Philosophy & Technology (2023) 36:78
https://doi.org/10.1007/s13347-023-00685-w

RESEARCH ARTICLE

The Authority to Moderate: Social Media Moderation and its Limits

Bhanuraj Kashyap[1] · Paul Formosa[1]

Received: 17 August 2023 / Accepted: 2 December 2023

Abstract

The negative impacts of social media have given rise to philosophical questions around whether social media companies have the authority to regulate user-generated content on their platforms. The most popular justification for that authority is to appeal to private ownership rights. Social media companies own their platforms, and their ownership comes with various rights that ground their authority to moderate user-generated content on their platforms. However, we argue that ownership rights can be limited when their exercise results in significant harms to others or the perpetration of injustices. We outline some of the substantive harms that social media platforms inflict through their practices of content moderation and some of the procedural injustices that arise through their arbitrary application of community guidelines. This provides a normative basis for calls to better regulate user-generated content on social media platforms. We conclude by considering some of the political and legal implications of our argument.

Keywords Social media · Content moderation · Public sphere · Platform governance · Free speech · Private property rights

1 Introduction

Increasing concerns around social media's impacts on political belief formation (Lewandowsky et al., 2019), extremism (West, 2021), radicalisation (Alfano et al., 2020), privacy violations (Sahebi & Formosa, 2022), lowering self-esteem (Cingel et al., 2022), and the formation of echo chambers (Terren & Borge-Bravo, 2021) have all lent increasing weight to recent efforts to better regulate the way that social media companies moderate content on their platforms. However, less attention has

✉ Bhanuraj Kashyap
bhanuraj.kashyap@mq.edu.au

Paul Formosa
paul.formosa@mq.edu.au

[1] Department of Philosophy, Macquarie University, Sydney, Australia

Published online: 13 December 2023

been given to the philosophical questions that underwrite these efforts. In particular, greater focus is needed on three foundational normative questions that social media moderation raises within the context of liberal democratic states. First, do social media companies have the *right sort of authority* to moderate user-generated content and speech on their platforms *at all*? Secondly, in so far as they do have that authority, what are the *limits* of their authority and what *responsibilities* do they have to moderate that content and speech? Thirdly, what are the *limits* of the *state's authority*, and what *responsibilities* does the state have, to regulate how social media companies moderate (or fail to moderate) content on their platforms?

The most prominent answer to the first two questions posed above involves an appeal to the social media companies' ownership rights over their own platforms (e.g., Cohen & Cohen, 2022). In Section 2, we outline what we mean by moderation, before exploring in Section 3 the extent to which social media companies' ownership of their own platforms ground their authority to moderate speech and content on their platforms. Next, in Section 4, we explore the limits of that authority by examining cases where social media companies' moderation practices and policies generate substantial user harms and instances of procedural injustice. Finally, in Section 5, we look at the state's responsibility to regulate social media companies to ensure that those harms and procedural injustices are limited or prevented.

This paper contributes to and furthers the literature on content moderation in two important ways. First, while many researchers in this literature allude to social media companies' commercial interests as a justification, or an explanation (or both), for their moderation of user-generated speech (see, for example, Gillespie, 2018; Suzor, 2019; Cobbe, 2021; Howard, 2021; Keller, 2021; and Cohen & Cohen, 2022), our paper innovatively develops and unpacks this relationship in detail from the standpoint of private property rights. In particular, we argue that it is through the prism of the powers granted by ownership of a platform that appeals to commercial interests can be made sense of as a possible justification for moderation practices. To this end, we offer a novel full-fledged account of *how* and *why* social media companies' ownership claims over their platforms ground their authority to moderate user-generated speech, the limits of this authorisation, and the normative upshots that should follow when social media companies violate those limits. Second, in providing our account, we not only strengthen the call for better content moderation systems, policies, and practices, but we also offer a novel argument against appeals to "commercial interests", which we ground in property rights, as a justification for unfettered control by social media companies over the governance of speech and communication in the digital world.

2 Social Media and Three Types of Moderation

Social media platforms, such as Facebook, YouTube, and TikTok, refer to digital intermediaries that allow their users to interact and communicate with one another through the sharing of user-generated content (Srnicek, 2017). Social media has billions of users across all regions of the world, and these users include individuals, groups, media outlets, local and international public bodies, as well as "bots"

(Gorwa & Guilbeault, 2020) and even terrorist organisations (West, 2021). User-generated content not only refers to traditional forms of speech involving written text, but also includes memes, audio, images, symbols, and videos, as well as comments and reactions (e.g., "Likes") to and reposting (e.g., "retweeting") of content from other users. In short, user-generated content takes many forms and includes engagement with the content of other users, although the types of content users can generate and the forms of engagement that they can undertake with that content tends to differ from platform to platform. There is another important form of content on social media, namely paid advertising content, that we will not focus on here, since advertisements have traditionally referred to *commercial* speech, which is distinct from other types of speech, and which has its own set of legal rules and regulations, although these vary from jurisdiction to jurisdiction (Johnson & Ho Youm, 2008). Of course, with the rise of "influencers" (Hudders et al., 2021) on social media, it is becoming increasingly difficult in practice to separate out commercial speech from other forms of speech on a platform (Bhagwat, 2019). Even so, we shall attempt to keep the two forms of speech conceptually distinct as much as possible.

User-generated content must then be shown to other users of a social media platform. This typically requires some form of content curation, as decisions must be made about what content is and is not shown, to whom it is shown, how it is shown, in what order it is shown, how often new content is shown, and so on. Due to the scale that most social media platforms operate at, content moderation typically depends, in part, on automation through the use of algorithms and Artificial Intelligence (AI) (Gillespie, 2020). Given the information sorting problems involved, content moderation can be understood as having three key basic processes: censorship, demotion, and amplification. These involve, respectively, hiding, decreasing (or burying), or increasing the visibility of content. We shall consider each in turn.

Censorship[1] by a platform refers to the removing, banning, or suspending of both user-generated content and user accounts from the platform. Censorship on social media typically relies on the following three apparatuses: a) moderation facilitated by human moderators in different countries who sift through content daily[2]; b) user-reporting mechanisms to report posts, such as hate speech, that may not meet a platform's published community standards (Lim & Ghadah, 2021); and c) algorithmic censorship of content that does not meet a platform's standards, such as scanning for images of naked bodies on platforms that do not allow such content (Cobbe, 2021). Closely related to censorship is user-level content blocking or silencing, whereby users choose not to see content from, for example, a particular user (Merten, 2021). Since

[1] The term "censorship" has recently been co-opted by the American right and is now used to push back against the left's socially progressive agenda, which the right unfavourably views as an attack on users' speech rights online (Srinivasan, 2023). We use the term "censorship", not in this way, but in the specific technical sense defined above that is common in the academic literature.

[2] Many moral issues regarding the exploitation of content moderation workers arise (see Roberts, 2019; and Barnes, 2022). It is also worth mentioning that some platforms, such as Reddit, give their users the authority to moderate content on forums (also called subreddits). These volunteer moderators are estimated to provide Reddit with $3.4 million worth of unpaid labour annually (Stokel-Walker, 2022), leading to demands for payment.

this amounts to user-generated moderation, we will not consider it further here given our focus on moderation by platforms.

