

THE ETHICS OF SOCIAL MEDIA: WHY CONTENT MODERATION IS A MORAL DUTY


JEFFREY W. HOWARD
University College London

This article defends platforms' moral responsibility to moderate wrongful speech posted by users. Several duties together ground and shape this responsibility. First, platforms have duties to defend others from harm when they can do so at reasonable cost. Second, platforms have a moral duty to avoid complicity with users' wrongfully harmful or dangerous speech. I will argue that one can be complicit in wrongs committed by others by supplying them with a space in which they will foreseeably commit them. For platforms, proactive content moderation is required to avoid such complicity. Further, platforms have an especially stringent complicity-based duty not to amplify users' wrongful speech, thereby increasing its harm or danger. Finally, platforms have a duty not to enable new wrongs by amplifying otherwise innocuous speech that becomes wrongfully harmful only through amplification. I close by considering an objection—that content moderation by platforms constitutes an objectionable form of private censorship—explaining how it can be answered.

Keywords: social media, content moderation, free speech.

1. Introduction

Should Facebook ban videos depicting graphic violence? Should X remove posts that spread misinformation about COVID-19? Should YouTube down-rank videos that promote conspiracy theories? Should TikTok ban war propaganda? These questions, which have flooded the news headlines over the past several years, concern the ethics of content moderation—the systems through which platforms govern the speech of their users. These rules are made by

Contact: Jeffrey W. Howard <jeffrey.howard@ucl.ac.uk>
 <https://orcid.org/0000-0002-6521-9228>

trust-and-safety teams working within the companies, who legislate a vast array of policies concerning threats, incitement, graphic content, hate speech, sexual content, misinformation, bullying and harassment, spam, self-harm content, and much else.¹ Once their feasibility is tested by engineers, these rules are subsequently enforced by a complex bureaucracy of content moderation workers and artificial intelligence systems (Klonick 2018; Douek 2022; Gorwa et al. 2020; Roberts 2019).

This private governance of online speech is hugely consequential; billions of people use social media, and millions upon millions of posts are removed each year.² As Evelyn Douek has aptly noted, ‘Facebook alone makes more speech decisions every day, perhaps even every hour, than the Supreme Court ever has in its entire history’ (Douek 2022). Some have argued that this form of *content moderation* is an illicit form of private censorship, violating the moral rights of speakers and their prospective audiences to communicate.³ Yet a burgeoning consensus holds that platforms indeed ought to engage in content moderation (see Suzor 2019; Gillespie 2018). There has been substantial debate as to how such moderation should be subjected to greater oversight and accountability, as with the European Union’s Digital Services Act and the United Kingdom’s Online Safety Act. Even so, a core foundational issue has not, in my view, been addressed with sufficient philosophical precision: why platforms have a duty to engage in content moderation in the first place. Without an adequate philosophical theory to answer that question, we are poorly positioned to defend content moderation against its critics who would prefer the Internet to be a Wild West of unfettered speech.

This article seeks to analyze the grounds of platforms’ obligations to moderate content. I will defend the thesis that social media companies (SMCs) indeed have a moral duty to engage in content moderation of wrongfully harmful or dangerous speech. I will argue that one can be complicit in wrongs committed by others by supplying them with a space in which they will foreseeably commit them, unless one takes reasonable measures to minimize those wrongs’ occurrence. Content moderation is required to avoid such complicity. I also argue that platforms have an especially stringent complicity-based duty not to further amplify users’ wrongful speech, which increases the (risks of) harm it

1. See, for example, ‘Facebook Community Standards’, Meta, n.d. <https://transparency.fb.com/en-gb/policies/community-standards/> [accessed July 3, 2024], which governs speech on Facebook and Instagram.

2. See, e.g., ‘How Meta Enforces Its Policies’, Meta, n.d., <https://transparency.fb.com/en-gb/enforcement/> [accessed July 3, 2024]; ‘Rules Enforcement’, X, n.d., <https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec> [accessed July 3, 2024].

3. This argument underpins the recent Texas statute forbidding ‘viewpoint discrimination’ by social media platforms, upheld by the Fifth Circuit Court of Appeals. See *NetChoice, LLC, v. Paxton*, No. 21–51178 (5th Cir.), No. 22–555. A variation of this position is defended in Kramer (2021), which I discuss later.

threatens—enabling it to reach larger audiences and crowding out ameliorative counter-speech. Finally, while my main concern is with speech that would be wrongful independent of its broader amplification, I argue that platforms have a duty not to impose new wrongful harms by amplifying otherwise innocuous speech that becomes wrongfully harmful only through amplification.

In section 2, I will begin to elaborate my case that platforms have a moral duty to combat this speech. Sections 3 and 4 then focus on duties to refrain from complicity with others' wrongful speech. Section 5 focuses on noncomplicity duties not to create new harms by aggregating and amplifying speech that would otherwise be innocuous. Finally section 6 concludes by addressing the objection from free speech.

Before proceeding, a final preliminary point is in order. In this article I will explore when and why platforms have a moral duty to remove wrongfully harmful or dangerous speech on their platforms. I take it for granted that some speech, such as racist bullying, causes wrongful harm; other speech, like encouraging violence against innocent persons, wrongfully endangers others (whether or not it ultimately eventuates in harm). For example, elsewhere I argue that one such category is dangerous advocacy, speech that endangers others by advocating clearly wrongful harms (Howard 2019). Another is the closely related (and overlapping) category of hate speech, which defames members of socially vulnerable groups as inferior or dangerous. A further category is dangerous disinformation or misinformation, which endangers people by communicating falsehoods (usually either by fabricating a nonexistent threat or denying the existence of a real threat).⁴ Plausibly, speech encouraging or otherwise promoting grievous self-harm—at least that directed toward vulnerable audiences such as children—is also wrongfully dangerous. No doubt there are other categories to be added to this list, as with forms of bullying, harassment, and credible threats of wrongful harm, as well as criminally instructional speech (e.g., bomb-making recipes). I will use *wrongful speech* as a placeholder term of art for all such categories.⁵ The list is illustrative, but it is not exhaustive. Moreover, those who disagree with me about what speech is wrongfully harmful or dangerous, or the conditions under which it is,⁶ can nevertheless accept the general framework I offer in what follows.

