Received: 30 November 2023 | Accepted: 2 September 2024

DOI: 10.1002/poi3.425

RESEARCH ARTICLE

P&I Policy & Internet PSO WILEY

Content moderation and the digital transformations of gatekeeping

Ralph Schroeder

Oxford Internet Institute, University of Oxford, Oxford, United Kingdom

Correspondence

Ralph Schroeder, Oxford Internet Institute, University of Oxford, Oxford OX1 3JS.
Email: ralph.schroeder@oii.ox.ac.uk

Abstract

This essay provides an overview of the current state of content moderation on social media platforms. The question the essay addresses is why there are a number of unresolved issues in tackling dysfunctional content. The argument is that there are two intersecting new phenomena which make effective content moderation difficult: one is that social media platforms lack the gatekeeping of content that was characteristic of traditional news media. The second is that the regulation of this un-gatekept content is still unsettled; it falls between social media companies that span the globe and the regulations or absence thereof bounded by nation-states. To understand both, an analysis restricted to law and regulation is insufficient. Instead, it is necessary to examine the role of media systems in society in a holistic way, and in a way that distinguishes between gatekept media and the absence of gatekeeping or new forms of gatekeeping. Such a broader account points to why the institutionalization of content moderation is likely to be a protracted and uneven process. The conclusion spells out how the tensions that have arisen with new media could be resolved, but also why they are likely to remain imperfectly resolved.

KEYWORDS

content moderation, freedom of expression, harmful speech, regulation, social media

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2024 The Author(s). *Policy & Internet* published by Wiley Periodicals LLC on behalf of Policy Studies Organization.

INTRODUCTION

Content moderation and associated debates about hate speech and free speech have recently moved to the top of the media policy agenda. Content moderation seems to involve a bewildering number of topics and questions that include geo-economic and geopolitical competition, regulatory regimes, laws applying to the internet and social media platforms, and how laws and regulations apply to content in countries where foreign social media operate but where domestic laws and norms may conflict with them. There are also larger questions about the legitimacy of content moderation, and whether it should be based on democratic inputs or on processes that social media platforms can implement themselves. Furthermore, these questions seem to apply in a different way in different parts of the world, like in China for example, since state control across all media content there entails that social media platforms are subject to nondemocratic state regulation from the start. Finally, there is the question of how human moderation and automated moderation can work together to ensure the implementation of regulation and norms.

These difficulties are compounded by the absence of separate transnational organizations that are specifically tasked with governing content or enforcing the rules even when they exist. As we shall see, different models of potential transnational regulatory orders have emerged, and these have been labeled the Brussels, Silicon Valley, and Beijing 'effects' (to be explained below). These are unlike previous transnational regulatory bodies like the International Telecommunications Union and the Internet Governance Forum and the UN's International Covenant on Civil and Political Rights, among others. There is nothing for content moderation like the long-standing transnational organizations described in Slaughter's 'A New World Order' (2005) which have facilitated the making of dense webs of rules, standards, treaties, and more. Platforms and the internet generally partly fall outside of these bodies, except with certain laws like 'safe harbor' and how section 230 of the American Telecommunications Act of 1996 have come by default to extend beyond the US. But although the social media platforms claim to be enforcing so-called 'community standards' for content moderation (these have been archived at https://www.platformgovernancearchive.org/) and require agreement to terms of service with certain rules for content, this form of self-regulation has been questioned and it is difficult to enforce. And again, it is unclear how it meshes with the norms and regulations of the countries in which the platforms operate.

The rules applying to content moderation and its legitimacy thus operate in overlapping ways and there are invariably trade-offs. At the same time, it is not clear which bodies or organizations are responsible. Tensions pull in different directions and it is largely left to platforms as the main actors that need to implement content moderation. The plurality of open questions also makes mapping the social science analysis complicated since the various tensions are open-ended and there are a number of academic disciplines involved, such as law, international relations, and communications research, an issue which we will return to below. Furthermore, there are philosophical issues about the limits and extent of free expression, about what constitutes harmful content, and about whether rules can apply nationally or transnationally and so enable or constrain a public sphere for citizens and civil societies. And, as Jungherr and Schroeder (2021) have argued, following Nancy Fraser, the concept of a 'public sphere' which aims at rational consensus is too restrictive; a more appropriate concept is the 'public arena' which is subject to contestation, including about the scope of legitimate media content. In this essay I develop a holistic approach that examines the role of media in society and the gatekeeping function of news media. I focus on how online media are non-gatekept or how the gatekeeping that exists does not yet have established and well-working norms and practices.

To begin with, a distinction can be made between the empirical basis for platform governance as against the normative basis. The empirical basis consists of how public norms about what is

acceptable expression have evolved and how they have been codified and applied. The normative basis consists of the political theory of freedom of expression and its boundaries in the media. The normative side is not the focus here, but it will be contextualized since these norms vary across social contexts. On the empirical side there has been extensive survey and other research about what publics think about media expression. There has also been academic research about how well content moderation works in practice, and some examples will be discussed below. Finally, there has also been much recent debate in the media about the nature of expression on tech platforms, which gives an indication of public concerns and of the shortcomings of platforms. Still, although there are well-established (if not settled) ideas about free expression and its limits in traditional media, these ideas have not yet been consistently applied to platforms.