Demotion refers to the down-rating of user content through recommendation features. This can also include practices such as “shadow banning” (also known as partial blocking) of user-generated content and users (Gorwa et al., 2020). For example, Instagram may bury content from accounts that buy followers or use hundreds of irrelevant hashtags without alerting holders of those accounts, unless those users directly follow those accounts (Savolainen, 2022). While the distinction between censorship and demotion is not always clear, we take the former to involve the removal or hiding of content, and the latter to involve the demotion or burying of content that remains available (to at least some users). Demotion mostly results from the algorithms that social media platforms use, which aim to amplify and demote content to achieve certain goals, such as increasing advertising income or maximising user engagement (Gillespie, 2022). However, demotion can also occur through user actions, such as downvoting posts on Reddit or disliking content on Facebook (Davis & Graham, 2021).

By contrast, amplification practices popularise content through recommendation features so that it reaches a wider audience. This includes amplification directly by the platform’s algorithms and due to the actions of users. For example, YouTube’s recommendation system recommends videos that will attract more views, Facebook’s Feed (formerly News Feed) is algorithmically personalised to each users’ preferences, and X (then known as Twitter) at one point algorithmically boosted tweets by Elon Musk to increase their reach (Schiffer & Newton, 2023). Amplification can also occur through user actions, such as using hashtags, Like buttons, up voting, and other recommendation boosts (Llansó et al., 2020).

One important point worth mentioning here is that large platforms *need* to moderate. This is because social media sites need some way to order, filter, update, and sort user-generated content, as well as some way to remove or bury problematic (for that platform) and illegal content. Further, given the scale of most social media platforms, algorithmic moderation is a necessary part of the moderation process (Gillespie, 2020). Even distributed platforms such as Mastodon that present posts in a chronological order from feeds that a user follows (Cobbe & Singh, 2019), rather than based on an algorithm designed, for example, to maximise user engagements, still depend on the use of hashtags to amplify content to other like-minded users, and the use of server level (rather than platform level) moderation practices, such as censoring content or banning users (Rozenshtein, 2022). However, the focus of our discussion will be on the algorithmic moderation of content by commercial platforms, rather than on either distributed moderation on decentralised platforms such as Mastodon or on individual user-level moderation through reporting, upvoting and downvoting content, as this is in line with the scholarship’s focus on algorithmic mediation employed by Big Tech on internet platforms.

3 The Authority to Moderate

3.1 The Authority to Moderate User-Generated Content

Now that we are clear what moderation of user-generated content is, we can ask what the normative ground of social media platforms' *authority* to moderate that content is. This question has, surprisingly, received comparatively little attention. A recent paper by Cohen and Cohen (2022) attempts to address this gap through a focus on the moral permissibility of non-state censorship that is based on the private property rights over their platforms that social media companies possess. In this section, we explore the argument that private property rights can ground the right of platform owners to moderate user-generated content on their platforms.

Cohen and Cohen (2022, 16–17) understand private property rights as including a bundle of claims, including the right to exclude, as necessary features of ownership. The inclusion of a right to exclude is needed to draw a distinction between a liberty and a right (Schmidtz, 2010, 79–80). For example, when someone says they own their house, this means that they are not only at liberty to live in their house and use it, but others are also morally required not to occupy and use their house against their wishes. Here a property right carries a liberty to use the property, plus a duty for others that they refrain from using this property without proper consent (Schmidtz, 2010, 80), except (arguably) in special circumstances such as an emergency (Cohen & Cohen, 2022, 28). For example, if someone is fleeing a gunman and they need a place to hide, it is (arguably) permissible for them to enter and hide in someone else's house, without the owner's explicit permission (see Schmidtz, 2010; and Honoré, 2013). The right to exclude others from private property is thus an important element in what it means to *own* something.

According to Cohen and Cohen (2022), social media companies are private nonstate actors that own their digital platforms. Their ownership over their platforms thereby confers on them "stringent private rights to exclude", and this in turn gives them "rights to exclude others who seek access to their property, including for the purposes of expression or communication" (Cohen & Cohen, 2022, 17). To continue with the previous example, if a white supremacist decides to enter someone's house without his permission to talk to him about white racial superiority, he can ask them to leave. In refusing to let them use his property to express their views, he may be censoring them in some sense, but this type of censorship is permissible since it is consistent with his right to exclude others from his private property. Similarly, so the argument goes, it is permissible for a social media company to delete user-generated content or to ban a user from having access to their platform, such as Twitter banning former US President Trump from using its platform (Dwoskin & Tiku, 2022), on the same private property ownership grounds, even if this amounts to censorship (in some sense).

While Cohen and Cohen's account explicitly focuses only on "censorship" by social media companies, we shall now argue that their account can be extended to also cover content demotion and amplification, and thus all three elements of moderation. This addition is needed, since the right to exclude that Cohen and Cohen focus on does not directly (without further argument) ground a right to amplify or demote content, as opposed to a right to censor or ban content. To expand their account, we shall draw on accounts of private property rights developed by Honoré's (2013) and Katz (2008) and their associated ideas of a right to use one's property to generate income and the authority to determine the agenda for a resource. This helps us to further spell out the rights and powers that ownership grants.

According to Honoré (2013), to whom Cohen and Cohen (2022, 16) refer when developing their own account, ownership is underpinned by a range of rights and duties, including the right to use and the right to manage one's owned resource. The right to use entails that the owner can enjoy their property in multiple ways, including using it to generate income. This aspect of a property right is helpful in grounding moderation rights across all three elements. In particular, different features of amplification are important means for platforms to attract advertisers to make money off their users (Gillespie, 2018). For example, Facebook introduced reactions to give its users more options to express themselves with reference to user-generated content. The posts that receive more reactions are, in turn, amplified to attract more users, thereby increasing user-engagement duration online (Vaidhyanathan, 2018, 35–37). Similarly, platforms such as YouTube have recommendation systems to suggest content to its users to increase user-engagement so that the platform can serve more advertisements to its users, thereby generating income (Alfano et al., 2020). Further, demoting or banning content that advertisers may not wish to have associated with their brand is also an important part of this strategy (Cobbe, 2021, 752–757).

Another aspect of private property rights is what Katz (2008) calls the "agenda-setting authority" that gives owners the position of "the exclusive agenda setter for the owned thing" (Katz, 2008, 275) or resource. Agenda-setter authority refers to the authority that owners exclusively possess to determine how an owned resource is to be used. If we apply this to social media companies, it seems to grant them a right to determine how their platforms are used through their authority to set the agenda for their owned resource. Moderating content, through censoring, amplifying, or demoting user-generated content, seems to fall under this broad agenda-setting authority of a platform owner. This authority also extends to setting the terms of services on platforms which can outline the sort of content, or community standards, that is acceptable on that platform, as well as further rights (if any) that users may have in terms of appealing moderation practices on that platform. These same terms of service may also be used to outline data mining activities, and other related tracking activities, undertaken by social media companies of both user-generated content and user activity on their platforms for commercial purposes (Magarian, 2021, 355–356).

As platform owners, social media companies have the right to exclude others, use their platforms to generate profit, and set the agenda for how their platform is used. This allows us to move from the property rights that social media companies have as owners of their platforms, to their authority to moderate, including censoring, amplifying, and demoting, user-generated content on their platforms. If users do not like a platform's moderation policies or practices, then they are free to leave the service.

This argument's focus on private property rights would seem to grant social media companies the moral right and practical authority to wield an incredible amount of power over their users. But, given their immense power over users, the large role that social media plays in the functioning of democracies (Andrejevic, 2020, 44–72), and the quasi-monopoly of social media companies that exist due to the difficulty of overcoming network effects that favour incumbent platforms (Srnicek, 2017), we might wonder (as Cohen and Cohen's argument requires) whether it makes sense to directly apply an account of *private* property rights to social media platforms. After all, it is one thing to own a small block of land, and another thing altogether to own the means of production of significant social entities, such as social media platforms. Should we instead think of these platforms as more akin to public utilities than private property? If so, how does this impact the above argument about the platforms' authority to moderate content on their platform?