4. These rules need not be identical for all users. Perhaps state officials have special responsibilities to refrain from lies and other culpable misrepresentations (Shiffrin 2022).

5. Some speech might be wrongful yet neither harmful nor dangerous. I set this complication aside for present purposes.

6. For example, later I argue that some speech only becomes wrongful when aggregated and amplified.

2. Platforms' Defensive Duties

The public debate on content moderation has historically proceeded as a quarrel about whether digital intermediaries like SMCs should be analyzed as publishers or platforms. The standard assumption was that if we view them as publishers, they are to be held morally and so potentially legally responsible for the content posted by their users, whereas if they are mere platforms, they are not to be held responsible for the content their users' post. But this debate was misguided from the start. Even if we reject the status of platforms as publishers, as I think we should, it does not follow that they escape obligation. The question instead concerns what it is reasonable to demand of platforms.

The basic insight from which to begin is that social media companies—like all corporations—are corporate agents (Pettit and List 2011: 182; Hess 2013; Pasternak 2017; Dan-Cohen 1986),⁷ and as agents, they have a range of positive and negative duties assigned to them in virtue of that fact. Like all agents, they have positive duties to defend others from wrongful harms and negative duties not to cause or contribute to wrongful harms. These austere premises, I will argue, have enormous explanatory power in accounting for the moral demands that we rightly place on social media companies.

The least controversial and most straightforward duty with which SMCs are saddled is a duty to defend (or rescue) people from (risks of) harm. Platforms are simply in the right place at the right time, with the right capacity, to protect people. Consider a standard case of easy rescue: someone is drowning in a pond, and you can save them at reasonable cost to yourself. Or consider a case of easy other-defense: a child is being attacked by an aggressive bully, and you can intervene to stop the bully without incurring serious costs yourself. It is uncontroversial among philosophers that we all have such duties. My contention is that this simple insight provides one source (albeit a limited one) for the idea that platforms should engage in content moderation. Faced with speech that endangers life and limb, and the opportunity to mitigate that danger at reasonable cost, they have a duty to do so. While this duty cannot justify most that we might expect SMCs to do, it can justify a minimal baseline. Specifically, it can justify a duty to remove wrongful speech of which the platform becomes aware.

Conceiving content moderation, in part, as a rescue effort illuminates something fundamental about its nature. The purpose of deleting a user's post is not, in the first instance, some backward-looking effort to punish or sanction a wrongdoer (see, e.g., Goldman 2021). Even if the general deterrent effects of

7. It is immaterial whether we conceive them as bona fide group agents or not. See Pettit and List (2011) for the view that groups can be genuine agents. Of course, the fact that groups can be agents does not mean they hold primary moral rights.

content moderation are salutary side effects, they are not, on the view I am advocating, its justifying goal. Rather, the purpose of content moderation is defensive; it is an attempt to defuse the danger posed by an ongoing threat, to protect prospective victims from (wrongful) harm.⁸ My suggestion is that SMCs' natural duties require them to engage in some amount of content moderation of wrongful speech, to defend those harmed or endangered by it.

This minimal duty seems well-placed to justify a practice of notice-and-takedown, whereby platforms remove certain harmful or dangerous content once notified about it. This is the most common contemporary form of SMCs' legal obligations (under current United Kingdom and European Union law); while this article does not fully address the issue of legal enforcement, one fruitful way to begin justifying this kind of legal duty would be to point to an underlying natural moral duty to defend (or rescue) others from (risks of) harm. Yet it seems less likely that these duties can justify more demanding obligations—to set up monitoring systems powered by a combination of sophisticated AI bots and thousands of human moderators—to proactively police the platform for wrongful speech. To lean further on the rescue analogy, a natural duty of rescue enjoins us to rescue drowning children when we stumble upon them, but it does not require us to become or to hire a network of lifeguards. Nor can these duties justify any requirement to redesign platforms to minimize the likelihood of the transmission and spread of illicit content. These more demanding obligations could only be justified by appealing to further, more demanding moral duties.

What would be so objectionable about stopping here and contending that defensive rescue duties were the sole basis of platforms' moderation responsibilities? The answer is that this would miss something fundamental about the relationship between companies and the wrongful speech that they platform. Consider a traditional rescue case, where misfortune or malice imperils some victim and a prospective bystander can intervene to save him. Here, the rescuer lacks any morally significant causal relationship with the original threat. Now suppose I hire a hitman to kill you and then (after a moral epiphany) intervene to rescue you. Or suppose I negligently cause a boulder to fall down a hill toward you and I rush to rescue you from its path. These are no ordinary rescue cases; in such cases, I incur a far more stringent duty to defend you from the threat precisely because of my wrongful causal role in creating it (see, e.g., Tadros 2011). While the social media case is not perfectly analogous to either of those examples, the general point is that platforms have a more morally fraught causal

8. With respect to the duty of rescue or other-defence, it doesn't matter much that the relevant harm is wrongful (since this duty can require us to save people from tornadoes and wolves, who can't be wrongdoers). The fact that certain speech is wrongfully harmful (or dangerous—i.e., risks harm) is mostly relevant for explaining (a) why speakers have moral duties to refrain from the speech and (b) why platforms that do nothing to combat such speech count as complicit with it.

relationship with incendiary content than the rescue duty alone can illuminate. This relationship, I will now argue, generates more demanding obligations to act against such content, to which I now turn.

3. Platform Complicity

What is the nature of the relationship between social media companies and wrongful speech that is posted on their platforms? It is a relationship of complicity. My proposal is that an agent can be complicit with wrongs committed by others simply by providing them with a space in which they will foreseeably commit them. Just by dint of one's ownership or control over a space, one has certain duties to minimize the likelihood that the space will be used for wrongdoing.⁹

Consider the parking garage company who refuses to install appropriate lighting or other security mechanisms to reduce the likelihood that crimes will occur. Consider the landlord who knows, or should know, that his tenant stashes trafficked children in his flat yet does nothing in response. Or consider the political leaders who offer haven to terrorists, allowing them to use their land to plan further attacks—as the Taliban did for al-Qaeda in the years before September 11, 2001. Across these examples, the relevant owner or controller enabled wrongdoing, by providing a space propitious for its occurrence, while doing nothing to mitigate the likelihood of its occurrence (and, in some cases, actively encouraging its occurrence). While this duty uncontroversially arises in physical, offline space, there is no reason to think it applies less forcefully in cyberspace. Accordingly, when a social media platform provides a platform that users will foreseeably use to engage in wrongdoing and then fails to take reasonable steps to minimize the likelihood of its occurrence, it is complicit in the speaker's wrongdoing. The second duty that justifies content moderation responsibilities, then, is a duty to avoid complicity with wrongdoing.

