Rethinking Regulation and Responses.* Cambridge: Cambridge University Press.

Glaesar E and Sunstein CR (2014) Does more speech correct falsehoods? The Journal of Legal Studies **43**(1), 65–93.

Gordon J (1997) John Stuart Mill and the marketplace of ideas. *Social Theory and Practice* **23**(2), 235–249.

Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review* **108**(4), 814–834.

Hay C (2011) The obligation to resist oppression. *Journal of Social Philosophy* **42**(1), 21–45.

Heinze E (2016) *Hate Speech and Democratic Citizenship*. Oxford: Oxford University Press.

Hope S (2014) Kantian imperfect duties and modern debates over human rights. Journal of Political Philosophy **22**(4), 396–415.

Howard J (2019) Free speech and hate speech. *Annual Review of Political Science* **22**, 93–109.

Iyengar S and Hahn KS (2009) Red media, blue media: evidence of ideological selectivity in media use. *Journal of Communication* **59**, 19–39.

Commented [Techset18]: AQ: Please provide page range in Gelber (2012a), Gelber (2012b), 34, Maitra and McGowan (2012), McGowan (2018), Mill, Moscovici (1980) and Wenar (2011).

Kahan D et al. (2017) Motivated numeracy and enlightened self-government. *Behavioral Public Policy* **1**(1), 54–86.

Kahneman D (2011) *Thinking Fast and Slow*. London: Penguin.

Krause S (2008) *Civil Passions: Moral Sentiment and Democratic Deliberation*. Princeton, NJ: Princeton University Press.

Langton R (2018) Blocking as counter-speech. In Fogal D, Harris DW and Moss M (eds), *New Work on Speech Acts*. Oxford: Oxford University Press.

Leader MJ and Benesch S (2016) Dangerous speech and dangerous ideology: an integrated model for monitoring and prevention. *Genocide Studies and Prevention: An International Journal* **9**(3), 70–95.

Lepoutre M (2017) Hate speech in public discourse: a pessimistic defense of counter-speech. *Social Theory and Practice* **43**, 851–885.

Lepoutre M (2018) Rage inside the machine: defending the place of anger in democratic speech. *Politics, Philosophy & Economics* **17**, 398–426.

Maitra I and McGowan MK (2012) Introduction. In Maitra I and McGowan MK (eds), *Speech and Harm: Controversies Over Free Speech*. Oxford: Oxford University Press.

McGowan MK (2018) Responding to harmful speech. In Johnson CR (ed.), *Voicing Dissent: The Ethics and Epistemology of Making Disagreement Public*. Abingdon: Routledge.

McMahan J (2005) The basis of moral liability to defensive killing. *Philosophical Issues* **15**(1), 386–405.

Mill JS (1991[1859]) On liberty. In Gray J (ed.), *On Liberty and Other Essays*. New York: Oxford University Press.

Moscovici S (1980) Toward a theory of conversion behavior. In Berkowitz L (ed.), *Advances in Experimental Social Psychology*, vol. 13. New York: Academic Press.

Nickerson R (1998) Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology* **2**(2), 175–220.

Nyhan B and Reifler J (2010) When corrections fail: the persistence of political misperceptions. *Political Behavior* **32**(2), 303–330.

Oswald M and Grosjean S (2004) Confirmation bias. In Pohl and Rüdiger (eds), *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. Hove, UK: Psychology Press, pp. 79–96.

Parekh B (2006) Hate speech: is there a case for banning? *Public Policy Research* **12**, 213–223.

Pew Research Center (2013) Changing attitudes on same sex marriage, gay friends and family. Available from http://www.people-press.org/2013/06/06/changing-attitudes-on-same-sex-marriage-gay-friends-and-family.

Quong J (2012) Liability to defensive harm. *Philosophy & Public Affairs* **40**(1), 45–77.

Rawls J (1999[1971]) A Theory of Justice. Cambridge, MA: Harvard University Press.

Rawls J (2005) *Political Liberalism*. New York: Columbia University Press.

Schauer F (1992) Uncoupling free speech. *Columbia Law Review* 92(6), 1321–1357.

Schauer F (2014) Review: '*When the State Speaks, What Should it Say?'* by Corey Brettschneider. *Political Theory* **42**(4), 498–513.

Commented [AQ19]: AQ: Please provide initials for editor names 'Pohl and Rüdiger' in Oswald and Grosjean (2004).

Schwartzman M (2012) The ethics of reasoning from conjecture. *Journal of Moral Philosophy* **9**, 521–544.

Shavitt S (1990) The role of attitude objects in attitude formation. *Journal of Experimental Social Psychology* **26**(2), 124–148.

Shelby T (2018) *Dark Ghettos: Injustice, Dissent, and Reform*. Cambridge, MA: Harvard University Press.

Shiffrin S (2014) *Speech Matters*. Princeton, NJ: Princeton University Press.

Silvermint D (2013) Resistance and well-being. *Journal of Political Philosophy* **21**(4), 405–425.

Singer P (1972) Famine, affluence, and morality. *Philosophy & Public Affairs* **1**(1), 229–243.

Stern J (2010) Mind over martyr: how to deradicalize Islamist extremists. *Foreign Affairs* **89**(1), 95–108.

Stone G (2004) *Perilous Times*. New York: W.W. Norton & Company.

Sunstein C (2009) *Going to Extremes: How Like Minds Unite and Divide*. New York: Oxford University Press.

Sunstein C (2018) *#Republic*. Princeton, NJ: Princeton University Press.

Tadros V (2016) Permissibility in a world of wrongdoing. *Philosophy & Public Affairs* **44**(2), 101–132.

Vasanthakumar A (2018) Epistemic privilege and victims' duties to resist their oppression. *Journal of Applied Philosophy* **35**(3), 465–480.

Wellman CH (1996) Liberalism, samaritanism, and political legitimacy. *Philosophy & Public Affairs* **25**(3), 211–237.

Wellman CH (2013) *Liberal Rights and Responsibilities: Essays on Citizenship and Sovereignty*. New York: Oxford University Press.

Wenar L (2011) Poverty is no pond. In Illingworth P, Pogge T and Wenar L (eds), *Giving Well: The Ethics of Philanthropy*. Oxford: Oxford University Press.

Wood T and Porter E (2019) The elusive backfire effect: mass attitudes' steadfast factual adherence. *Political Behavior* **41**, 136–163.