Even putting this point aside for a moment, it is worth noting that the conception of private property ownership borrowed here from Honoré and Katz does place *some* limits on the owner's authority to use and manage their property. In the case of monopolies over key aspects of the means of production and other components that have critical impacts on the functioning of a democratic society, such as newspapers and advertisements in the media, Katz acknowledges that the state possesses "the power to regulate the use of resources and to control their allocation" (2008, 295). This means that private owners of such entities, which would seem to include owners of large social media platforms, have a public justification requirement to meet to keep enjoying the state's endorsement of their agenda-setting authority over their owned resources relative to non-owner users. Further, their authority to moderate content on their platforms can be regulated and controlled by the state. Another relevant limitation, this time from Honoré's account (2013), is the prohibition of harmful use. Honoré agrees that the owner is not allowed to do whatever they please with their owned property if it will result in harm to other members of the society. For example, one may own the knife in one's kitchen, but that does not give one the right to use it to stab someone in the back. Thus, although social media companies own their platforms, we believe that social media companies' ownership rights do not exempt them from the moral requirement not to harm others when using their owned resource. In the next section, we explore further the limitations to the moderation powers that private property rights can authorise.

4 Limitations to the Authority to Moderate

Drawing on the previous section, we have identified at least two key cases where it is appropriate to interfere with private property rights. The first case focuses on establishing the presence of monopolies over key resources. While we think that the first route via regulation of monopolies is open to critics of unfettered moderation powers of social media companies, we shall not pursue this line of argument further here since it has been dealt with in detail elsewhere (see, e.g., Ranttila, 2020; Alfano & Sullivan, 2022; and Balkin, 2022). Even though we do not directly address the case of social media monopolies, it is worth stressing that the monopoly status of social media platforms does significantly contribute to some of the injustices and harms users experience resulting from the moderation policies and practices of platforms, as we will discuss in detail below. Furthermore, the monopoly status of social media platforms establishes that a viable exit strategy from social media does not exist for many users and, if such an exit is nonetheless taken, it can be extremely costly to both a person's prospects and social connectivity (Flew & Wilding, 2020). However, even presuming that an individual user can leave a social media platform without any *direct* adverse personal outcomes does not, as we will show below, mean that she can also escape the large-scale effects of social media moderation on others, including the broader democratic society and culture of which she is part, that may still harm her in *indirect* yet important ways even though she does not use any platform.

Instead, we focus here on the second case that looks at the harms and injustices that are inflicted on others by the unfettered ownership of property. For this line of argument, whether the harm inflictor also has a monopoly over a key resource is often relevant but not (in all cases) essential. Further, we consider this focus on harms and injustices to be complementary, rather than a competitor, to the alternative focus on whether social media platforms constitute monopolies. We pursue this issue below by considering first substantive harms and second procedural injustices, as those are the two types of primary concerns that are raised in the legal and media studies literature on platform content moderation (see Bollinger & Stone, 2022). While this discussion has clear links with the broader literature on Mill's harm principle (for discussion see, e.g., Turner, 2014), we will not be exploring those links as this literature raises larger issues that are beyond our direct focus. However, future work could explore these links in more detail.

4.1 Substantive Harms

We shall document here some of the harms of social media. While there are also various benefits of social media, such as improved means of social connection, we shall not attempt to outline these here since these do not typically raise regulatory concerns or give us reasons to limit the use of private property. While, of course, we need to consider the relationship between the harms and benefits caused by social media, if the harms are significant enough or of the right form, the fact that there may also be benefits may not matter. For example, the fact that slavery may have the benefit of providing low-cost goods or giving sufficient time to slave owners to engage in meaningful political activity does not matter given the significant injustices and harms involved in slavery. Something similar, we suggest, might be the case with the harms caused by social media, although that argument will need to be explicitly made. Further, it may not be a case of weighing up the harms and benefits of social media, since it may be possible to keep the benefits (e.g., constructive communication with others) while also minimising or eliminating the harms (e.g., by limiting the amplification of harmful content for financial gain).

Given our focus on the issue of platform moderation of content, we concentrate on large-scale patterns of harm that have emerged at a societal level. This means that we do not focus on individual instances of harm perpetrated on social media, although these are of course significant for the people involved, but rather on the broader patterns of harm that emerge in the aggregate. Further, it is important to note that harms can be caused by all three components—censorship, demotion, and amplification—that constitute moderation on our account. For example, the harms associated with moderation may sometimes be caused by ineffective censorship or demotion practices by platforms, such as failing to consistently and accurately remove or demote hate speech against users, and sometimes by platform's amplification practices, such as its amplifying of hate speech content.

We now turn to both empirical and theoretical evidence to support our claims about the harms of social media. Research shows that social media companies' focus on increasing user engagement tends to: undermine democratic politics (Forestal, 2020; Vaidhyanathan, 2018); increase political distrust and intergroup conflict (Hong & Kim, 2016; Alfano & Sullivan, 2022); lead to the formation of epistemic bubbles and homogenisation of important information networks (Nguyen, 2018); and lead to the proliferation of violent and extremist content and conspiracy theories (Klein et al., 2018, 2019; Alfano et al., 2018, 2020; and Hao, 2021). Simultaneously, certain social media platforms have become a breeding ground for hate speech, which is in and of itself harmful (see, for example, Waldron, 2012), and which is often directed at women, members of the LGBTQI + community, and racial minorities (Are, 2020; Suzor, 2019). The largely opaque algorithmic moderation of user-generated content can also undermine and disrespect user autonomy (Sahebi & Formosa, 2022), lead to widespread and serious mental health issues in vulnerable populations, especially young people (Klein, 2023), and promote the circulation of misinformation (Barnes, 2022) and magnify affective polarisation (Cho et al., 2020).

One might, however, object that the cited evidence is biased against social media companies and does not accurately describe the impact of social media on political distrust, polarisation, and democratic politics. This objection would state that contrary evidence persists, and it is not, at this stage, possible to *conclusively* state that social media companies' moderation of speech leads to substantive harms (see, for example, Newman et al., 2017; Benkler et al., 2018; Silver & Huang, 2019; Barberá, 2020; and Boxell et al., 2021). Something similar applies to the research on social media's negative impacts on youth mental health, where the evidence is again strong but not conclusive (for evidence in favour of the existence of this impact, see Klein, 2023 and Harriger et al., 2023; for counter evidence, see Orben & Przybylski, 2019 and Vuorre & Przybylski, 2023).

In response to this objection, several points are worth making.

First, as Lazar (2023) notes, the acontextual empirical approach that researchers often use to measure polarisation can paint an inaccurate empirical picture. For example, it may be true that traditional media primarily drove election-related disinformation in the US during the 2020 Presidential election cycle (Benkler et al., 2020). However, this piece of information paints an empirically incomplete picture since it does not grapple with how social media companies have shifted the media ecology towards more sensationalistic content, often forcing traditional media to compete with right-wing digital media outlets such as Breitbart (Lazar, 2023). Second, empirical research often focuses on the extent to which types of information is available and accessible online, as opposed to understanding how platforms enable and encourage social relationships (see, for example, Vaidhyanathan, 2018, and Settle, 2018). For example, online users may have access to more diverse information than before (Newman et al., 2017), but this cannot be interpreted as unconditionally positive if those users come across diverse information in hostile social networks, such as in comments sections where they often engage in abusive arguments with one another. In terms of the mental health harms of social media, there are clearly many complex intersecting elements that are also at play, which makes isolating the role of social media alone empirically difficult. It is therefore appropriate to concede that establishing *conclusive* evidence about the causation of social media on democratic politics, political distrust, and youth mental health is indeed an exceedingly complex task. Consequently, it is unsurprising that empirical research on these issues is still in flux. However, unless the evidence drastically changes to demonstrate that our approach is thoroughly unfounded, the quoted evidence is arguably sufficient to motivate our present concerns by supporting the claim that social media companies' regulation of speech is likely to result in harmful outcomes that we need to take seriously now (even in the absence of conclusive evidence). Combined with our earlier claim that causing significant harms to others is a presumptive ground for limiting or regulating the use of private property, we get what we call here "the limiting condition of substantive harms":

> When the moderation practices and policies of social media platforms lead to substantive harms to users and societies, then this creates a presumptive ground for limiting the private property-based authority to moderate user-generated content that owners of social media platforms otherwise possess.