What is complicity? Here are the crucial ideas. First, I assume a causal conception of complicity, whereby one is morally complicit just in case one's acts or omissions wrongly and foreseeably risk contributing to wrongs committed by others (Lepora and Goodin 2013; Gardner 2007; Mackie 1974).¹⁰ To discharge

9. In my initial reflections on this topic, I argued that social media companies have duties in virtue of their role as *curators* of public discourse (Howard 2021), but the argument here does not rely on this claim.

10. Note that a complicit act might be potentially essential to the principal wrong or definitely essential. A potentially essential complicit act is an INUS condition (to use Mackie's term) of a principal wrong (i.e., an insufficient but necessary condition of an unnecessary but sufficient condition of the primary act's occurrence). A definitely essential complicit act is a particularly strong version of an INUS condition: it is an insufficient but necessary condition of every unnecessary but sufficient condition of the primary act's occurrence. For discussion, see Lepora and Goodin 2013: 61.

the duty to avoid complicity, one must take reasonable steps not to make such causal contributions.

Next, cases of complicity always involve some principal wrongdoer, on the one hand, and those who contribute to his wrongdoing, on the other. In our case, a speaker who posts a wrongful post is the principal wrongdoer, and the social media platform is a secondary wrongdoer, complicit in the wrongdoing of the principal. This is a purely conceptual point; it need not follow that the principal wrongdoer is morally worse than those who aid him. It is perfectly possible that, in the final analysis, the complicit party is more blameworthy than any one principal.¹¹ This is especially likely when we contrast any single wrongful speaker, with a relatively few number of harmful posts, with a massive platform that enables the communication of millions of harmful posts by many people.

Second, I am assuming that complicity is a partly moralized notion; for something to count as complicit, it must involve a wrongful causal contribution to the wrongdoing of others. The reason is that not all causal contributions to the wrongdoing of others are wrongful.¹² Consider again the parking garage company. If it had a duty to guarantee that there was a zero probability of crimes ever being committed in its space, it would likely be obligated to close up shop, given the inordinate difficulty of achieving such a goal. There is no categorical duty to eliminate one's contributions to wrongdoing, only to take reasonable steps to minimize one's contributions (which may or may not involve eliminating them fully). What counts as reasonable will depend, first, on how much cost the agent can be expected to bear toward this purpose. Like any duty, the duty to avoid complicity will be cost-sensitive, in that its demands cannot be unreasonably burdensome. For corporations, these costs will involve costs to owners, employees, and even shareholders. But it will also (often more importantly) involve costs to third parties. If parking garages all had to close down, since this was the only way to eliminate wholly their use as platforms for crimes, this would involve an obvious social cost, to be borne by the many who would otherwise benefit from their existence. The main reason to think that parking garage companies don't have such a stringent duty, then, is not merely that it would put garage company owners and employees out of work; it also reflects the burdens such a duty would place on the public. Or consider the example of firearms and

11. Consider the example of a police officer who abets a petty theft; surely the officer, while merely an accomplice, is nevertheless morally worse in this case than the petty thief (Lepora and Goodin 2013: 34).

12. Some acts of complicity might be pro tanto wrongful but all-things-considered justified. Consider the humanitarian organization that must bribe a warlord to access and thus provide aid in some territory—a bribe that helps the warlord buy bullets. The humanitarian organization is complicit with the killings their bribe made possible—there is genuinely pro tanto wrong done here, triggering remedial and compensatory obligations—yet in many cases there could well be an all-things-considered justification, depending on just how many lives are thereby saved.

ammunitions manufacturers, who also (on the view I am defending) have duties to take reasonable steps to reduce the likelihood that the guns and bullets they sell will be used in crimes (e.g., by making bullets traceable to their owners, thereby helping to deter their criminal use). If they had a duty to eliminate such a possibility, they could not sell guns or ammunition. Aside from any complaints the companies might have, the significant objection to that requirement would be that legitimate police and military services would then be unable to acquire weapons they need to discharge morally compulsory purposes.¹³

So, too, it goes for social media platforms (potentially among other platforms¹⁴). For platforms to reduce platforming complicity to zero, they would likely need to close down. What makes such a demand unreasonable is not merely its impact on those who work at the companies; just as salient (if not much more salient) would be the unacceptable costs on the broader public. Given the enormous potential good served by social media—the substantial opportunities these platforms offer for valuable expression, connection, and mobilization for just ends—we should tolerate some amount of unavoidable platforming of wrongful speech to enable these benefits.¹⁵ As agents, we have interests both in access to forums in which we can pursue our weighty communicative interests and in reducing the likelihood that we will be wronged while in those forums. A plausible theory will accommodate both interests.

Finally, I stress that complicity involves contributions that are *foreseeable*. The requirement of foreseeability steers a middle course between two implausible poles. One pole eschews any mental condition altogether: one is complicit just in

13. It might be suggested that the intervening agency of those who buy guns and then use them for wrongdoing means that they, rather than the companies, are responsible for any ensuing harm. Similarly, one might think that it is those who abuse platforms to harm or endanger others who should be held responsible, rather than platforms themselves. My own view is that while criminal gunmen and online abusers are certainly responsible for their own wrongdoing, this fact does not let others off the hook; manufacturers and platforms remain potential accomplices if they have failed to take reasonable steps to minimize their contributions to that principal wrongdoing. I defend the claim that intervening agency has less significance than commonly supposed in Howard 2019, following Tadros (2016).

14. While I focus here on social media platforms, the framework I offer here plausibly applies, *mutatis mutandis*, to other platforms like search engines, and even to offline platforms. I thank a referee for raising this issue.

