Deleted: ORR ARTICLE HANDOVER COVER SHEET¶
Article details
... [1]

Extreme Speech, Democratic Deliberation, and Social Media

Jeffrey W. Howard

Abstract:

Social media are now central sites of democratic discourse among citizens. But are some contributions to social media too extreme to be permitted? This entry considers the permissibility of suppressing extreme speech on social media, such as terrorist propaganda and racist hate speech. It begins by considering the argument that such restrictions on speech would wrong democratic citizens, violating their freedom of expression. It proceeds to investigate the moral responsibilities of social media companies to suppress extreme speech, and whether these ought to be enforced through the law. Finally, it explores an alternative mechanism for combatting extreme speech on social media—counter-speech—and evaluates its prospects.

Keywords:

extreme speech; freedom of expression; deliberative democracy; social media; content moderation; incitement; hate speech; counter-speech

Introduction

Deliberation among citizens is a touchstone of contemporary normative democratic theory. For better or worse, online networks are now the principal site of civic deliberation. What the coffee house was to the public sphere in the eighteenth century (Habermas 1962), social media surely is to the twenty-first. Peer-to-peer sharing platforms have achieved an amplification of ordinary citizens' voices in a manner unthinkable just a few years ago. In everyone's pocket is a device enabling countless encounters with one's fellow citizens, on anything and everything of public political concern. Citizens who, in an earlier era, would have had their views on some policy matter heard by just their neighbours can now find their speech re-tweeted to millions.

The fact that social networks constitute a central site of democratic discourse seems to militate against the legal regulation of speech on these platforms. A central commitment of deliberative democracy is precisely that it ought to be an open exchange of citizens' authentic convictions. This democratic argument looms large in the scholarly literature on freedom of speech, sitting prominently alongside a raft of other arguments that aim to protect citizens' right to express their viewpoints (and hear the viewpoints of others). From this position, it seems to follow that *indirectly* suppressing citizens' speech—by legally commanding social media companies to do so through their content moderation practices—cannot be morally justified.

Yet the very amplification of varied voices that social media make possible—and that fuels optimistic sentiments about the democratizing power of the internet—has a dark side. Hateful speakers, hostile to the values of free and equal citizenship that underpin liberal democracy, can weaponize these platforms to cause a wide variety of harms. Racist conspiracy theories have inspired mass shootings around the globe, just as online sermons advocating religious extremism have encouraged suicide attacks. Given the state's duty to protect its citizens from wrongful harm, cases like these motivate the argument for restricting extreme speech on social media—controversially, by requiring social media companies to purge such content from their platforms.

This chapter examines the state of the debate on the fraught question of whether extreme speech should be suppressed or otherwise legally combatted on social media, or whether doing so would be incompatible with fundamental principles of democracy and free speech. I begin by reviewing the argument that freedom of expression, properly understood, protects a wide range of extreme speech, such as terrorist incitement and hate speech. Central to this discussion is the aforementioned thesis that *because* social media constitute a crucial

venue of democratic discourse, it is all the more important that citizens be free to express their views, however noxious.

The next section turns to the moral and legal responsibilities of social media companies regarding extreme speech. Even if, as I believe, individual speakers have no right to propagate extremist content online (such that individual criminal or civil liability for such speech could, in principle, be justified), it does not follow that social media companies should be legally subjected to a duty to remove such content. A recurrent complaint is that because social media networks are mere platforms on which users post their own content, it would be perverse to hold these companies responsible as if they were publishers. After exploring the extremely young philosophical debate on this difficult question, I explore various philosophical challenges that arise in the course of specifying an adequate regulatory model.

The final section then explores what measures there are for combatting extreme speech online beyond the legal regulation of social media networks. One ubiquitous suggestion in the scholarly literature on freedom of expression—one with deep affinities with the ideal of democratic deliberation—is that the best way to combat extreme speech is not to ban it, but rather to argue back through *counter-speech*. This proposal raises the question of who, exactly, ought to argue back against extremist voices, how they ought to go about it, and why it is reasonable to demand of them that they do it. I explore both state-centric and citizen-centric responses to this question in the scholarly literature. And I discuss how to think about the policy choice between banning extremist content and permitting it so it can be challenged.

Too extreme for democratic discourse? Extreme speech and the right to freedom of expression

Before we can assess what moral duties social media companies might have to combat extreme speech, we need to assess its normative status. Is it the kind of speech that is

protected by the moral right to freedom of expression? If so, then while private companies may decide that it has no place on their platforms, they cannot be forced by law to suppress it. Thus, before we can turn to the issue of what social media companies can be forced to do, one needs to set the stage for that discussion by rehearsing a set of debates over the limits of free speech.

Few believe that freedom of speech is *absolute*; all accept that some speech—be it soliciting a hitman to kill one's nemesis, or intentionally libelling a private citizen to destroy his reputation and livelihood—does not fall under the protective ambit of the right to freedom of expression. Yet liberal democracies disagree about whether speech that exhibits contempt for the values of liberal democracy itself—so-called *extreme speech*—is protected by a properly constituted principle of free speech.[1]

What is extreme speech? The phrase is a common one in the scholarly literature (e.g. Weinstein and Hare 2009), naming speech that expresses hostility to basic commitments of liberal democracy. The most basic subcategory of extreme speech is speech that incites violence and other serious violations of fundamental rights—what I have elsewhere called *dangerous speech* (Howard 2019b; following Benesch 2012). In the UK, for example, there is legislation forbidding the encouragement of terrorism (even implicitly, through speech 'glorifying' past terrorist acts) (Barendt 2009; Choudhury 2009). Under this law, British citizens have been arrested for tweeting praise of the Islamic State of Iraq and Syria (ISIS) and Al-Qaeda (Press Association 2016). Contrast this with the United States, where the Supreme Court has held, since 1969, that speech advocating criminal violence and other lawbreaking is constitutionally protected as free speech—except in emergency cases, in which speech is *intended* and *likely* to lead *imminently* to illegal conduct (*Brandenburg v. Ohio* 395 U.S. 444, 1969). Because online content seldom incites violence *imminently*, as there is time for the audience to ponder whether to act on its exhortations, vast swaths of

Commented [FB1]: Typesetter: please hyperlink.

terrorist propaganda is held protected in the United States (see Tsesis 2017, who thinks this is partly mistaken, for discussion).

The same divergence between democracies is on display with respect to the overlapping subcategory of so-called *hate speech*, that is, expressions of hatred toward vulnerable groups.[2] I call it an *overlapping* category, since a principal reason to be concerned about hate speech is precisely that it can inspire violence (see Howard 2019a: 104, 2019b)—though the empirical assumption on which this hypothesis rests is controversial (Heinze 2016: 125ff). This is only one rationale for restricting hate speech, where harm is caused via an intervening agent. But hate speech can also cause harm directly (see Schauer 1993 for this distinction). For example, some hate speech takes the form of intimidating threats or harassment, constituting an especially objectionable form of such speech (Delgado 1982; Howard 2019a: 101–102). Further, hate speech directly communicated to vulnerable groups can also be objectionable because it undermines the assurance of dignity and equal standing (Waldron 2012). The internet contains both forms of hate speech, and has a distinctive capacity to propel and prolong their harms when contrasted with offline hate speech. (For discussion of the distinctive power of hate speech online, see Tsesis 2001; Delgado and Stefanic 2014; Cohen-Almagor 2015**;** Brown 2018).