Private property rights do not grant unfettered rights to utilise owned resources in ways that evidence shows creates significant harms.

Two prima facie duties follow on from this limiting condition. First, in light of the harms they produce, social media companies have an additional duty to justify their continuing authority to moderate user-generated content on their platforms. But, as argued above, content needs to be moderated, since information needs to be ordered, presented, and updated. With this in mind, we can understand social media companies as providing important moderating services (Gillespie, 2018). If the public wants the content on platforms to be properly moderated, then the owners of those platforms are best placed to provide those moderation services, whether that occurs directly through algorithmic moderation run by platforms or more indirectly through platforms providing the infrastructure and oversight for users to have greater input into moderation on the platform. Thus, the additional justification requirement of the first duty can, arguably, be met on grounds of practicality. Second, social media platforms should moderate content in ways that try to limit or eliminate the substantive harms caused by moderation. As we have argued, the empirical evidence suggests that social media companies are failing to uphold this obligation to both their users and the broader public. Having made our positive argument here, we now consider some objections.

One objection that is often made against attributing responsibility to social media companies for online harms emerges from within the 'no liability' model (Howard, 2021). According to this model, social media companies are not responsible for the actions and speech of its users. If a harm is produced because of how P has acted on platform Q, it is P's actions that need to be held responsible for the harm produced, as opposed to holding Q responsible for permitting P's actions on its platform. Q letting a harm occur on its platform is conceptually and morally distinct from P's direct act of perpetrating harm against another.

While this objection may seem initially plausible, the many reasons for rejecting the 'no liability' model are covered extensively in the media studies literature on this topic. These include everything from the practical difficulties of tracking and holding individuals responsible for harmful actions and speech perpetrated online, to the misconception that platforms are neutral towards content (for further discussion see, e.g., Gillespie, 2018 and Andrejevic, 2020). But the most relevant reason for rejecting this model is that it rests on the mistaken view that digital intermediaries, such as Facebook and Twitter (now X), do not make causal contributions to how their users engage with one another on their platforms (Forestal, 2020 and 2022). Platform moderation and architecture plays a causally important and distinctive role in laying out the ways in which communication and engagement between its users start, continue, end, and restart. The 'no liability' model tends to be overfocused on questions of permissibility regarding the accessibility of extreme and violent content to a platform's users (Lazar, 2023). But this ignores questions of the distribution of extreme content, hate speech, misinformation, and conspiracy theories on platforms, and the ways that platform architectures and moderation policies both encourage the creation and causally facilitate the spread and amplification of such content (Andrejevic & Volvic, 2020; and Lazar, 2023). Platforms do not merely provide a neutral space for user engagement with one another; rather, they provide an entire social ecosystem that not merely enables, but also encourages activities on its platforms that are harmful in the ways outlined above. This suggests, in opposition to the 'no liability' model's focus on individual responsibility *only*, an alternative 'engagement' model which *additionally* focuses on intermediary responsibilities attributable to digital platforms when patterns of user and broader large-scale harms are produced, reinforced, and maintained due, in part, to those platform's architecture and moderation practices (Lazar, 2023). It is only by ignoring this 'engagement' model that social media platforms can operate under the assumption that they do not have any type of *intermediary responsibility* towards their users.

This notion of intermediary responsibility lends some moral force to, for example, the European Union's regulatory model of 'notice-and-takedown', where hate speech is legally bound to be taken down by social media companies when they become aware of it (Digital Services Act, 2022). While we believe that this is a good first step, the European Union's regulatory model may result in a host of related issues, from poor identification of hate or illegal speech to overenforcement of rules to censor speech (Keller, 2021), as we discuss further below. It also focuses on types of direct speech harms, such as racial abuse, but fails to consider the broader range of other large-scale harms, such as poor mental health in children, that emerges as a network effect of the amplification of content on platforms, such as promoting an unrealistic body image, that in isolation may be morally innocuous but at scale cause widespread harm. However, whatever legal regulatory framework states decide to pursue to ensure social media companies refrain from causally contributing to substantive harms, states will still have to strike a balance between users' rights to free speech and social media companies' ownership right to curate and disseminate content, including extreme speech, for commercial purposes. For example, if social media companies are legally obliged to filter all user-generated content before allowing it to become visible and distributed on their platforms, this could have significant and unintended negative consequences for user engagement and communication.

Finally, it is worth adding that holding individual users responsible for directly perpetrating harm against other users on a platform is not incompatible with also holding social media platforms responsible for acting as an intermediary of such harms by amplifying or not blocking such content. However, one could object that social media companies may not be aware of these harms occurring, and thus not count as bearing any intermediary responsibility for them given this lack of knowledge. But this objection is unfounded as ample evidence shows that social media companies are aware of these harms and have often done too little to adequately respond to them (Vaidhyanathan, 2018). For example, Frances Haugen's revelations show that Facebook was aware of the negative impacts of Instagram on the mental health of teenage girls driven by its engagement-based moderation practices (Milmo, 2021). Further, some of the strategies that social media companies have introduced to squash the circulation and amplification of fake news and misinformation is arguably too weak, late on the scene, and has not produced the desired results (Collins et al., 2021). It is therefore reasonable to claim that social media platforms are, by now, knowingly complicit in these unjustified and large-scale patterns of harms, and thus meet any relevant epistemic requirements for intermediary responsibility.

4.2 Procedural Injustice

Social media platforms publish community guidelines that lay out standards regarding user-generated content that those platforms consider permissible or impermissible (Gillespie, 2018). All the major social media companies—Facebook, YouTube, and Instagram, for instance—are committed to upholding similar community standards across their platforms (45–47; 52–62). For example, all platforms prohibit child pornography and almost all platforms[3] discourage content gesturing to self-harm, violence against others, and hate speech, though they implement these standards and rules differently, depending on their platform architecture and design. Many platforms, such as Facebook, also offer reasons and justifications for classifying different types of content as appropriate or inappropriate (Gillespie, 2018, 47–52).

It is important to note that community guidelines are partner documents to the formal legal terms and conditions that users agree to when signing up to a social media platform. While the primary aim of terms and conditions documents is to ensure that social media companies are protected against potential future litigation, community guidelines are a *performative endorsement* of a platform's values

[3] The social media platform Gab may be an exception to this observation (see Stanford Internet Observatory, 2022).

and goals that is used to guide, manage, and curate its users' speech (Gillespie, 2018, 47–52). Further, it is to these community guidelines that users often turn to in cases of disputes, concerns, and confusions regarding platform moderation outcomes.