15. Section 230 of the Communications Decency Act in the United States, which confers a broad immunity on platforms for most illegal content posted by users, captures this concern; for example, if platforms were liable for all defamatory remarks posted by users, they would have to shut down. While Section 230 raises a wide range of policy issues, it is coherent to argue that platforms should retain 230 immunity (shielding them from an onslaught of private action claims) while also arguing that platforms should be subjected to the kinds of oversight and risk assessment by regulatory bodies that will soon be in operation in the United Kingdom and European Union. This, it seems to me, holds platforms accountable without making their business untenable (and thereby compromising the social value it brings) (Kosseff 2019).

case one causally contributes, even if one doesn't know and couldn't be expected to have known one's decisions had this property.¹⁶ This position is implausible, given the interconnectedness of the world; it would counterintuitively hold that the taxi driver who drives the passenger to the nightclub where he will end up killing someone in a barfight will thereby count as an accomplice. A second pole requires the full-blown intentionality of the accomplice – what Christopher Kutz calls a 'participatory intention' (Kutz 2000: 74, 89).¹⁷ On this view, those who causally contribute to the wrongdoing of others, even with full knowledge that they do so, are not complicit because they do not share a common purpose with the principal wrongdoers. Yet intuitively, the arms vendor who sells weapons to the tyrannical dictator, who does not share an intention with the dictator but who knows (or should know) what the dictator intends to do, is complicit in the killings subsequently perpetrated. What truly matters, as Larry May puts it, is that the complicit agent 'knows, or should have known, that . . . he or she will advance whatever intentions the principal has' (May 2010; Lepora and Goodin 2013: 42). Accordingly, even if SMCs do not intend to enable hateful speakers to engage in wrongful communications, they qualify as complicit, just in virtue of the wrongful and foreseeable causal contribution they make. However, where certain wrongful uses of the platform are not foreseeable, it is a mistake to think that SMCs are violating a duty by platforming them.¹⁸

The duty to avoid complicity is more demanding than the general defensive duties invoked in the previous section. The duty not to be complicit in murder and other serious crimes, by platforming those who incite them, is very stringent.¹⁹ Accordingly, platforms should be expected to bear greater costs to discharge it. While notice-and-takedown protocols may be sufficient to discharge the former duty, they are plainly inadequate to discharge the latter. In addition to such protocols, my contention is that SMCs have a responsibility to actively police their networks for such communicative wrongdoing and to delete it when they find it. Only by doing so can they avoid the charge of complicity.²⁰

16. We might call this 'fact-relative complicity'. The language of 'fact-relative', in contrast to 'evidence-relative' or 'belief-relative', traces to Parfit (2013).

17. Goodin and Lepora (2013: 80) reject Kutz's view. Kutz's view is also rejected by Gardner (2004).

18. For example, some instances of platforming wrongful speech will not be reasonably foreseeable (e.g., for new cases of 'coded' language), and so it would be infeasible to prevent it.

19. Of course, not all duties are this stringent; different categories of wrongful speech will trigger differentially stringent moderation responsibilities, depending how harmful/dangerous the category is.

20. Given the scale of these platforms and the sheer volume of speech within them, such a herculean feat cannot be accomplished through humans alone. It requires the deployment of artificial intelligence trained to hunt for violations. The use of AI raises distinctive normative problems, which I discuss elsewhere. For a critical take on the use of AI, and the ways the tech sector overplay its benefits, see Barnes (2022). For arguments that AI struggles to satisfy our right to

4. Amplification Duties

So far, I have discussed the phenomenon of merely providing a platform on which a wrong will be perpetrated. The duty to avoid complicity, I argued, requires that SMCs take reasonable steps to reduce the likelihood that their platform will be used for wrongdoing. Content moderation reduces that likelihood.

But SMCs do more than merely provide a platform for dangerous speech. Through their algorithms, they have the power to amplify it—to increase its visibility—and when they do, their complicity only deepens. Depending on a speaker’s authority and message, the wrong of an incendiary communication can be more or less grave.²¹ But the gravity of this wrong also depends on the size and susceptibility of its audience. In the case of social media, platforms have substantial influence on precisely these variables. How many or how few people see a speaker’s incendiary post is precisely a function of platforms’ engineering, which determines what content is amplified.

What counts as amplification? Amplification will always be relative to some baseline, and the selection of a baseline will always be arbitrary, at least within a range of plausible options.²² In this sense, there is no such thing as merely providing a platform, since providing a platform counts as amplification relative to a baseline of not providing a platform. Thus when we refer to amplification, it must be comparatively, with reference to some baseline that involved less visibility. So, suppose we stipulate the relevant baseline as a configuration where users can post content that is then findable by others who follow or search for them. Any platform-enabled visibility increases beyond this baseline (e.g., through recommender-algorithms that pipe the content into other users’ newsfeeds) counts as amplification relative to that baseline. In some cases, amplification may be an intentional decision by platform designers (e.g., the decision to amplify trustworthy COVID news content during the pandemic). But nearly always, amplification is the spontaneous result of platform algorithms, which are designed to show people more of the content that is likely to optimize their engagement with

explanation (given the inscrutability of their ‘black box’ deliverances), see Vredenburg (2022) and Lazar (2022).

21. For example, hate speech is especially pernicious when uttered by a (perceived) practical authority, who has the (perceived) illocutionary power to authorize, permit, and license the subordination of targeted groups (Langton 2018). Likewise, harmful misinformation is especially harmful when communicated by (perceived) epistemic authorities.

22. For an accessible overview of how recommender/amplification systems work and a good discussion on what counts as amplification, with attention to the baseline issue, see Thorborn, Stray, and Bengani (2023). For related analysis of the difficulties in measuring amplification, and the unavailability of any neutral baseline, see Lum and Lazovich (2023). See also Miller (2021).

the platform—and hence increase their advertising revenue. This story—of how social media manages our attention—is by now quite familiar.²³

My claim is that when a platform amplifies wrongful speech, increasing its visibility, it thereby makes a greater causal contribution to the speaker's wrongdoing—making his principal wrongdoing worse than it would otherwise be. Like the gun vendor who sells the terrorist a larger weapon, enabling him to kill more people, platform amplification enables wrongful speakers to commit a greater wrong. This is so even if a platform does so unintentionally (e.g., when it is merely the foreseeable result of the platforms' algorithmic design).