Notwithstanding the harms it may cause, is extreme speech protected by the moral right to free speech?[3] The standard way to answer this question is to review the arguments that serve to justify the moral right to free speech, and then ask whether those arguments count in favour of including extreme speech within the right's ambit (Howard 2019a: 96). For the purposes of this entry, I will focus on arguments that relate to democracy, and in particular that appeal to the idea of *democratic legitimacy*. I suggest that these arguments do not supply decisive protection for extreme speech; we would not wrong democratic citizens by restricting extreme speech on social media.

According to deliberative democrats, the legitimacy of laws flows from the fact that those laws were conceived in a process of open debate among citizens (e. g. Cohen 1989; Habermas 1992). This debate has both non-instrumental and instrumental value (Gutmann and Thompson 2004). The non-instrumental value of the deliberation inheres in the value of respecting our fellow citizens' equal moral status. By permitting one another to speak—and listening to one another—on questions of public concern, we respect each other as possessing the capacity for judgement over complex questions of public concern. Further, we respect each other as agents who are entitled to a justification for the coercion that is exercised over us. The instrumental value of deliberation inheres in the way it improves the quality of policy outcomes—for example, by enabling decision-making to incorporate the diversity of citizens' perspectives (Landemore 2012).

An apparent implication of this view is that the *legitimacy* of democratic decisions—by which we might mean either the permissibility of their enforcement, or their morally binding, authoritative status—is attenuated as more citizens are prevented from expressing their convictions (however wrongheaded) in democratic discourse. This view has been defended, in subtly different forms, by a wide array of scholars (Brettschneider 2012; Heinze 2016; Heyman 2009, Meiklejohn 1948, 1960; Post 1991, 2009, 2011; Sunstein 1993; Weinstein 2009). On this view, contributions to public discourse cannot be supressed simply because of their hateful or extreme character. As Ronald Dworkin puts the point, 'The majority has no right to impose its will on someone who is forbidden to raise a voice in protest or argument or objection before the decision is taken' (2009: vii). If social media are together a central forum of democratic discourse, then, it follows that governmentally imposed, viewpoint-based restrictions on what can be said on these platforms are undemocratic.

What should we make of this family of arguments? Consider the non-instrumental variation of the argument first. One possible reply is to grant that limiting extreme speech diminishes the democratic character of a polity, but to insist that this loss can be justified. Perhaps we *pro tanto* wrong extreme speakers by suppressing their speech, or indeed wrong *all citizens*, given everybody's interest in maximal democratic legitimacy, but this, all things considered, can be justified, given the comparable importance of preventing serious harm. So, for example, when we stop terrorists from advocating terrorism by enacting statutes restricting such advocacy, we do something wrong, but it might be justified nonetheless.[4]

Another possible reply is that even if limiting extreme speech diminishes democracy, this need not commit any *pro tanto* wrong at all, as not all instances of democracy have non-instrumental value (Beerbohm 2012: 36). So, for example, Jonathan Quong notes that when we are determining whether a moral right protects a particular action, we must 'ask whether the particular act that is alleged to be protected by a right is consistent with the overall moral ideal which the system of rights is meant to uphold' (2010: 308; see also Waldron 1989: 518). Because extreme speech expresses hostility towards the system of rights, on this view, it is not protected. A similar strategy, which I have defended, asks whether the moral right in question—in this case, democratic citizens' rights to express extreme speech—is compatible with moral duties that democratic citizens have (Howard 2019b: 232). So, for example, if democratic citizens have duties *not* to advocate for the violation of other citizens' fundamental rights, then these duties constrain what moral rights they have.

Even if we grant that *speakers* have no right to express extreme speech, it may be that efforts to suppress such speech wrong its prospective *listeners*. According to one influential theory, 'a legitimate government is one whose authority citizens can recognize while still regarding themselves as equal, autonomous, rational agents' (Scanlon 1972: 214). This view does not mean that no speech can ever be restricted; but it does rule out certain *justifications*

for restricting speech. As T. M. Scanlon argues, 'those justifications are illegitimate which appeal to the fact that it would be a bad thing if the view communicated by certain acts of expression were to become generally believed' (1972: 209; see also Dworkin 1996: 200). While seldom explicitly connected to the idea of deliberative democracy, this argument articulates the deliberative democrat's central concern about legitimacy: that if the state has rigged what ideas were allowed to be aired in a deliberation prior to a law's passage, the legitimacy of that law is called into serious question.

Yet this argument, too, faces important objections (so much so that Scanlon largely rejected it; 1979: 532ff). The most significant objection is that autonomous citizens *also* have an interest in avoiding the wide variety of harms that extreme speech can cause (Amdur 1980: 299; Brison 1998: 329). It is not clear why the interest in being respected as an autonomous thinker ought to have priority over this other weighty interest (for further discussion, see Howard 2019a: 97, 2019b: 236).

There are other reasons for why listeners may value exposure to extreme speech—not because of any deontological constraint on state power, but because of putative benefits that might flow from such exposure. This brings me to the *instrumental* version of the democratic legitimacy argument. The central point here is that by enabling a wide variety of voices' perspectives to enter the democratic discourse, it is thereby epistemically enriched, leading to better policy outcomes. This militates in favour of a truly capacious public discourse, but does it require an unlimited one? One reply here is to deny that extreme speech genuinely contributes to the epistemic value of public discourse; there is little to learn from engagement with neo-Nazis and white supremacists (*pace* Mill 1859/1978, who insisted that there was much to gain from our 'collision with error').[5] This point seems especially apt on social media, where a preponderance of those exposed to extremist views are those *already sympathetic to them*, and so inclined to visit the relevant websites, chat rooms, and pages (see

Sunstein 2017). But even if there *is* considerable epistemic value by permitting extreme views to be aired, this value is not infinite. Surely, we should accept some kind of trade-off between whatever epistemic value is achieved by permitting extreme speech and the obvious value of protecting citizens from the various harms such speech can inspire (Howard 2019a: 244).