Community guidelines help a platform to strike a balance between respecting its users' communicative interests and executing its moderation services. By communicative interests, we mean the interests a person has in expressing their own views to another person (or persons) and having that other person (or persons) pay attention to those expressions (Cohen, 1993; Bejan, 2020; Cohen & Fung, 2021; and Lazar, 2023). While we do not deny that communicative interests can have instrumental grounds (where like-minded individuals aim to get together and achieve desired outcomes, for example, in setting up a contract or organising an event), we will focus on the non-instrumental justification of an individual user's communicative interests. To be able to speak and express your views to other people and be listened to, genuinely heard, and responded to captures an important, non-instrumental dimension of relational equality and belonging to a moral community (Lazar, 2023). It is part of being taken seriously as a person with dignity who deserves respect and consideration (Formosa, 2017). An individual's communicative interests, as an expression of their dignity, ought to be respected and valued, unless doing so is incompatible with respecting the dignity and worth of other agents, such as when executing one agent's communicative interests would result in harm to other agents either interpersonally or by contributing to a societal harm. Users refer to community guidelines to learn about the publicly available rules and procedures that platforms employ to censor and demote content (Gillespie, 2018). So long as users abide by this document when creating content, the assumption is that a platform will not arbitrarily interfere with their speech by either censoring it or suspending them from the platform for exercising it. In this way, social media companies can respect their users' communicative interests (see, for example, Facebook Community Standards, 2023) in the digital public sphere.

However, in practice platforms often do arbitrarily interfere with their users' speech. We can understand arbitrary interference here as a form of *procedural injustice* that refers to instances where platforms unreasonably fail to treat similar cases of users' speech similarly and different cases of user's speech differently. For example, imagine that two users post relevantly similar content on a platform that seems to comply with that platform's published community standards. However, one user's post is taken down (i.e., censored) without a proper explanation and the user is issued with a warning, whereas the other user's post is not only not censored but is rather amplified to others. Instances that fit this example are, unfortunately, not uncommon on social media platforms (see Gillespie, 2018 for examples), often leaving users deeply unhappy with the moderation policies of a platform (West, 2018; and Cook et al. 2021). While options to challenge moderation outcomes exist on some platforms, these options are slow, confusing, not user friendly, and difficult to navigate (Everett, 2018).

In moderating arbitrarily, platforms act in a procedurally unjust way and express disrespect for their users by failing to give proper weight to their communicative interests. Combined with our earlier claim that inflicting injustices on others is also a presumptive ground for limiting or regulating the use of private property, we get what we call here “the condition of arbitrary platform interference in users’ speech”:

> When the moderation of user-generated content by social media platforms unreasonably fails to treat similar cases of users’ speech similarly and different cases of users’ speech differently, this amounts to arbitrary interference in users’ communicative interests and should be considered a procedural injustice. This creates a presumptive ground for limiting the private property-based authority to moderate user-generated content that owners of social media platforms otherwise possess.

We now consider and respond to some objections to this claim.

First, it is not clear why a user’s communicative interests should override social media platform’s ownership rights to moderation. According to Cohen and Cohen (2022, 18), to argue for this claim we need to show that there are no genuine substitutes available for users to express their communicative interests. Cohen and Cohen call this the “substitution objection”. Implicit in this objection is the assumption the communicative interests of users do not need to be respected on a platform when genuine substitutes for it are available. However, we believe that genuine substitutes are unavailable since social media platforms constitute a digital public sphere (Cohen & Fung, 2021; Lazar, 2023) that acts as an extension (or another part of) the broader public sphere. Empirical evidence suggests that social media companies host and curate different types of political viewpoints, from COVID-19 vaccine (mis)information to discussions of political officials, elections, and public policies (Bollinger & Stone, 2022). In the last decade, it has become evident that social media companies are an incredibly important resource for social movements to begin and spread (Tufekci, 2017). Furthermore, the entities that are often associated with upholding the traditional public sphere, such as media houses and newspapers, typically rely on social media platforms for distributing and amplifying their content. Research indicates that the advent of social media has unprecedently altered the traditional news, opinion, and information landscape (Andrejevic & Volvic, 2020), in ways that may be irreversible. Given that traditional news platforms and organisations have long been understood as helping to form part of the public sphere even though many of them are commercial entities, it follows that social media platforms can also form part of the public sphere even if they engage in commercial activities (Andrejevic, 2020, 45–72). It is the social and political role that matters, and given the above evidence, it seems clear that social media platforms, among other things, do in fact host and curate an important part of the digital public sphere. As a public sphere, users who participate on social media platforms to express themselves to others have strong communicative interests that ought not to be interfered with arbitrarily. This does not mean that there are no *alternatives* to social media, since clearly there are, such as talking directly to friends or engaging in a public debate in-person. Rather, the point is that these alternatives are not genuine *substitutes* for the crucial functional role that social media plays in helping to constitute the broader public sphere.

The second objection that arises is an extension of the first objection. Even if we presume that social media companies provide the services of hosting and curating the digital public sphere, it is not clear whether they owe individual users *procedural justice*. This objection states that the ownership rights to moderation that platforms possess overrides users' demand for procedural justice (see Cohen & Cohen, 2022, 17–23, where this is called the "town square objection"). To see the force of this objection, consider a related example. Imagine that a powerful individual invites several people to her house to discuss different political viewpoints, but she decides to arbitrarily exclude some potential guests in a procedurally unjust way (e.g., she randomly excludes half the potential guests with brown hair). Many of her guests may feel genuinely wronged by her choices and believe that her way of choosing guests is procedurally unfair, but they also reason that since she possesses ownership rights to her house, she gets to decide the types of guests that are given permission to attend and express themselves at her private party. If they do not like it, the invited guests can always choose not to attend the party or attend some other party instead.

But a social media platform is not anything like a small, private social gathering. Unlike an individual owner of a house, social media companies are quasi-monopolies that own significant infrastructure and digital entities that collectively host and curate the digital public sphere and their published community guidelines play a role in helping users navigate the way that they exercise their communicative interests online. When those community guidelines are applied in an arbitrary manner, they violate the like cases maxim—generally considered an important maxim for any robust account of procedural justice (Zimmermann & Lee-Stronach, 2022)—and this leaves its users unable to navigate platforms that are essential to helping them realise their communicative interests. This could, in turn, leave them potentially alienated from the public sphere, or at least from key parts of the digital component of the public sphere. Unless social media companies fix this problem and make reliable options available for recourse, we should not empower private institutions such as these with unfettered legitimacy to moderate user-generated content.

Finally, note that it is possible to separate our claim about respecting users' communicative interests from our claim about users experiencing procedural injustice when content moderation rules are arbitrarily applied to their speech. It may be the case that rules are applied consistently and correctly, but they still end up interfering with some users' communicative interests for unjustified reasons, thereby disrespecting those users. For example, photographs of breastfeeding women were heavily censored and removed by Facebook since Facebook took those photographs to violate norms around public nudity and thereby its own community standards. This (arguably) simply reinforced patriarchal norms that control displays of women's bodies in public which are harmful to women (Gillespie, 2018), as well as wrongly disrespect their communicative interests. This points to the broader issue that beyond the procedural injustices that we focus on in this section, there are also cases of moderation, such as the above breastfeeding example, which are substantively unjust without also being procedurally unjust as they involve the fair and consistent application of a rule that is itself unjust. While we agree that such cases are clearly significant, for the purposes of our paper they are best addressed through our first limiting condition of substantive harm, since such moderation practices are substantively harmful (and thereby potentially substantively unjust for that very reason), even if they are not also procedurally unjust. Of course, this also leaves a category of case where platform moderation is both procedurally *and* substantively just, such as the fair and consistent application by platforms of publicly available rules requiring the removal of certain forms of very harmful content.

This completes our section on the two types of limitations—the limiting condition of substantive harms and the limiting condition of procedural injustice—that we have argued curtail the social media companies' private property-based authority to moderate user-generated content. We will now turn to the question of the responsibility of the state to regulate social media companies to ensure that these harms and procedural injustices are limited or prevented.