Amplification can occur in different ways. A standard form of amplification increases visibility by *expanding* the audience, making more people see it. This plausibly makes the speech more dangerous or harmful since there are more people likely to be exposed to it. This is true both for indirectly and directly harmful speech. In the case of speech that incites violence, for example, such amplification increases the pool of people who may be inspired to act on such exhortations. For speech that promotes hatred, amplification increases the pool of people who may be persuaded to adopt bigoted views. For other kinds of harmful speech—like the subset of hate speech that functions as a direct psychological attack—this expansion has the effect of increasing the number of victims who see, and are thereby directly harmed by, the speech.

Even if there is no *net* visibility increase for the general platform population, platforms may amplify speech in a second sense—not by increasing the number of people exposed to it but by *flooding* particular users' feeds with such speech, overwhelming them with exposure to it. For indirectly harmful speech-acts like incitement to racial hatred, it stands to reason that those flooded with the relevant speech are at greater risk of radicalization by it, as their feed is dominated by speakers promoting bigoted views. For directly harmful speech-acts like bullying and racist harassment, flooding targets' feeds with such speech is manifestly more harmful. Moreover, such flooding crowds out forms of ameliorative counter-speech—in which others 'talk back' against the wrongful speech, offering their own countervailing points of view. Such counter-speech (e.g., arguing against an extremist speaker's violent interpretation of their religion or affirming an anti-racist narrative) can both dissuade listeners from acting on incitements and can offer reassurance to the direct targets of harmful speech and even block its wrongful effects (Langton 2018; Lepoutre 2021). In this way, amplification of wrongful speech by platforms can stymie the efforts of counter-speakers to prevent or block the relevant harms—by reducing the likelihood that those counter-speakers' posts will be seen.

23. For incisive philosophical work on the attention economy, see Lazar (2023). For a historical perspective, see Wu (2016). See also Thorborn (2022).

The normative upshot of this may seem redundant. After all, I have already established that SMCs have a moral duty to police their networks for wrongful speech and remove it. The point here is that when a company provides a platform for wrongful speech, and then takes measures to increase its audience beyond what it would otherwise be, they position the speech to cause or risk even greater harm than it otherwise would. Accordingly, the platform makes a greater causal contribution to wrongdoing, breaching a more stringent duty. It is reasonable to demand that they bear greater costs to avoid that outcome through more intensive content moderation of wrongful content—since, if that content is not removed, it will (given what we know about the platforms’ design) almost certainly be amplified.²⁴

To be sure, these costs are not unlimited. As I argued earlier, to reap the benefits of a digital public sphere, platforms cannot reasonably be expected to eliminate the prospect that their products will be abused for wrongdoing, since they therefore would not be able to exist at all. Still, given the profits that some of these platforms earn, they have ample resources from which to draw to ensure their products are safe to use and their negative externalities are reduced. It is therefore fully appropriate to demand that platforms conduct assessments of the risks posed by their products, setting out clear plans to reduce the risks—and be held accountable for doing so.²⁵

5. Transforming Innocuous Speech

My main concern so far has been speech that is wrongful already independent of any further amplification; the further amplification serves simply to make it worse. For example, terrorist incitement is wrongful even when it is not amplified to a mass audience. But not all forms of wrongful speech are like this. Some speech only becomes wrongfully harmful when *aggregated* and *amplified* alongside lots of similar speech.²⁶ No single unit of speech causes substantial harm.

24. A reviewer asks why it wouldn’t be enough simply for platforms to alter their algorithmic systems such that this content is not amplified in the first place. While this would be preferable to doing nothing, I stress that such speech is wrongfully harmful even if it is not amplified. Reducing amplification for this content, then, is not enough. (In contrast, the next section discusses cases where reducing amplification would be sufficient, at least potentially.)

25. This is a basic insight of the EU Digital Services Act. Note that, in principle, governments could require risk assessments while remaining content-neutral about what exact speech ought to be removed—thereby (better) avoiding (in the US context) a First Amendment challenge. In other jurisdictions, content-based directives can be consistent with free speech protections; for a philosophical defence of the claim that content- and even viewpoint-based restrictions on (some) wrongful speech can be compatible with a proper commitment to freedom of expression, see Howard 2019.

26. Strictly speaking, we can imagine four sets of cases: (1) speech that is wrongful even when not aggregated and amplified, (2) speech that is wrongful only when aggregated with similar

Some forms of dangerous misinformation may fall into this category. Consider the example of climate change denial, peddled by people duped into believing it is or might be true. Such misinformation is dangerous, but only when aggregated and amplified such that it floods the information ecosystem. In small doses, it is no more objectionable than speech suggesting the Earth is flat—something that hardly violates a moral duty. Unlike the terrorist incitement case, the individual user is not plausibly described as a culpable wrongdoer. It is only when the speech is aggregated and amplified that such speech becomes wrongful, because only in such circumstances does their speech help to constitute a flood of misinformation that can genuinely lead to harm (namely, in propagating a dangerous falsehood that endangers countless living and future people by stymying climate action). Or consider the example of militaristic, violent rhetoric ('We're going to slaughter our political opponents') or rhetoric using hyperbolic insults of one's political opponents ('They're a bunch of traitors'); such speech is ubiquitous in divided democracies, but it seems unlikely that any one unit of such speech constitutes a moral wrong. Yet when amplified and aggregated, it may well be that such speech becomes wrongful, by serving as a constituent component of a wave of speech that coarsens the public discourse and gradually attenuates citizens' inhibitions against violence. Some 'pile-on' campaigns of harassment work like this; a single critical post from a stranger is innocuous, yet when receiving thousands of such posts from countless strangers, it can do real harm (Billingham and Parr 2019).²⁷ If this conjecture is correct, this would be a further way in which platforms act wrongly through their amplification practices.