Some deliberative democrats are likely to offer the following rejoinder: we cannot identify *ex ante* which views are true and which truths are false; a central point of deliberation is to precisely separate the wheat from the chaff. But this is why the most plausible characterization of extreme speech will pick out only that content whose falsehood is *beyond reasonable dispute*—for example, white supremacism. So, for example, when we're concerned about speech that endangers others by advocating the violation of their moral rights, it is proper to focus only on violations of rights that are properly incontrovertible (Howard 2019b: 215), that is, that no one could reasonably deny counts as a genuine violation of a genuine moral right. This means that citizens would retain broad scope to advocate views within the ambit of reasonable disagreement, but not to harm citizens by inciting rights violations outside of that ambit.[6]

I have inspected some of the most prevalent arguments in the free speech literature that connect to the political autonomy of democratic citizens. Importantly, I have not discussed all of them; for example, I have not discussed the powerful theory developed by Seana Shiffrin, who argues that freedom of expression traces to our fundamental interests as thinkers (2014; for discussion, see Scanlon 2011).[7] Nor have I dealt with the fact that while deliberative democracy is generally an account of the proper nature of public discourse within a nation state, social media is inherently global—a fact that raises a host of complications. Still, I have established two crucial points: that the scholarly literature is far from settled on the question of whether extremist speech must be permitted into democratic

discourse; and that there are several powerful reasons on offer to think that it should not, such that democratic citizens would not be wronged by such content's removal from social media. Whether social media platforms are morally obligated to remove such content—and whether such obligations should be enforced through law—are further questions, depending on further moral considerations, to which I now turn.

Platforming hate: on the duties of social media companies

Should social media platforms be required to suppress extreme content? It will not suffice simply to point out that such platforms are managed by private corporations that accordingly have a right to do whatever they wish. The idea that private corporations are appropriately saddled with various legal duties not to contribute to unjustified harm to the broader public is nothing new; just consider the raft of regulations corporations face with respect to environmental protection. Nor is it anything new to suppose that corporations should be obligated to look after the well-being of those who are using its products. Corporations—whether they are conceived as bona fide group agents or simply fictitious agents—are duty-bound to refrain from perpetrating wrongful harms, just like individual agents.

In the case of harmful speech on social media, controversy arises because corporations do not *directly* cause harm through their services; rather, they *enable others* to cause harms by supplying them with a platform through which to cause them. Especially significant is the fact that technology companies do not (typically) *intend* that users deploy their platforms to incite terrorism or racist hatred. One way to defend the legal immunity of social platforms for users' illegal posts, then, is to argue that they are simply 'like a billboard: anybody can sign or display anything on it' (Koltay 2019: 157). They are, in this way, privately owned versions of Speakers' Corner in Hyde Park, where anyone can show up to say whatever they want. If those who show up engage speech that is illegal, then *they* may be prosecuted or sued for saying it, but the owner of the platform is not to be held responsible.

In stark opposition this 'no liability' view is the thesis that social media networks are akin to either traditional media companies or book publishers, such that they are jointly responsible with users for illegal content. So, just as a publishing house can be sued alongside an author for a book's defamatory content, perhaps social media platforms could be liable alongside users for the illegal extreme speech they post.[8]

Both models are unattractive. The first 'no liability' model is defective because it relies upon mistaken views about a widely discussed philosophical idea known as *intervening agency*. A standard view is that if I act in a certain way, and this leads to harm only because of the intervening decisions of a responsible agent, then this fact (usually) immunizes me from moral responsibility. Yet this seems to me to be false; just consider cases in which someone sells an automatic weapon to someone who foreseeably plans to use it to violate others' rights. When I foreseeably cause harm to innocents through the conduct of others, I am accountable for these decisions and can be blamed for them (Tadros 2016). I have argued elsewhere that this moral truth applies to speech, just as it applies to all conduct (Howard 2019b: 216–217).

It might be replied that so long as the speaker does not *intend* for any intervening agent to engage in wrongful harm, he or she is off the hook. But this cannot be right; a lack of intention is seldom sufficient to immunize a party from moral responsibility (as our intuitions about selling weapons suggests). Leading philosophical work on complicity (understood as causally contributing to the wrongs of others) holds that, to be (*pro tanto*) wrongfully complicit in the wrong of another, one need not *intend* to aid the primary wrongdoer in his or her wrongful project. Rather, it is enough that one knew, or ought to have known, that one was causally contributing to its realization (Lepora and Goodin 2013: 83). (Given the ubiquity of journalistic reporting on the problems of extreme speech on social media, and platforms' efforts to limit extreme content on their networks *voluntarily*, it would appear that

this knowledge condition is satisfied.) Such a view offers a rationale for the current 'notice-and-takedown' approach currently prevalent in the European Union, whereby platforms are required to remove illegal content of which they are notified.[9] This is not necessarily the right regulatory model (more on that below), but it reflects the insight that it is perfectly possible to aid and abet the wrongs of another unintentionally.[10]

But the 'publisher' model is defective, too. Simply because social media companies might be complicit in the wrongs perpetrated by users (e.g., by providing a platform for the incitement of violence), it does not follow that social media companies are *co-principals* in these wrongs. Social media companies are not publishers in the traditional sense; Facebook plainly does not publish the content placed on it by its several billion users in anything like the way the *New York Times* publishes its editorials—authoring and standing by their content—or the way in which Penguin Random House publishes its books—vouching for their merit, and while not necessarily endorsing their content, suggesting that their content is worth one's time and money. Were regulations to treat social media firms like publishers, they would need to engage in extraordinary levels of so-called 'upload filtering'—screening all content for all potential legal issues before it ever hit the internet. This would radically alter the nature of social media, potentially for the worse.

We should likely opt, then, for a middle ground in how we are to conceive of social media networks—not neutral platforms, nor proper publishers. But what? Even if social media companies are not publishers, there is a sense in which they are nevertheless a new kind of *editor*. As Tim Berners-Lee, the inventor of the World Wide Web, put it, commenting on what is presently the world's largest social media network: 'Facebook makes billions of editorial decisions every day . . . The fact that these decisions are being made by algorithms rather than human editors doesn't make Facebook any less responsible . . .' (Lee 2016). To be sure, as Koltay notes, a platform is not a *fully fledged* editor; 'it does not initiate or

commission the production of content'. Yet 'it is an editor in the sense that it makes decisions concerning pieces of content and filters, removes content or keeps it available. It also controls all communications through the platform. All in all, it is clearly not neutral toward content' (Koltay 2019: 189). It is plausible to think of social media networks neither as platforms, nor as publishers, but rather as *curators* (cf. Herman 2016).

This leads to the question: what are the moral constraints on curating users' content, and how should they be enforced? Consider the natural habitat of this term: an art gallery. Imagine a peculiar kind of enormous 'open' art gallery, where members of the public are free to display their art, in their millions. Based on its knowledge of visitors' preferences, it directs them to the sections of the gallery that are likely to command their attention the most, giving grander spaces to those artworks that attract the greatest interest. Now suppose that artwork inciting hatred (e.g. racist propaganda art) is illegal. It seems plausible that the gallery should make a reasonable effort to limit the display of this art—making a reasonable effort to identify it (e.g. investigating complaints) and removing it in an expeditious manner upon discovering it. The duty to do so would simply be the duty not to be complicit in the harms such hateful content inspires, by providing a platform to do it.