5 What should States do about it?

Two questions need to be addressed in this section. First, *can* the state override social media companies' unfettered authority, based in their private property rights, to moderate speech on their platforms as they see fit, given their violations of the two limiting conditions that we defended in the previous section? Second, *should* the state *in fact* override this authority to regulate how platforms moderate online content? The first question focuses on *permissibility*, and the second on political *action*. In this section, we will briefly explore both these questions.

A liberal democratic state's obligation towards its citizens is clear: it can rightfully intervene to prevent harm and injustice from falling upon its citizens (e.g., Mill, 2001 and Heinze, 2016). If a citizen decides to engage in an activity that harms others, such as murder, assault, or robbery, the state is permitted to step in to prevent that citizen from engaging in their desired activity to protect others. This abrogation of their freedom of activity to ensure others are not harmed is, therefore, an instance of justified political power (Fabienne, 2017). While our analysis is restricted to a liberal democratic framework, other political frameworks also consider such interventions by the state to be pro tanto permissible. For example, in the neo-republican framework advanced by Pettit, 2002; 2013), if private entities engage (or have the ability to engage) in acts of private domination, where they can arbitrarily interfere with the choices available to a state's citizens, the state is obliged to use its coercive powers to regulate those private actors.

Many activities and items can be harmful (for example, regularly eating junk food) or pose significant risks to others (for example, driving a car in the presence of other cars and pedestrians). While the regulation of activities that are only harmful to oneself has a long history of contestation (e.g., see the discussion of seatbelt laws for adults in Formosa & Mackenzie, 2014, 890 and Nussbaum, 2000, 95), it is generally accepted that the state may regulate products and activities that are harmful to others. For example, speeding and drink driving laws, and safety standards and ratings, help to minimise the risks to others of driving cars. Permissible restrictions on commercial speech are also often made to ensure that harmful products (for example, cigarettes) cannot be advertised to consumers (Tuchman, 2019) or must come with clear warnings (for example, gambling advertising—see Gainsbury et al., 2016). Something similar, we hold, should apply to social media moderation practices, insofar as they risk significant harms to others, especially to members of vulnerable groups such as children and minorities (Mackenzie et al., 2014), or involve perpetrating injustices against others, whether substantive or procedural. The fact that social media platforms are privately owned properties does not, we argue above, override these imperatives. This entails that the state *can* intervene to regulate how social media companies moderate speech on their platforms. Our arguments, therefore, demonstrate that the commercial interests of platforms, grounded in private property rights, do not offer overriding reasons in favour of unfettered platform content moderation practices, and offers a conceptual foundation for the permissibility of state intervention in this regard.

However, this still leaves open the practical question of whether the state *should* intervene, given that it is permissible for it to do so. While our arguments set out a normative foundation for this further discussion which is beyond our direct scope to resolve conclusively here, there are several points worth noting that follow from our discussion. The question of state intervention is complicated because there are other rights and interests at stake, and it is not clear whether state intervention to moderate user speech on platforms will always respect these other rights (for example, rights to speech and association). Further, it is unclear if state interventions will have unintended negative consequences and will actually be effective in achieving the aims of the intervention given the dynamic and global nature of social media that eludes the control of any single jurisdiction.

For the state to deal with substantive harms, it will have to decide the limits of permissible speech in the digital public sphere, which often invokes free speech concerns. The democratic right to free speech underpins viewpoint neutrality as a constitutive feature of a democratic society, since in a democracy people need “free speech in order to engage in the enterprise of self-government" (Howard, 2019, 98). By viewpoint neutrality, we mean that speech is not targeted based on its content and the message it conveys. The literature on free speech is highly controversial, with many commentators arguing for some types of censorship and regulation of hate speech (e.g., Waldron, 2012; and Delgado & Stefancic, 2018), whereas other commentators argue against viewpoint-based restrictions on speech (see, for example, Post, 2011; and Heinze, 2016) since they worry that viewpoint-based restrictions risk flirting with democratic backsliding, as has arguably been seen in countries such as India. Any viable model of speech regulation must determine how users' right to speech can be respected while simultaneously legally requiring platforms to censor and remove speech that may be deemed harmful or unjust.

The liability model—holding platforms' responsible for hosting and amplifying illegal speech, or legal but harmful speech, or both these types of speech—has recently gained momentum in Europe. Many European countries have written and implemented laws that have given them unprecedented power over moderating content on platforms such as Facebook and YouTube (see Digital Services Act, 2022). However, these new laws often do not make available avenues to individual users for recourse in cases of decision disputes, thereby raising the moral cost of enforcing these speech laws (Keller, 2021). Without recourse options, over-enforcement of these laws restricting speech by platforms seeking to avoid potential liabilities could further violate the communicative interests of their users. As the liability model is rolled out over the next few years, more empirical and theoretical research will have to be carried out to help better assess the efficacy of this model at dealing with substantive harms.

Many commentators think that value-neutral content moderation policies might be the best way forward here (see, for example, Keller, 2021; Balkin, 2022; and Benkler, 2022). It is permissible (and sometimes obligatory) for states to enforce value-neutral time, place, and manner restrictions on speech; for example, fining the proprietor of a bar playing loud music (regardless of its content) after a certain time in the evening in a residential neighbourhood. Similarly, content-neutral policies, some commentators contend, should be used to moderate content on social media platforms. This content neutral approach to moderation lends weight to our claim that states should require social media companies to apply their community standards fairly across all user-generated content, treating similar instances of speech similarly and different instances of speech differently. This will not only protect users against procedural injustice, but it will also help to respect their communicative interests.

It is worth noting that some practical difficulties might arise with both these value-based and value-neutral approaches to content moderation, especially related to the problem of scale. Social media companies moderate an inordinately large amount of online speech, making it practically impossible for them to avoid making numerous mistakes (Gillespie, 2018). Given this, it is an open question how we can best strike a balance between respecting each user's communicative interests and accepting the reasonableness of some mistakes being made by a moderation system enforcing community guidelines on platforms at scale (for discussion of this point, see Douek, 2021). While responding to this problem of scale is outside our direct scope, if social media companies act with transparency and in good faith and give justifiable reasons for their decisions (and mistakes), punitive actions in terms of fines may not be reasonable. However, part of responding in good faith and ensuring procedural justice will involve setting up systems by which moderation decisions made by platforms can be appealed by users and addressed in a reasonable timeframe. While demands to consistently apply policies and require recourse and explanation would incur costs for platforms and (potentially) their users, given the importance of the issues at stake, the imposition of these costs seems reasonable.

6 Conclusion

In this paper we have focused on exploring the claim that private ownership rights can ground the unfettered right of social media platforms to moderate user-generated content on their platforms. We argue that there are at least two cases where ownership rights to moderate content can be overridden, and that is when the moderation results in significant large-scale patterns of harms and when the moderation practices perpetuate procedural injustices. There may be other cases too, such as platforms constituting monopolies, which we do not explore in detail here beyond noting that these cases complement and strengthen the force of our overall argument. We then argue that social media platforms inflict substantive harms and perpetuate procedural injustices, which means that in those cases we have grounds for overriding the ownership rights of platforms to moderate content through regulation. Finally, having made the normative case for the permissibility of state regulation of moderation practices on social media platforms, we briefly considered some of the practical implications that arise from our argument in terms of its implementation.

Authors' Contributions BK and PF have both made substantial contributions to this manuscript.

Funding None.

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not Applicable.

Consent for Publication Consent from authors obtained.