In such cases, algorithmic systems transform speech that would otherwise be innocuous into speech that is genuinely wrongful. They make the moral difference.²⁸ Conversations that might be utterly anodyne when occurring offline

speech (regardless of whether amplified), (3) speech that is wrongful only when amplified (regardless of whether aggregated with similar speech), and (4) speech that is wrongful only when amplified and aggregated with similar speech. I have focused above on 1 and in this subsection am focused on 4, but that is not to deny the importance of 2 and 3. What I say here applies, *mutatis mutandis*, to them, too. Seth Lazar has independently developed some similar insights, distinguishing between stochastic harm (whereby aggregation increases the probability of harm) and collective harm (whereby lots of small harms aggregate into a large harm). See Lazar (2023).

27. Thanks to an editor for suggesting this point.

28. It is striking that platforms arguably already do this for speech that is wrongful independent of amplification. For example, if a speaker posts a slur to a platform, it is because the platform's systems enable it to be posted that the speaker thereby becomes a wrongdoer. Had the post remained for her own private viewing, it wouldn't have constituted a wrong (at least, not an instance of wrongful harm). In this way, platforms already make the difference between whether some people are wrongdoers or not. What is distinctive about the amplification power is that it takes communications that would likely be innocuous offline and puts them in a context where they can cause or risk real harm.

become seriously harmful online in virtue of algorithmic systems. Consider again the naïve citizen airing misguided questions about climate change; through the aggregation of her speech with similar speech, and its consequent amplification, she becomes transformed into a co-constitutive producer of a flood of falsehoods that endangers the planet.

I am assuming here that such citizens have reason to foresee that their speech will be aggregated and amplified in just this way; accordingly, when it is aggregated and amplified, they become liable. But we could imagine cases in which users *had no reason to anticipate* that their speech would be aggregated and amplified in a way that would then cause harm. Where this is so, they are not plausibly described as wrongdoers (not, anyway, on an evidence-relative standard). Yet if the platform nevertheless had grounds to believe that such speech would emerge through its algorithms, *it* — the platform — remains a wrongdoer, breaching duties it owes to others. In such cases, it is the *only* wrongdoer.

Because such aggregatively harmful speech only becomes wrongful once amplified, it isn't clear that *removal* is necessary to redress its harm. Suppose that a policy of *deamplifying* climate misinformation — deliberately reducing its visibility down to the baseline where it is allowed and findable but not promoted — adequately redressed its harms, as seems plausible. If so, deamplifying such speech would mean that those engaging in it would no longer qualify as wrongdoers, since their speech would not be innocuous (Gillespie 2022; Keller 2021).

6. The Objection from Free Speech

I have argued that SMCs have a weighty moral responsibility to engage in vigorous content moderation of wrongful speech. This responsibility is justified by a suite of underlying moral duties: natural rescue duties to defend those wronged by such speech, duties to avoid complicity with users' wrongful speech (which is exacerbated through greater amplification), and duties to refrain from rendering otherwise innocuous content harmful through amplification.

Showing that platforms have this moral responsibility is necessary to justifying its codification into a legal responsibility. Yet it is not sufficient; one could accept everything I have argued about platforms' moral duties while also denying that those moral duties ought to be legally enforced. A strong version of such a view would hold that while speakers have moral duties to refrain from wrongful speech, and platforms have duties not to platform or amplify it, the coercive enforcement of such duties would violate the moral right to freedom of expression. On this view, speakers' have moral rights to communicate all sorts of wrongful messages, and prospective audiences have moral rights to hear those messages, notwithstanding the danger or harm they pose. A more contingent,

instrumental version of the view would hold that legal enforcement is not in principle impermissible, but in practice, it is simply too risky to grant the state the authority to enforce platforms' and speakers' moral duties, given the potential for abuse and overreach. American liberals who champion the orthodox interpretation of the First Amendment yet insist on robust content moderation, hold one or both of these views (Miller 2021). (Note that the prevailing interpretation of the First Amendment largely forbids the state from punishing wrongful speakers or forcing SMCs to remove much, though importantly not all, wrongful speech.²⁹)

Suppose free speech concerns (whether noninstrumental or instrumental) militate against the legal enforcement of speakers' and platforms' moral duties. This fact does nothing to reduce the significance or stringency of those moral duties. Even if SMCs are off the legal hook, this does not immunize them from their moral requirements. And this has normative implications: SMCs are justifiably subjected to public opprobrium, advertiser boycotts, and other social sanctions in response to their failures to moderate dangerous content with sufficient vigor. It is even permissible for officials to pressure companies to discharge their responsibilities, even though it would be impermissible (on the view I am entertaining) to *force* companies to do so. Take a different example: the duty to vote, which let us suppose is a weighty moral duty but, for various reasons, shouldn't be legally codified and enforced. It is nevertheless wholly appropriate for officials to strongly encourage citizens to vote. Likewise, there is nothing incoherent in supposing that all things considered, the state shouldn't force SMCs to remove incitement, yet it may express and defend its view that they have a moral duty to do so.

The upshot, then, is that concerns of free speech do not pose an obstacle to the thesis that social media companies have a moral duty to engage in vigorous content moderation of wrongful speech. But now consider an objection to this view. The objection holds that *if* it would be impermissible for the state to restrict certain speech, it is likewise impermissible for social media platforms to restrict it. On this view, it is not even permissible for social media platforms to adopt rules against (most) wrongful content.³⁰

The suggestion, now prominent among Republican lawmakers in the United States, is that the large social media platforms constitute *public forums*—spaces so central for the operation of public discourse and the exercise of communicative freedoms that they are thereby obligated to refrain from restrictions on users' legitimate speech. On this view, if it is impermissible for a government to restrict a certain category of speech, so too is it impermissible for a large social media

29. For incisive reflection on the First Amendment in the age of social media, with an emphasis on ways in which online speech can be weaponized (especially by governments) to the detriment of public discourse, see Wu (2018).

30. This piece was accepted before the publication of Messina (2023), so I regrettably do not engage with that book's argument here.

platform to restrict it.³¹ Matthew Kramer has recently defended just this claim: ‘Social media platforms such as Facebook and Twitter and YouTube have become public fora. Although the companies that create and run those platforms are not morally obligated to sustain them in existence at all, the role of controlling a public forum morally obligates each such company to comply with the principle of freedom of expression while performing that role. No constraints that deviate from the kinds of neutrality required under that principle are morally legitimate’ (Kramer 2021: 58–59).