If this *curator model* is right, everything hangs on what it is reasonable to expect curators to do. What, exactly, is it reasonable to expect social media companies to do to combat extreme speech? At the time of publication, there is a flurry of political debate on exactly this topic. I am writing this entry shortly after Germany has enacted a law that ramps up the 'notice-and-takedown' mechanism with respect to hate speech (*Netzwerkdurchsetzungsgesetz*, or NetzDG), which requires a robust complaints mechanism requiring companies to remove 'manifestly unlawful' content within twenty-four hours or risk fines up €50 million. The UK is presently contemplating its own legislation to saddle social media companies with a 'duty of care', whereby companies must 'take reasonable

steps to keep users safe' (UK Government 2019)—to be specified by a government regulatory body, such as Ofcom, the telecommunications regulator. So, just as amusement parks must pursue various precautionary measures to keep visitors safe, so, too, should online networks (Woods and Perrin 2019). This legislation has provoked fierce debate, and at the time of writing it is difficult to predict what the final law will involve (for discussion, see Woods 2019; Nash 2019; Tambini 2019; see Theil 2019 for a comparison of UK and German approaches). These real-world developments will no doubt provoke further philosophical reflection on what, exactly, a duty of care is and what it demands[11] (e.g. see Herstein 2010, for reflection on that general issue).

What is striking is that, in academia, philosophers have scarcely weighed in; those who are making the greatest contribution at present are (theoretically minded) lawyers, especially scholars of media law (e.g. Rowbottom 2018b: 341ff; Koltay 2019; PoKempner 2019). But there are central philosophical questions here that require greater attention from the philosophical community. A central topic of burgeoning debate, I suspect, will concern the role of artificial intelligence (AI) in content moderation. If using AI is the only efficient way to take down large quantities of extreme content, we face the challenge that highly imperfect algorithmic moderation processes are likely to be either *overinclusive*—taking down more content than we would want—or *underinclusive*—taking down less, or indeed both in different respects (see Douek 2021 for related discussion). This, then, raises a wide set of important questions about what collateral costs of speech restrictions (e.g. reduction in the amount of sarcasm on the internet) it would be reasonable to expect citizens to bear. While I seek to address such questions in my future work, there is little philosophical attention to them at present.

Another important question concerns whether social media networks *themselves* enjoy expressive rights that immunize them from interference. The fact that social media networks

Deleted: ors

are not neutral—that they inevitably take responsibility for the algorithms that determine what content is promoted or quieted—has a particular interesting implication. To the extent that social media firms *do* share some of the properties of publishers or editors, they may be entitled to the protections of freedom of speech and freedom of the press. As Koltay notes, 'In a sense, its news feed is Facebook's "opinion" on what its users might be most interested in and how the platform's business interests could be best served in that context. If a platform has an opinion, it is afforded protection under the constitutional rules.' Of course, that does not mean the platforms are allowed to enable any speech they like; it still means that their speech 'may also be subject to restriction, pursuant to applicable legal principles' (2019: 159). But an intriguing upshot of this view is that *if* a certain category of speech is protected as free speech for users, it would be impermissible for the state to require social media companies to suppress it. Indeed, we should be very suspicious of requiring social media companies to suppress speech that individual users have a legal right to express.

In reply, it may be a mistake to view a social media company as a merely private entity entitled to express 'its' own views. This is partly because of doubts about the expressive rights of corporations. But more fundamentally, social networks have enormous power over the public discourse (Klonick 2018). The nexus of social media is clearly part of what Rawls called *the basic structure* of society, given that 'its effects are so profound and pervasive' (Rawls 1971/1999: 82) on the shape of democratic deliberation. Suppose a social network started to ban the expression of certain religious or political views on the grounds that it disfavoured the view. Given the role of these networks in curating the democratic discourse of contemporary societies, there is a powerful argument for thinking that *the same* free speech principles that bind governments should also regulate social media platforms (see Jackson 2014 for discussion).[12] Those inclined to view social media networks as a kind of communications utility, albeit privately owned as a legal matter, would be inclined to support

such a position (see Lentz 2011 for related discussion). In my view, there is nothing *morally* incompatible with viewing these networks as a public utility while also requiring them to remove extreme speech (even though this would raise constitutional issues in the United States).

The final philosophical puzzle that remains unsolved concerns the issue of *asymmetric enforcement of duties*. It seems clear that we are entering a world in which the *primary* way in which extreme speech is combatted is by removing it on social media, rather than prosecuting the extreme speakers themselves. This raises a worry: shouldn't the initial speakers be held accountable if anyone is? In reply, it must be noted that even if speakers *initiate* extreme content by posting it, it is social media platforms that enable its widespread dissemination. Further, there may be something morally desirable about a world in which social media companies limit the dissemination of extreme speech, but individual citizens are nevertheless free to express it. As Jacob Rowbottom notes, this is, in part, a matter of efficiency; given the huge amount of illegal content, 'it is easier to ask a gatekeeper to control the flow of such content than to bring a legal action against each individual publisher' (Rowbottom 2018a: 2). But a more principled worry is that prosecuting individual speakers for each and every extreme statement—however careless—will disproportionately interfere in public discourse and undermine conversation (Rowbottom 2012). What is more, insofar as we think there is *some* interest to engage in extreme speech (even if not weighty enough to justify a moral right, as I have argued in Howard 2019b), permitting such speech—but then requiring intermediaries to limit its dissemination—'may strike a balance between the free flow of conversation and any potential harm' (Rowbottom 2018b: 2). The upshot, then, would be that while individual speakers *do* have moral duties to refrain from posting extreme content, we would refrain from enforcing these duties directly.

Deleted: b

In closing, I should note that there are exceptions to the general trend of increasing social media regulation, which raise their own philosophical complications. In the United States, restricting extreme speech on social media will not be so straightforward. Even if Section 230 of the Communications Decency Act were altered in various ways to make platforms liable for illegal speech posted by users (e.g. libel), extreme speech is mostly legal in the United States. And it is highly unlikely that this fact will change any time soon; as mentioned, the prevailing interpretation of the First Amendment to the US Constitution protects extreme speech except in cases of imminent harm (though cf. Tsesis 2017).[13] Even so, many media networks are nevertheless taking it upon themselves to remove extreme content *voluntarily*. As Koltay puts it, 'This means the enforcement of a "pseudo legal system", with its own code, case law, sanctions . . . taking place in a privately owned virtual space' (2019: 3). We have reason to be concerned about the prospect of what republican political theorists term *domination* by these entities (Pettit 2012); if these social media firms are going to become the *real* arbiters of what people are permitted to say in the public sphere, a question arises as to whether they enjoy the right kind of *legitimacy* to wield this form of power. If public discourse is to be curtailed to prevent harm, perhaps it should be done by a legitimate democratic state, or by no one at all.