Competing Interests PF has received funds for an unrelated project from Meta and Facebook.

References

Alfano, M., Carter, J. A., & Cheong, M. (2018). Technological seduction and self-radicalization. *Journal of the American Philosophical Association, 4*(3), 298–322.

Alfano, M., Fard, A. E., Carter, J. A., Clutton, P., & Klein, C. (2020). Technologically scaffolded atypical cognition: The case of youtube's recommender system. *Synthese, 199*, 835–858.

Alfano, M. & Sullivan, E. (2022). Online Trust and Distrust. In Hannon and Ridder J.D. (Eds), *The Routledge Handbook of Political Epistemology* (pp. 480–491). Routledge.

Andrejevic, M. (2020). *Automated Media*. Routledge.

Andrejevic, M., & Volvic, Z. (2020). From Mass to Automated Media. In N. Witzleb, M. Paterson, & J. Richardson (Eds.), *Big Data, Political Campaigning and the Law* (pp. 17–33). Routledge.

Are, C. (2020). How Instagram's Algorithm is Censoring Women and Vulnerable Users But Helping Online Abusers. *Feminist Media Studies, 20*(5), 741–744.

Balkin, J. M. (2022). To Reform Social Media, Reform Informational Capitalism. In L. Bollinger & G. R. Stone (Eds.), *Social Media, Freedom of Speech and the Future of Our Democracy* (pp. 233–254). Oxford Publishing Press.

Barberá, P. (2020). Social Media, Echo Chambers, and Political Polarization. In N. Persily & J. A. Tucker (Eds.), *Social Media and Democracy: The State of the Field, Prospects for Reform* (pp. 34–55). Cambridge University Press.

Barnes, R.M. (2022). Online extremism, AI, and (human) content moderation. *Feminist Philosophy Quarterly*, *8*(3/4). Article 6.

Bejan, T. M. (2020). Free expression or equal speech? *Social Philosophy and Policy Foundation, 37*(2), 153–169.

Benkler, Y. (2022). Follow the Money, Back to Front. In L. Bollinger & G. R. Stone (Eds.), *Social Media, Freedom of Speech and the Future of Our Democracy* (pp. 255–272). Oxford Publishing Press.

Benkler, Y., Farris, R., & Roberts, H. (2018). *Network Propaganda*. Oxford Publishing Press.

Benkler, Y., Tilton, C., Etling, B., Roberts, H., Clark, J., Faris, R., Kaiser, J., and Schmitt, C. (2020). Mail-in Voter Fraud. *Berkman Klein Center*, 2020–6.

Bhagwat, A. (2019). Free Speech Categories in the Digital Age. In K. Gelber & S. J. Brison (Eds.), *Free Speech in the Digital Age* (pp. 88–103). Oxford University Press.

Bollinger, L., & Stone, G. R. (Eds.). (2022). *Social Media*. Oxford Publishing Press.

Boxell, L., Gentzkow, M., and Shapiro, J.M. (2021). Cross-Country Trends in Affective Polarization. *Review of Economics and Statistics*, 1–61. https://doi.org/10.1162/rest_a_01160

Cho, J., Ahmed, S., Hilbert, M., Liu, B., & Luu, J. (2020). Do Search Algorithms Endanger Democracy? *Journal of Broadcasting and Electronic Media, 64*(2), 150–172.

Cingel, D.P., Carter, M.C., and Krause, H.V. (2022). Social Media and Self-Esteem. *Current Opinion in Psychology*, 45. https://doi.org/10.1016/j.copsyc.2022.101304.

Cobbe, J. (2021). Algorithmic Censorship by Social Platforms. *Philosophy & Technology, 34*(4), 739–766.

Cobbe, J., and Singh, J. (2019). Regulating recommending: motivations, considerations, and principles. *European Journal of Law and Technology*, *10*(3).

Cohen, I. A., & Cohen, A. J. (2022). The Permissibility and Defensibility of Nonstate 'Censorship.' In J. P. Messina (Ed.), *New Directions in the Ethics and Politics of Speech* (pp. 13–31). Routledge.

Cohen, J., & Fung, A. (2021). Democracy and the Digital Public Sphere. In L. Bernholz, H. Landemore, & R. Reich (Eds.), *Digital Technology and Democratic Theory* (pp. 23–61). University of Chicago Press.

Cohen, J. (1993). Freedom of expression. *Philosophy & Public Affairs*, pp. 207–263.

Collins, B., Hoang, D. T., Nguyen, N. T., & Hwang, D. (2021). Trends in combating fake news on social media – a survey. *Journal of Information and Telecommunication, 5*(2), 247–266.

Cook, C. L., Patel, A., & Wohn, D. Y. (2021). Commercial versus volunteer. *Frontiers in Human Dynamics*, *3*. https://doi.org/10.3389/fhumd.2021.626409

Davis, J. L., & Graham, T. (2021). Emotional Consequences and Attention Rewards. *Information, Communication & Society, 24*(5), 649–666.

Delgado, R., & Stefancic, J. (2018). *Must we defend Nazis?* New York University Press.

Digital Services Act, (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065&qid=1689750399256. Accessed 28 Jul 2023.

Douek, E. (2021). Governing Online Speech: From "Post-As-Trumps" to proportionality and probability. *Columbia Law Review, 121*(3), 759–834.

Dwoskin, E., & Tiku, N. (2022). How twitter, on the front lines of history, finally decided to ban trump. *Washington Post*. https://www.washingtonpost.com/technology/2021/01/16/how-twitter-banned-trump/. Accessed 20 Jul 2023.

Everett, C. M. (2018). Free speech on privately-owned fora. *Kansas Journal of Law & Public Policy, 28*(1), 113–145.

Fabienne, P. (2017). Political Legitimacy. In Zalta, E.N.'s (Ed), *The Stanford Encyclopedia of Philosophy*.

Facebook Community Standards. (2023). *Transparency Center*. https://transparency.fb.com/en-gb/policies/community-standards/. Accessed 28 Jul 2023.

Flew, T., & Wilding, D. (2020). The Turn to Regulation in Digital Communications. *Media, Culture, and Society, 43*(1), 48–65.

Forestal, J. (2020). Constructing Digital Democracies. *Political Studies, 69*(1), 26–44.

Forestal, J. (2022). *Designing for Democracy: How to Build Community in Digital Environments*. Oxford University Press.

Formosa, P. (2017). *Kantian Ethics*. Cambridge University Press.

Formosa, P., & Mackenzie, C. (2014). Nussbaum, Kant, and the capabilities approach to dignity. *Ethical Theory and Moral Practice, 17*(5), 875–892.

Gainsbury, S. M., Delfabbro, P., King, D. L., & Hing, N. (2016). An exploratory study of gambling operators' use of social media and the latent messages conveyed. *Journal of Gambling Studies, 32*, 125–141.

Gillespie, T. (2018). *Custodians of the Internet*. Yale University Press.

Gillespie, T. (2022). Do Not Recommend? Reduction As a Form of Content Moderation. *Social Media+ Society, 8*(3), 20563051221117550.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2). https://doi.org/10.1177/2053951720943234

Gorwa, R., & Guilbeault, D. (2020). Unpacking the Social Media Bot. *Policy & Internet, 12*(2), 225–248.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic Content Moderation. *Big Data & Society, 7*(1), 2053951719897945.

Hao, K. (2021). How Facebook Got Addicted to Spreading Misinformation. *Technology Review*. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/. Accessed 1 Dec 2022.

Harriger, J. A., Thompson, J. K., & Tiggemann, M. (2023). TikTok, TikTok, the time is now: Future directions in social media and body image. *Body Image, 44*, 222–226.