Given that Kramer endorses the First Amendment view that substantial amounts of wrongful speech is intrinsically protected, he thereby holds that social media platforms act unjustly when they adopt prohibitions against it. For Kramer, this encompasses speech advocating criminal violence such as terrorism, speech advocating racial and religious hatred, misogynistic pornography, and much else.

There is reason to doubt this position. First, it is certainly plausible that *government* channels or pages on social media networks are public forums in the sense that triggers the duty not to discriminate against legitimate speech.³² But it is not clear why this is a general duty that applies across the entirety of a social media network. Social media networks routinely restrict all manner of content that is protected by any plausible theory of free speech—such as low-quality commercial advertising (spam) and sexual content. Virtually everyone agrees that it would be unacceptable for the state to punish this speech; yet it is very counterintuitive to think that Facebook violates its users’ rights to free speech by doing so. Yet if Kramer were correct, Facebook would be duty bound to allow pornography on its platform—an implausible result.³³

Next, while Kramer is surely right that SMCs are important sites for public discourse, so too are newspaper op-ed pages, television news and debate shows, and much else. Their status as such seems to trigger a demand for public justification and accountability for their activities, but it is not clear why they must showcase the same viewpoint neutrality that presumptively binds the state’s speech-impacting decisions. In the case of traditional media, their contribution to public discourse seems to depend on making a wide range of editorial judgments about what views are reasonable and worth taking seriously. While social media companies of course have a different role than traditional media, that role (however one wishes to specify it) seems compatible with the moral duty I have sketched to

31. For relevant litigation in the state of Texas that is likely to be heard eventually before the US Supreme Court, see *NetChoice v. Paxton* (2022).

32. This was the central claim at issue in *Knights First Amendment Institute v. Trump* (2019). For relevant discussion, see Nunziato (2019).

33. Of course, one could hold that pornography falls outside the legitimate ambit of free speech. Yet Kramer himself rejects that position. See Kramer (2021: p. 160).

guard against their spaces' weaponization by nefarious actors. Note that such a view is compatible with the claim that platforms' voluntary commitment to principles such as 'user voice', and general obligations as public-facing businesses, generates an obligation not to discriminate against certain views or groups without justification. One can complain when a social media platform's content decisions seem to place excessive burdens on speech by racial minorities, without thinking that the platform has duties to respect free speech akin to the state's.

Suppose, however, Kramer were to come up with an argument establishing that platforms are, in fact, under the same free-speech duties as states. It would matter why states' duties are what they are. Consider those who think the state shouldn't ban hate speech because of instrumental anxieties about the abuse of state power; on such a view, hate speech is not intrinsically protected as free speech but it remains all-things-considered impermissible for the state to restrict it. Such a view recognizes the distinctive evils involved when states misuse their awesome capacities. But of course platforms do not have anything like the awesome capacities of states (e.g., to imprison people for the views they express). A policy of subjecting certain speakers to criminal punishment, and a company policy of moderating speakers' posts, involve manifestly different costs, plausibly leading to different standards for what speech may be restricted. Those (unlike Kramer) who endorse the orthodox First Amendment view on instrumental grounds (because of the dangers of giving the state too much power), then, could accept some kind continuity between states' and platforms' duties without thinking their speech policies need to be identical. One could, in other words, think that states shouldn't ban hate speech while also thinking that platforms should.

7. Conclusion

Many citizens are increasingly inclined to argue that platforms have duties to moderate wrongful speech. Yet we have lacked a compelling argument as to why. My aim in this article has been to set out the positive case for such a responsibility. Much more would need to be said to offer a complete theory of the duty to engage in moderation; in particular, such a theory would need to explore the importance of avoiding bias, especially in the use of automated technologies for moderation. It would also need to address the harms that befall content moderation workers themselves.³⁴ Further, the duty to moderate is only one of many

34. Both issues are astutely addressed in Frost-Arnold (2022: ch. 2), which uses tools from social epistemology to explore both the exploitation of workers and the epistemic injustices that current moderation systems produce.

moral duties held by platforms. The analysis I have offered here is broadly compatible with a vision of single-mindedly profit-maximizing platforms, hellbent on maximizing engagement and thus advertising revenue—but (inconveniently) side-constrained by the stringent moral duties I have set out. Yet it would be a mistake to think this is where the story ends. As agents, platforms do not simply have the positive duties of rescue and negative duties to refrain from imposing or contributing to wrongful harms that I have canvassed here; they also plausibly have civic duties to help maintain a salutary public discourse—to bring the best out of their users rather than the worst. Thinking through what those duties involve will require a fuller theory of the positive role that platforms should play in a liberal democracy—and possibly greater disruption to the business model of social media as it currently functions.

Acknowledgments

Thanks to the editors and referees of this journal for very helpful feedback, which led to several improvements. I am grateful to audiences at the 2022 American Philosophical Association Pacific Division conference in Vancouver, the 2022 American Political Science Association conference in Montreal, the 2022 MANCEPT Workshops at the University of Manchester, the Law & Philosophy Colloquium at Pompeu Fabra University, the CELPA Seminar at the University of Warwick, the Oxford Institute for Ethics in AI's Lunchtime Research Seminar, the Conceptual Foundations of Conflict Project seminar at the University of Southern California, and the Legal & Political Theory Colloquium at UCL. I am grateful to written comments from, or detailed conversations with, Diana Acosta, Etienne Brown, Juan Espindola, Leslie Kendrick, Seth Lazar, Spencer McKay, Tom Parr, Seana Shiffrin, Robert Simpson, and John Tasioulas. I am thankful to UKRI for research funding (UKRI grant MR/V025600/1).