The ethics of online counter-speech

What else might be done to combat extreme speech on social media? A recurrent suggestion in the scholarly literature on free speech—indeed, for some, a rationale for free speech itself—appeals to the importance of *counter-speech*. As Justice Louis Brandeis of the US Supreme Court put it, reflecting on the best way to confront speech that is harmful or otherwise disagreeable: '[T]he remedy to be applied is more speech, not enforced silence' (*Whitney v. California* 274 U.S. 357 (1927)). In other words, rather than suppress extreme speech, we need to *argue back against it*. This strategy is an especially fitting one in the

context of a deliberative democracy, in which public deliberation (followed by voting) among citizens and their representatives is the default mechanism for dealing with disagreement. And while deliberative democrats tend to suggest that the main substance of their disagreement is *reasonable* disagreement about what justice requires, there is no reason why *unreasonable* disagreements (in which one side is manifestly mistaken) should not also be dealt with through the same strategy. This section explores the moral status of this strategy of response.

Why might counter-speech be preferred to the use of legal coercion? The traditional argument in defence of counter-speech is simply that it is the only option morally available. If the moral right to free speech protects extreme speech, as many philosophers contend, then counter-speech is the only recourse we have left to combat the harms such speech can generate. As discussed above, I am not convinced that this traditional argument succeeds. However, simply because extreme speech is unprotected by the moral right to free speech does not automatically mean that we should prefer the coercive use of state power to peaceful alternatives. A better reason to prefer counter-speech, I have proposed, appeals to the philosophical *principle of necessity*, familiar from the ethics of self-defence (Howard 2019b: 248ff). According to the necessity principle, one ought to avert a threat using the least amount of harm or force, *ceteris paribus*. So, if police can successfully deescalate a dangerous situation through *talking*, they should do that rather than deploying violence. Likewise, if one can attain an important social goal without deploying the coercive power of the state, then *ceteris paribus* we ought to prefer the non-coercive strategy.

Ceteris paribus is an important qualification here. If counter-speech is a significantly less effective strategy, or if it is morally unreasonable to demand that the relevant counter-speakers engage in the requisite counter-speech, then the use of law may turn out to be preferable, after all. An adequate defence of counter-speech thus must attend to the issues of

who ought to engage in counter-speech, *why* it is reasonable to demand that they undertake, and *how* they ought to undertake it in order to be both ethical and effective (Howard 2019c). I will discuss these issues in turn.

Start with the question of *who* should engage in counter-speech. One possibility is *the state.* Challenging the false dichotomy that the state must either ban extreme content or otherwise sit back and let it proliferate unchecked, Corey Brettschneider argues that the state ought to take on the central role of engaging in counter-speech against extreme views (2012). According to his account of 'democratic persuasion', Brettschneider contends that the state should endeavour to persuade citizens in the grip of extremist views 'to adopt the values of equal citizenship' (2012: 72). In the digital era, this kind of counter-speech could come in many different forms; we might imagine the state recording YouTube videos defending the values of the freedom and equality, or publicizing politicians' speeches doing so on its social media channels. It is tempting to see this as a form of propaganda, though when deployed in the service of just end, we might instead see it as a form of what Jason Stanley has called 'civil rhetoric' (2015).

What is the argument for insisting that the state ought to engage in counter-speech against extreme views? For Brettschneider, the argument appeals to the idea that *if* the state sat back and did nothing in the face of extremist speech, its silence would constitute a form of *complicity* (2012: 71). We might also appeal to the state's obligation to reduce the likelihood of the wrongful harms such extreme speech can inspire—the very same obligation underpinning the case for banning such speech (Howard 2019b). Further, in the case of extreme speech that directly attacks the dignity of vulnerable citizens (of the sort that concerns Waldron 2012), we might think that state counter-speech can more effectively *block* such dignitarian harm by authoritatively affirming the dignity of the attacked citizens (Lepoutre 2017).

Commented [FB2]: Typesetter: please hyperlink.

Even if state speech is useful in upholding the dignity of citizens who are directly smeared by hate speech, and even if it plays some role in dissuading susceptible citizens from embracing hateful views, it has its limits. Most notably, the liberal state is unlikely to be successful at convincing those in the grip of anti-liberal ideologies to abandon their deeply held convictions (Howard 2019c).

That the state is unlikely to convince opponents naturally leads to the suggestion that *citizens* ought to take up the task of engaging in counter-speech (though, of course, the state could support them in this role in various ways; see Gelber 2012). There have been a variety of attempts in the scholarly literature along these lines. Focusing on the problem of religiously inspired terrorism, Clayton and Stevens argue that liberal adherents to a particular religion ought to engage with those in the grip of an intolerant version of that same religion, since they are those best positioned to persuade (2014: 75). Relatedly, Micah Schwartzman has defended the practice of what Rawls called 'reasoning from conjecture', whereby citizens reason *as if* they shared the argumentative starting points of their interlocutors (2012; cf. Badano and Nuti 2020). In other work, focusing on the problem of right-wing xenophobic populism, Badano and Nuti (2018) defend the claim that citizens have a 'duty of pressure' to try to dissuade their fellows from populist views. And I have argued that all citizens in any position to talk someone out of a dangerous view have powerful reason to do so (2019c).

What is the justification of requiring ordinary citizens to engage in counter-speech, given its difficulties? Some authors have appealed to the *natural duty of justice* (Clayton and Stevens 2014: 81), the moral requirement to support and help advance just institutions. In a similar spirit, others have appealed to the *liberal principle of legitimacy*, which requires citizens to deploy public reason in their engagements with one another on matters of law and policy (Badano and Nuti 2018: 148). I have offered what I take to be a more austere

argument, which simply appeals to the natural moral *duty to rescue* others from harm when one can do so at reasonable cost to oneself (Howard 2019c).

One reason to worry about saddling ordinary citizens with duties to engage in counter-speech is that it suggests that even the *victims* of extreme speech have duties to argue back against the speech that degrades and endangers them—which seems unfair (Maitra and McGowan 2012: 9). A possible reply is to argue that even victims of injustice have duties to resist their own oppression (Hay 2011). But a more plausible reply is to recognize that any duty's existence is sensitive to costs; if it is extremely demanding for victims of extreme speech to engage in counter-speech, then it cannot reasonably be required of them (Howard 2019c).

If citizens have moral duties to engage in counter-speech, what do these duties require of them in the digital era? The answer to this question is, to put it mildly, unclear. For example, consider extreme speech propagated on white supremacist websites, chat rooms, or threads. A principal danger of such speech is that it will inspire violence against non-whites. So how should we combat it? Should anti-racists infiltrate these chat rooms, subtly inserting seeds of doubt—or engaging in outright counterargument? The difficulty of answering such questions is compounded by the fact that it is unclear what kinds of counter-speech are actually effective at achieving their aim (see Lepoutre 2019 for relevant discussion). In cases in which the aim is to protect the dignity of vulnerable groups by standing up for them, thereby 'blocking' the hateful speech (Langton 2018; Lepoutre 2017), the aim is achieved just in case the communication is successful. But when the aim of the counter-speech is to change hearts and minds, to persuade susceptible listeners or hardened extremists to reject extremist views, it is simply an open empirical question what strategies are most effective. (For relevant empirical discussion on counter-speech generally, see Benesch et al. 2016 and

Brown 2016, and for particular attention to strategies for the online context, see Gagliardone et al. 2015).