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

Hong, S., & Kim, S. H. (2016). Political polarization on twitter. *Government Information Quarterly, 33*(4), 777–782.

Honoré, M.A. (2013). Ownership. In Coleman, J.L.'s (Ed), *Readings in the Philosophy of Law*. Routledge.

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

Howard, J. W. (2021). Extreme Speech, Democratic Deliberation, and Social Media. In C. Véliz (Ed.), *The Oxford Handbook of Digital Ethics.* Oxford University Press.

Hudders, L., De Jans, S., & De Veirman, M. (2021). The commercialization of social media stars. *International Journal of Advertising, 40*(3), 327–375.

Johnson, B. E., & Ho Youm, K. (2008). Commercial speech and free expression: The United States and Europe compared. *Journal of International Media & Entertainment Law, 2*(2), 159–198.

Katz, L. (2008). Exclusion and exclusivity in property law. *The University of Toronto Law Journal, 58*(3), 275–315.

Keller, D. (2021). Amplification and its discontents. *Journal of Free Speech Law, 1*, 227–268.

Klein, C., Clutton, P., & Polito, V. (2018). Topic modeling reveals distinct interests within an online conspiracy forum. *Frontiers in Psychology, 9*, 189.

Klein, C., Clutton, P., & Dunn, A. G. (2019). Pathways to conspiracy. *Plos One, 14*(11), 1–23.

Klein, E. (2023). The Teen Mental Health Crisis, Part 1 (Podcast). *The Ezra Klein Show*. Apple Podcasts.

Lazar, S. (2023). Communicative *Justice and the Political Philosophy of Attention*. https://hai.stanford.edu/events/tanner-lecture-ai-and-human-values-seth-lazar. Accessed 28 Jul 2023.

Lewandowsky, S., Cook, J., Fay, N., & Gignac, G. E. (2019). Science by social media. *Memory & Cognition, 47*(8), 1445–1456.

Lim, M., & Ghadah, A. (2021). Beyond a technical bug. *The Conversation.* https://theconversation.com/beyond-a-technical-bug-biased-algorithms-and-moderation-are-censoring-activists-on-social-media-160669. Accessed 20 Jul 2023.

Llansó, E., Hoboken, J. V., Leerssen, P., & Harambah, J. (2020). Artificial intelligence, content moderation, and freedom of expression. *Transatlantic Working Group on Content Moderation Online and Freedom of Expression*. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf. Accessed 20 Jul 2023.

Mackenzie, C., Rogers, W., & Dodds, S. (2014). *Vulnerability: New Essays in Ethics and Feminist Philosophy*. Oxford University Press.

Magarian, P.G. (2021). The Internet and Social Media. In Stone, A. and Schauer (Eds), *The Oxford Handbook of Freedom of Speech* (pp. 350–368). Oxford University Press.

Merten, L. (2021). Block, hide or follow—personal news curation practices on social media. *Digital Journalism, 9*(8), 1018–1039.

Mill, S. J. (2001). *On Liberty*. Electric Book Co.

Milmo, D. (2021). Facebook revelations. *The Guardian*. https://www.theguardian.com/technology/2021/oct/25/facebook-revelations-from-misinformation-to-mental-health. Accessed 30 Jul 2023.

Newman N., Fletcher, R., Kalogeropoulas, A., Levy, D.A.L., and Nielsen, R.K. (2017). *Reuters Insitute Digital News Report 2017*. Reuters Institute for the Study of Journalism.

Nguyen, C. T. (2018). Echo Chambers and Epistemic Bubbles. *Episteme, 17*(2), 141–161.

Nussbaum, M. C. (2000). *Women and human development*. Cambridge University Press.

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour, 3*, 173–182.

Pettit, P. (2002). *Republicanism: A Theory of Freedom and Government*. Oxford University Press.

Pettit, P. (2013). *On the people's terms*. Cambridge University Press.

Post, R. (2011). Participatory democracy and free speech. *Virginia Law Review, 97*(3), 477–489.

Ranttila, K. (2020). Social media and monopoly. *Ohio Northern University Law Review, 46*, 161–179.

Roberts, S. (2019). *Behind the screen: Content moderation in the shadows of social media.* Yale University Press.

Rozenshtein, A. Z. (2022). Moderating the fediverse. https://www.journaloffreespeechlaw.org/rozenshtein2.pdf. Accessed 20 Jul 2023.

Sahebi, S., & Formosa, P. (2022). Social media and its negative impacts on autonomy. *Philosophy and Technology, 35*(3), 1–24.

Savolainen, L. (2022). The shadow banning controversy. *Media, Culture & Society, 44*(6), 1091–1109.

Schiffer, Z., & Newton, C. (2023). Yes, Elon Musk Created a Special System for Showing You All His Tweets. *The Verge*. https://www.theverge.com/2023/2/14/23600358/elon-musk-tweets-algorithm-changes-twitter. Accessed 25 May 2023.

Schmidtz, D. (2010). Property and Justice. *Social Philosophy and Policy, 27*(1), 79–100.

Settle, J. E. (2018). *Frenemies: How Social Media Polarizes America*. Cambridge University Press.

Silver, L., & Huang, C. (2019). *Emerging economies, smartphone and social media users have broader social networks*. Pew Research Center. https://www.pewresearch.org/internet/2019/08/22/in-emerging-economies-smartphone-and-social-media-users-have-broader-social-networks/. Accessed 11 Oct 2023.

Srinivasan, A. (2023) Cancelled. *London Review of Books,* 45(13). https://www.lrb.co.uk/the-paper/v45/n13/amia-srinivasan/cancelled. Accessed 8 Oct 2023.

Srnicek, N. (2017). *Platform Capitalism*. Polity.

Stanford Internet Observatory. (2022). New report analyzes dynamics on alt-platform Gab. *Cyber Policy Center*. https://cyber.fsi.stanford.edu/io/news/sio-new-gab-report. Accessed 4 Oct 2023.

Stokel-Walker, C. (2022). Reddit Moderators do $3.4 million worth of unpaid work each other. New Scientist. https://www.newscientist.com/article/2325828-reddit-moderators-do-3-4-million-worth-of-unpaid-work-each-year/. Accessed 9 Oct 2023.

Suzor, N. (2019). *Lawless*. Cambridge University Press.

Terren, L., & Borge-Bravo, R. (2021). Echo Chambers on Social Media. *Review of Communication Research, 9*, 99–118.

Tuchman, A. E. (2019). Advertising and demand for addictive goods: The effects of e-cigarette advertising. *Marketing Science, 38*(6), 994–1022.

Tufekci, Z. (2017). *Twitter and Tear Gas*. Yale University Press.

Turner, P. N. (2014). "Harm" and Mill's Harm Principle. *Ethics, 124*(2), 299–326.

Vaidhyanathan, S. (2018). *Antisocial Media*. Oxford University Press.

Vuorre, M., & Przybylski, A. K. (2023). Estimating the association between Facebook adoption and well-being in 72 countries. *Royal Society Open Science, 10*, 221451.

Waldron, J. (2012). *The Harm in Hate Speech*. Harvard University Press.

West, S. M. (2018). Censored, suspended, shadowbanned. *New Media & Society, 20*(11), 4366–4388.

West, L.J. (2021). Counter-Terrorism, Social Media, and the Regulation of Extremist Content. In Miller, S., Henschke, A., and Feltes, J.'s (Eds), *Counter-Terrorism* (pp. 116–128). Edward Elgar Publishing.

Zimmermann, A., & Lee-Stronach, C. (2022). Proceed with Caution. *Canadian Journal of Philosophy, 52*(1), 6–25.

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