References

- Barnes, Michael. 2022. 'Online Extremism, AI, and (Human) Content Moderation', *Feminist Philosophy Quarterly*, 8, pp. 1–28
- Billingham, Paul, and Tom Parr. 2019. 'Online Public Shaming: Virtues and Vices', *Journal of Social Philosophy*, 51.3, pp. 371–90
- Dan-Cohen, Meir. 1986. *Rights, Persons, and Organizations* (University of California Press)
- Douek, Evelyn. 2022. 'Content Moderation as Systems Thinking', *Harvard Law Review*, 136, pp. 526–607
- Douek, Evelyn. 2022. 'The Siren Call of Content Moderation Formalism', in *Social Media, Freedom of Speech, and the Future of our Democracy*, ed. by Lee Bollinger and Geoffrey Stone (Oxford University Press), p. x

- Frost-Arnold, Karen. 2022. *Who Should We Be Online?* (Oxford University Press)
- Gardner, John. 2007. 'Complicity and Causality', in *Offences and Defences: Selected Essays in the Philosophy of Criminal Law* (Oxford University Press)
- Gardner, John. 2004. 'Review of Complicity: Ethics and Law for a Collective Age by Christopher Kutz', *Ethics*, 114, pp. 827–30
- Gillespie, Tarleton. 2022. 'Do Not Recommend: Reduction as a Form of Content Moderation', *Social Media + Society*, 8, doi:10.1177/20563051221117552
- Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (Yale University Press)
- Goldman, Eric. 2021. 'Content Moderation Remedies', *Michigan Technology Law Review*, 28, pp. 1–59
- Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', *Big Data & Society*, 7.1, pp. 1–15
- Hess, Kendy M. 2013. "If You Tickle Us . . .": How Corporations Can Be Moral Agents without Being Persons', *Journal of Value Inquiry*, 47, pp. 319–35
- Howard, Jeffrey W. 2019. 'Dangerous Speech', *Philosophy & Public Affairs*, 47, pp. 208–54
- Howard, Jeffrey W. 2021. 'Extreme Speech, Democratic Deliberation, and Social Media', in *The Oxford Handbook of Digital Ethics*, ed. by Carissa Véliz (Oxford University Press), pp. 181–200
- Keller, Daphne. 2021. 'Amplification and Its Discontents', Knight First Amendment Institute <<https://knightcolumbia.org/content/amplification-and-its-discontents>>
- Klonick, Kate. 2018. 'The New Governors: The People, Rules, and Processes Governing Online Speech', *Harvard Law Review*, 131, pp. 1599–1670
- Knight First Amendment Institute v. Trump*, 928 F.3d 226 (2019)
- Kosseff, Jeff. 2019. *The Twenty-Six Words that Created the Internet* (Cornell University Press)
- Kramer, Matthew. 2021. *Freedom of Expression as Self-Restraint* (Oxford University Press)
- Kutz, Christopher. 2000. *Complicity: Ethics and Law for a Collective Age* (Cambridge University Press)
- Langton, Rae. 2018. 'Blocking as Counter-Speech', in *New Work on Speech Acts*, ed. by Daniel Fogal (Oxford University Press), pp. 144–64
- Langton, Rae. 2018. 'The Authority of Hate Speech', *Oxford Studies in Philosophy of Law*, vol 3 (Oxford University Press), pp. 123–52
- Lazar, Seth. 2023. 'Communicative Justice and the Distribution of Attention', Knight First Amendment Institute at Columbia University <<https://knightcolumbia.org/content/communicative-justice-and-the-distribution-of-attention>>
- Lepora, Chiara, and Robert E. Goodin. 2013. *On Complicity and Compromise* (Oxford University Press)
- Lepoutre, Maxine. 2021. *Democratic Speech in Divided Times* (Oxford University Press)
- Lum, Kristian, and Tomo Lazovich. 2023. 'The Myth of "The Algorithm": A System-Level View of Algorithmic Amplification', Knight First Amendment Institute at Columbia University, <<https://knightcolumbia.org/content/the-myth-of-the-algorithm-a-system-level-view-of-algorithmic-amplification>>
- Mackie, J. L. 1974. *Cement of the Universe* (Clarendon Press)
- May, Larry. 2010. *Genocide: A Normative Account* (Cambridge University Press)
- Messina, J. P. 2023. *Private Censorship* (Oxford: Oxford University Press)
- Miller, Erin. 2021. 'Amplified Speech', *Cardozo Law Review*, 43, pp. 1–69
- NetChoice v. Paxton*, No. 21–51178 (5th Cir. 2022)

- Nunziato, Dawn. 2019. 'From Town Square to Twittersphere: The Public Forum Doctrine Goes Digital', *BU Journal of Science and Technology Law*, 25, pp. x–x
- Pasternak, Avia. 2017. 'From Corporate Moral Agency to Corporate Moral Rights', *Law & Ethics of Human Rights*, 11, pp. 135–59
- Pettit, Philip, and Christian List. 2011. *Group Agency* (Oxford University Press)
- Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale University Press)
- Seth Lazar. 2022. 'Legitimacy, Authority, and the Political Value of Explanations', *Computers and Society*, doi:10.48550/arXiv.2208.08628
- Shiffrin, Seana. 2022. 'Unfit to Print: Government Speech and the First Amendment', *UCLA Law Review*, 69, pp. 986–1026
- Suzor, Nicholas P. 2019. *Lawless: The Secret Rules That Govern Our Digital Lives* (Cambridge University Press)
- Tadros, Victor. 2016. 'Permissibility in a World of Wrongdoing', *Philosophy & Public Affairs*, 44, pp. 101–32
- Tadros, Victor. 2011. *The Ends of Harm: The Moral Foundations of Criminal Law* (Oxford University Press)
- Thorborn, Luke, Jonathan Stray, and Priyanjana Benani. 2023. 'Making Amplification Measurable', Medium <<https://medium.com/understanding-recommenders/making-amplification-measurable-2be548e5986c>>
- Thorborn, Luke, Jonathan Stray, Priyanjana Bengani. 2022. 'What Does it Mean to Give Someone What They Want? The Nature of Preferences in Recommender Systems', Medium <<https://medium.com/understanding-recommenders/what-does-it-mean-to-give-someone-what-they-want-the-nature-of-preferences-in-recommender-systems-82b5a1559157#:~:text=The%20definition%20of%20preferences%20we%20find%20most%20useful,but%20it%E2%80%99s%20designed%20to%20highlight%20two%20important%20ideas>>
- Vredenburg, Kate. 2022. 'The Right to Explanation', *Journal of Political Philosophy*, 30, pp. 209–29
- Wu, Tim. 2018. 'Is the First Amendment Obsolete?' *Michigan Law Review*, 117, 547–81
- Wu, Tim. 2016. *The Attention Merchants: The Epic Scramble to Get Inside Our Heads* (Knopf)