Much of the important work left to be done is indeed philosophical. For example, even if *publicly shaming* illiberal citizens on the internet were an effective way of standing up for liberal values, there remain important questions about whether it is morally permissible (see Billingham and Parr 2020). But as with so many applied normative topics, much of the important work that is yet to be done is not strictly philosophical, but rather empirical. This is why it is all the more important that philosophers engage with social scientists, to learn from but also crucially to inform their research agendas. With respect to the issue of online counter-speech, it is vital that we secure an evidence base with which to adjudicate whether online counter-speech is or is not an effective remedy. This is vital precisely because, if counter-speech is not effective, or if it is simply too difficult to do it effectively given constraints on people's time and resources, this could justify a recourse to legal measures.[14]

I sincerely hope this conclusion is false, and that we can indeed combat the harms on social media—as Justice Brandeis hoped for the offline world—with 'more speech'. It is never ideal when a liberal society cracks down on speech, even justifiably, and there is always the risk that it will counter-productively play into extremists' hands (Howard 2019b: 245). As with so many thorny problems in the burgeoning field of digital ethics, we are staring down the precipice at an uncertain new world.

Acknowledgements

I am grateful to the Leverhulme Trust for research funding and to Carissa Véliz and an anonymous reviewer for helpful comments.

References

Alexander, Larry. (2005), *Is There a Right to Freedom of Expression*? (Cambridge: Cambridge University Press).

Amdur, Robert (1980), 'Scanlon on Freedom of Expression', *Philosophy & Public Affairs* 9, 287–300.

Badano, Gabriele, and Nuti, Alasia. (2018), 'Under Pressure: Political Liberalism, the Rise of Unreasonableness, and the Complexity of Containment', *Journal of Political Philosophy*, 26, 145–168.

Badano, Gabriele, and, Nuti Alasia. (2020), 'The Limits of Conjecture: Political Liberalism, Counter-Radicalisation and Unreasonable Religious Views', *Ethnicities* 20, 293–311.

Bambauer, Derek E. (2006), 'Shopping Badly: Cognitive Biases, Communications, and the Fallacy of the Marketplace of Ideas', *University of Colorado Law Review* 77, 649–710.

Barendt, Eric (2009), 'Incitement to, and Glorification of, Terrorism', in James Weinstein and Ivan Hare, eds, *Extreme Speech and Democracy* (Oxford: Oxford University Press), 445–462.

Beerbohm, Eric (2012), *In Our Name: The Ethics of Democracy* (Princeton, NJ: Princeton University Press).

Benesch, Susan (2012), 'Dangerous Speech: A Proposal to Prevent Group Violence', World Policy Institute, 12 January, https://worldpolicy.org/wp-content/uploads/2016/01/Dangerous-Speech-Guidelines-Benesch-January-2012.pdf., accessed 11 August 2021.

Benesch, Susan; Ruths, Derek; Dillon, Kelly P; Saleem, Haji Mohammad; and Wright, Lucas. (2016), 'Considerations for Successful Counterspeech', Dangerous Speech Project, https://dangerousspeech.org/considerations-for-successful-counterspeech, accessed 11 August 2021.

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

Deleted: Baker, C. Edwin (2009), 'Autonomy and Hate Speech', in J. Weinstein and I. Hare, eds, *Extreme Speech and Democracy* (Oxford: Oxford University Press), 139–157.¶

Brison, Susan (1998), 'The Autonomy Defense of Free Speech', *Ethics* 108, 312–339.,

Billingham, Paul, and Parr, T. (2020), 'Enforcing Social Norms: The Morality of Public Shaming', *European Journal of Philosophy*, doi: https://doi.org/10.1111/ejop.12543.

Brown, Alexander (2017a), 'What Is Hate Speech? Part 1: The Myth of Hate', *Law & Philosophy* 36, 419–468.

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

Brown, Alexander (2018), 'What Is So Special about Online (as Opposed to Offline) Hate Speech?', *Ethnicities* 18, 297–326.

Brown, Rachel (2016), *Defusing Hate: A Strategic Communication Guide to Counteract Dangerous Speech* (Washington, DC: US Holocaust Memorial Museum).

Choudhary, Tufyal (2009), 'The Terrorism Act 2006: Discouraging Terrorism', in James Weinstein and Ivan Hare, eds, *Extreme Speech and Democracy* (Oxford: Oxford University Press), 463–487.

Clayton, Matthew, and Stevens, David. (2014), 'When God Commands Disobedience: Political Liberalism and Unreasonable Religions', *Res Publica* 20, 65–84.

Cohen, Joshua (1989), 'Deliberation and Democratic Legitimacy', in Alan Hamlin and Philip Pettit, eds, *The Good Polity* (Oxford: Basil Blackwell), 17–34.

Cohen-Almagor, Raphael (2015), *Confronting the Internet's Dark Side: Moral and Social Responsibility on the Free Highway* (Cambridge: Cambridge University Press).

Delgado, Richard (1982), 'Words that Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling', *Harvard Civil Rights–Civil Liberties Law Review* 17, 133–181.

Delgado, Richard, and Stefanic, Jean (2014), 'Hate Speech in Cyberspace', *Wake Forest Law Review* 49, 319–343.

Deleted: Brison, Susan, and Gelber, Katherine, eds (2019), *Free Speech in the Digital Age* (Oxford: Oxford University Press).¶

Douek, Evelyn (2021), 'Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability', *Columbia Law Review* 121, 759–834.

Dworkin, Ronald (1996), *Freedom's Law: The Moral Reading of the American Constitution* (Oxford: Oxford University Press).

Dworkin, Ronald (2009), 'Forward', in James Weinstein and Ivan Hare, eds, *Extreme Speech and Democracy* (Oxford: Oxford University Press), 123–138.

Franks, Mary Ann (2019), '"Not Where Bodies Live": The Abstraction of Internet Expression', in Susan Brison and Katherine Gelber, eds, *Free Speech in the Digital Age* (Oxford: Oxford University Press), 137–149.

Gagliardone, Iginio; Gal, Danit; Alves, Thiago; and Martinez, Gabriela (2015), *Countering Online Hate Speech* (New York: UNESCO).

Gelber, Katherine (2012), 'Reconceptualizing Counterspeech in Hate Speech Policy (with a Focus on Australia)', in Michael Herz and Peter Molnar, eds, *The Content and Context of Hate Speech: Rethinking Regulation and Responses* (Cambridge: Cambridge University Press), 198–216.

Gordon, Jill (1997), 'John Stuart Mill and the "Marketplace of Ideas"', *Social Theory & Practice*, 23, 235–249.

Gutmann, Amy, and Thompson, Dennis (2004), *Why Deliberative Democracy?* (Princeton, NJ: Princeton University Press).

Habermas, Jürgen (1962), *The Structural Transformation of the Public Sphere* (Cambridge, MA: Polity Press).

Habermas, Jürgen (1992), *Between Facts and Norms* (Cambridge, MA: MIT Press).

Hay, Carol (2011), 'The obligation to Resist Oppression', *Journal of Social Philosophy* 42, 21–45.

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

Herrman, John (2016), 'Social Media Finds New Role as News and Entertainment Curator', *New York Times*, https://www.nytimes.com/2016/05/16/technology/social-media-finds-new-roles-as-news-and-entertainment-curators.html, accessed 11 August 2021.

Hern, Alex (2020), 'Twitter Hides Donal Trump Tweet for "Glorifying Violence"', *The Guardian*, https://www.theguardian.com/technology/2020/may/29/twitter-hides-donald-trump-tweet-glorifying-violence, accessed 11 August 2021.

Herstein, Ori (2010), 'Responsibility in Negligence: Why the Duty to Care Is Not a Duty "To Try"', *Canadian Journal of Law and Jurisprudence* 23, 403–428.

Heyman, Steven (2009), 'Hate Speech, Public Discourse, and the First Amendment', in James Weinstein and Ivan Hare, eds, *Extreme Speech and Democracy* (Oxford: Oxford University Press), 123–138.

Howard, Jeffrey W. (2019a), 'Free Speech and Hate Speech', *Annual Review of Political Science* 22, 93–109.

Howard Jeffrey W. (2019b), 'Dangerous Speech', *Philosophy & Public Affairs* 47, 208–254.

Howard, Jeffrey W. (2019c), 'Terror, Hate, and the Demands of Counter-Speech', *British Journal of Political Science*, doi: https://doi.org/10.1017/S000712341900053X.

Jackson, Benjamin F. (2014), 'Censorship and Freedom of Expression in the Age of Facebook', *New Mexico Law Review* 44, 121–167.

Kendrick, Leslie (2017), 'Free Speech as a Special Right', *Philosophy & Public Affairs* 45, 87–117.

Klonick, Kate (2018), 'The New Governors: The People, Rules, and Processes Governing Online Speech', *Harvard Law Review* 131, 1599–1670.

Koltay, András (2019), *New Media and Freedom of Expression* (Oxford: Hart Publishing).

Langton, Rae (2018), 'Blocking As Counter-Speech', in Daniel Fogal, Daniel W. Harris, and Matt Moss, eds, *New Work on Speech Acts* (New York: Oxford University Press), 144–164.

Landemore, Hélène (2012), *Democratic Reason* (Princeton, NJ: Princeton University Press).

Lee, Timothy B. (2016), 'Mark Zuckerberg Is in Denial about How Facebook is Harming Our Politics', *Vox*, https://www.vox.com/new-money/2016/11/6/13509854/facebook-politics-news-bad., accessed 11 August 2021.

Lentz, Roberta (2011), 'Regulation as Linguistic Engineering', in Robin Mansell and Marc Raboy, eds, *The Handbook of Global Media and Communication Policy* (Oxford: Blackwell), 432–448.

Lepora, Chiara and Goodin, Robert E. (2013), *On Complicity and Compromise* (Oxford: Oxford University Press).

Lepoutre, Maxime (2017), 'Hate Speech in Public Discourse: A Pessimistic Defense of Counter-Speech', *Social Theory and Practice* 43, 851–885.

Lepoutre, Maxime (2019), 'Can "More Speech" Counter Ignorance Speech?', *Journal of Ethics and Social Philosophy* 16, 155–191.

Maitra, Ishani and McGowan, Mary Kate (2012), 'Introduction', in Ishani Maitra and Mary Kate McGowan (eds.), *Speech and Harm: Controversies Over Free Speech* (Oxford: Oxford University Press), 1–23.

Meiklejohn, Alexander (1948), *Free Speech and Its Relation to Self-Government* (New York: Harper and Brothers).

Meiklejohn, Alexander (1960), *Political Freedom* (New York: Harper and Brothers).

Mill, John Stuart (1859/1978), *On Liberty*, ed. Elizabeth Rapaport (Indianapolis, IN: Hackett).

Nash, Victoria (2019), 'Revise and Resubmit? Reviewing the 2019 Online Harms White Paper', *Journal of Media Law* 11, 18–27.

Pettit, Philip (2012), *On the People's Terms* (Cambridge: Cambridge University Press).

PoKempner, Dinah (2019), 'Regulating Online Speech: Keeping Humans, and Human Rights, at the Core', in Susan Brison and Katherine Gelber, eds, *Free Speech in the Digital Age* (Oxford: Oxford University Press), 224–245.

Post, Robert (1991), 'Racist Speech, Democracy, and the First Amendment', *William Mary Law Review* 32, 267–327.

Post, Robert (2009), 'Hate Speech', in James Weinstein and Ivan Hare, eds, *Extreme Speech and Democracy* (Oxford: Oxford University Press), 123–138.

Post, Robert (2011), 'Participatory Democracy as a Theory of Free Speech: A Reply', *Virginia Law Review* 97, 617–632.

Press Association (2016), 'Security Guard Jailed for Five Years over Tweets Glorifying Isis', *The Guardian*, https://www.theguardian.com/uk-news/2016/apr/28/security-guard-mohammed-moshin-ameen-jailed-for-five-years-over-tweets-glorifying-isis., accessed 11 August 2021.

Quong, Jonathan (2010), *Liberalism without Perfection* (Oxford: Oxford University Press).

Rawls, John (1971/1999), *A Theory of Justice* (Cambridge, MA: Harvard University Press).

Reid, Andrew (2020), 'Does Regulating Hate Speech Undermine Democratic Legitimacy? A Cautious "No"', *Res Publica* 26, 181–199.

Rowbottom, Jacob (2012), 'To Rant, Vent and Converse', *Cambridge Law Journal* 71, 355–383.

Rowbottom, Jacob (2018a), 'Written Evidence on Internet Regulation to the House of Lords Communications Committee', http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/

communications-committee/the-internet-to-regulate-or-not-to-regulate/written/82636.pdf., accessed 11 August 2021.

Rowbottom, Jacob (2018b), *Media Law* (Oxford: Hart Publishing).

Scanlon, Thomas M. (1972), 'A Theory of Freedom of Expression', *Philosophy & Public Affairs* 1, 204–226.

Scanlon, T. M. (1979), 'Freedom of Expression and Categories of Expression', *University of Pittsburgh Law Review* 40, 519–550.

Scanlon, T. M. (2011), 'Comment on Shiffrin's Thinker-Based Approach to Freedom of Speech', *Constitutional Commentary* 27, 327–335.

Schauer, F. (1993), 'The Phenomenology of Speech and Harm', *Ethics* 103(4), 6350–6653.

Schauer, Frederocl (2019), 'Recipes, Plans, Instructions, and the Free Speech Implications of Words that Are Tools', in Susan Brison and Katherine Gelber, eds, *Free Speech in the Digital Age* (Oxford: Oxford University Press), 74–87.

Schwartzman, Micah (2012), 'The Ethics of Reasoning from Conjecture', *Journal of Moral Philosophy* 9(4), 521–544.

Scoccia, Danny (1996), 'Can Liberals Support a Ban on Violent Pornography?', *Ethics* 106, 776–799.

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

Stanley, Jason (2015), *How Propaganda Works* (Princeton, NJ: Princeton University Press).

Sunstein, Cass (1993), *Democracy and the Problem of Free Speech* (New York: Free Press).

Sunstein, Cass (2017), *#Republic: Divided Democracy in the Age of Social Media* (Princeton, NJ: Princeton University).

Tadros, Victor (2016), 'Permissibility in a World of Wrongdoing', *Philosophy & Public Affairs* 44, 101–132.

Commented [OUP-CE5]: AQ: The reference "Rowbottom, Jacob (2018a)" has not been cross-referred in the text. Please provide the cross-reference, or remove the reference from the reference list.

Commented [HJW6R5]: It is now referenced in the text on p. 22
(see where it says Rowbottom 2018a: 2)

Tambini, Damian (2019), 'The Differentiated Duty of Care: A Response to the Online Harms White Paper', *Journal of Media Law* 11, 28–40.

Theil, Stefan (2019), 'The Online Harms White Paper: Comparing the UK and German Approaches to Regulation', *Journal of Media Law* 11, 41–51.

Tsesis, Alexander (2001), 'Hate in Cyberspace: Regulating Hate Speech on the Internet', *San Diego Law Review* 38, 817.

Tsesis, Alexander (2017), 'Social Media Accountability for Terrorist Propaganda', *Fordham Law Review* 86, 605–631.

UK Government (2019), 'Online Harms White Paper', https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper, accessed 11 August 2021.

Ullman, Stefanie, and Tomalin, Marcus (2020), 'Quarantining Online Hate Speech: Technical and Ethical Perspectives', *Ethics and Information Technology* 22, 69–80.

Waldron, Jeremy (1989), 'Rights in Conflict', *Ethics* 99, 503–19.

Waldron, Jeremy (2012), *The Harm in Hate Speech* (Cambridge, MA: Harvard University Press).

Weinstein, James, and Hare, Ivan, eds (2009), *Extreme Speech and Democracy* (Oxford: Oxford University Press).

Weinstein, James (2009), 'Extreme Speech, Public Order, and Democracy: Lessons from The Masses', in James Weinstein and Ivan Hare, eds, *Extreme Speech and Democracy* (Oxford: Oxford University Press),

Woods, Lorna (2019), 'The Duty of Care in the Online Harms White Paper', *Journal of Media Law* 11, 6–17.

Woods, Lorna, and Perrin, William (2019), *Online Harm Reduction: A Statutory Duty of Care and Regulator* (Dunfermline: Carnegie UK).

[1] I assume, for the sake of this entry, that such a free speech principle is defensible in the first place. Some scholars doubt this (see Alexander 2005, though cf. Kendrick 2017).

[2] There is much debate about how to define hate speech, which I do not pursue here; for some varying approaches, see Brison (1998: 313); Brown 2017a, 2017b; Quong 2010: 305n; Waldron 2012: 8–9).

[3] I will largely focus on the categories of terrorist advocacy and hate speech, as they are the most pernicious forms of dangerous speech online and raise the thorniest free speech issues. Violent pornography is yet another much-discussed category, which also raises difficult free speech issues (see Scoccia 1996). And there are other forms of dangerous speech online, too, such as recipes for building bombs and instructions on how to commit crimes effectively (see Schauer 2019).

[4] This possibility is implied by Heinze (2016); he notes that democratic legitimacy sometimes needs to be compromised to achieve fundamental governmental aims, such as security—though he doubts that this is ever empirically necessary in longstanding stable, prosperous democracies. For related discussion, see Reid (2020).

[5] The thesis that the truth is bound to prevail through an open 'marketplace of ideas'—a view strongly associated with Mill, albeit controversially (Gordon 1997)—has been highly discredited in light of the huge empirical literature on cognitive bias. For a terrific review of the relevant empirical literature, see Bambauer (2006).

[6] This leaves open the important question of what counts as expressing a view. For example, does it qualify as sharing extreme content to 'like' someone else's post sharing that content, thereby promoting it in one's feed? For discussion, see Koltay (2019: 148).

[7] My own view is that it is possible to interpret Shiffrin's theory in a manner compatible with restricting extremist speech; for this argument, see Howard (2019b: 228–230). One

important implication of Shiffrin's view is that *insincere* speech (e.g. by bots or those deliberately sewing discord by spewing inauthentic hateful sentiments) is largely unprotected by free speech, and so, in principle, permissibly regulated.

[8] It is precisely on the condition that social media platforms refrain from exerting strong control over users' speech that they are, at the time of this publication, granted considerable immunity for users' illegal speech by the widely disputed Section 230 of the Communications Decency Act in the United States. For philosophical reflection on Section 230, see Franks (2019).

[9] This is spelled out in Article 14 of the Electronic Commerce Directive. A notice-and-takedown approach presently applies in the United States as well, but is largely limited to issues of copyright infringement, as per the Digital Millennium Copyright Act. Notice-and-takedown also characterizes the controversial Network Enforcement Act in Germany (*Netzwerkdurchsetzungsgesetz*), enacted in 2017.

[10] The discourse of complicity is not typically used in conjunction with this debate, but I believe it is a plausible framework within which to capture the nature of the wrong as a moral matter. Whether we should think of social media companies as *genuine legal accomplices* in the crimes committed by their users, such that they could be criminally prosecuted for some new inchoate offence ('criminal platforming'), is a further policy question.

[11] One promising proposal is that content flagged as extreme by artificial intelligence could be 'quarantined' prior to its review by human moderators—whereby prospective viewers would be notified before seeing it that it is potentially hateful (Ullmann and Tomalin 2020).

[12] The idea that social media networks have positive responsibilities not simply to take down harmful speech, but also to keep up legitimate speech is certainly reflected in the popular backlash to cases in which networks remove clearly valuable content, as when Facebook mistakenly removed a famous photograph from the Vietnam War; see https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo, accessed 11 August 2021.

[13] Different complications are raised by the fact that authoritarian countries have pushed for clearly excessive and impermissible regulation of social media companies (e.g. demanding them to remove content critical of state policy). If the price of doing business in an authoritarian country is to serve as a tool for the repression of citizens' legitimate speech, this is too great a cost.

[14] While I have focused on citizens' counter-speech in this section, it is also possible for social media companies themselves to engage in counter-speech (e.g. by putting warning labels around certain content indicating that it violates their community standards). In the case of extreme speech that comes in the form of misinformation, companies can also post links to fact-checking websites. And companies can also combine counter-speech with other methods, such as when Twitter places extreme speech behind an interstitial screen, forcing users to click through to see it and limiting the possibility of re-tweeting without comments. This occurred in response to US President Donald Trump's claim—'When the looting starts, the shooting starts', which was interpreted as an incendiary threat against Black Lives Matter protesters (see Hern 2020).

Page 1: [1] Deleted Howard, Jeffrey 02/12/2021 16:04:00