Thus one step here will be to tease out the differences between traditional media and new platforms, in addition to the cross-country comparison across the three 'effects' (Matassi & Boczkowski, 2023). Before we do so, it is worth pointing out that content moderation is mainly a discussion among professionals. There is a broader discussion among the public about controversial content online which plays out in the media and where the topic is dealt with under different headers - such as online harassment, free speech, foreign interference in elections, pandemic misinformation, and the like. Professionals, which include policy makers, academics, and technology developers, in contrast, treat this issue under the heading of content moderation, and they debate the rules and the enforcement mechanisms pertaining to content moderation under that header, including what is known about the public norms regarding content. There are, of course, also the content creators themselves, who may want to engage with the platforms about the content they post online - and whether it is taken down or downranked or the like. Any discussion of content moderation will need to take both the ongoing debates among professionals and the wider context of the role of media and content creators and their publics into account.

This essay will proceed as follows: first, it will discuss a few key debates and select findings from the growing literature on this topic. Next, it will review some previous ways of thinking about tackling content moderation. Then, it will move to an overview of the problem globally, distinguishing the 'three effects' in how European, US, and Chinese regulatory approaches differ. Thereafter follows an account of how social media platforms should be considered as falling outside of - and yet having some similarities to - the rules that govern the traditional mediated public sphere, which is mainly related to the gatekeeping functions in different types of media systems. Finally, the essay turns to the legitimacy of gatekeeping, and how, even though a global convergence on legitimate gatekeeping norms for social media is unlikely, such norms should nevertheless be sought.

BACKGROUND AND PREVIOUS WORK

The literature on content moderation has grown rapidly in recent years and an exhaustive review that includes policies and practices of moderation and findings about implementation is not possible in a short space. It is also not necessary since Gorwa (2024) provides an up-to-date overview of the debates around online content moderation policies. He argues 'that platform regulation has become merely another (albeit increasingly important) space of transnational political contestation in the global context' (2024:3). But the future, he suggests, is still uncertain, and especially which actors – states, platforms, civil society – play a decisive role in what is still a very open-ended process. Gorwa's approach to tackling this uncertainty is to argue that governments should take greater control, including by setting up independent and accountable bodies that oversee platform regulation. He acknowledges, however, that such efforts will come up against the territorial fragmentation that is likely to lead, instead, to a variety

of national regulation frameworks and so, from a policy perspective, to an inconsistent and less than optimal proliferation of regulatory regimes (Gorwa, 2024: 169-170). Gorwa takes a policy perspective; here, again, the perspective instead is a holistic one.

The holistic perspective developed here entails that content moderation is analysed in the broadest or most encompassing context. This means, apart from the role of media in society, the essay must also address how geopolitical competition is changing how digital media companies are being regulated, which includes their reach across the borders and which, in turn, is not just a question of domestic law and regulation but also how it affects content producers 'on the ground'. Previous work in this area is extensive, but the literature on the topic can be discussed by reviewing a few key debates and approaches. The first of these is a key debate in law, which we shall then need to broaden out to media regulation, its international context, and impact on online expression.

The debate in law turns on a key disagreement concerning how to conceptualize social media and the norms that govern them in contrast with traditional media: as we shall see, this disagreement is decisive for thinking about platform regulation. The opposing positions were crystallized in the views taken by Douek (2022), who favors a systems approach to content moderation as against Klonick (2022) who argues that, if anything, there are multiple such systems. Douek argues that only a 'systemic' approach, which treats platforms as media in a system which must be regulated as such (ie. as a system) can hope to control the harms of these new media. Klonick, in contrast, argues that, as with other social problems arising from new or emerging technologies or other phenomena, a legal approach which treats these problems on a case by case or sectoral basis is more appropriate since the overall effects or harms (and so how to legislate or regulate them) cannot be known in advance, so a stepwise approach will prevent too much (over-) regulation.

The debate between Douek and Klonick is about law and regulation: both are academic lawyers, though Douek's perspective has a broader societal purview. But from a holistic perspective, the question is clearly about what types of systems are involved when such new problems arise, how these systems function, and where the tensions are. The two systems that are relevant here are the media system (not necessarily bounded nationally), the political system, often bounded by nation-states (again, with some transnational flows) and the economic system (there are variants of greater and lesser control of markets by states, but these are only relevant here insofar as they relate to politics and so deal with media and speech). If this conceptualization is correct, then Douek is also correct in pointing to a 'systemic' approach insofar as the media can be seen as a system which is conceptually distinct and plays a role in the political system (even if the boundaries are contested). Klonick might argue that there is complexity such that different laws pertain to different economic sectors or in different countries, and descriptively and from a legal perspective she is right to say that the US has often taken a sectoral approach to regulation. Still, from an analytical, as opposed to a practical legal, perspective, Douek is correct to say that a systemic approach is appropriate since only such an approach could identify how the tensions in the system require certain rules.

To give one example, it is clear that AI will be needed for content moderation (in combination with human input) due to the sheer scale of content, and no doubt the rules used by the AI will be biased in certain directions. But, as Esposito (2022) has argued, AI must be treated as a system of information exchange with humans. Therefore, she argues, to overcome bias, humans should provide more rules for AI rather than fewer if communication should be steered (systemically) so as to serve society better. And if that is the case, then those human rules for content moderation need to consider how media are part of the social system and what rules are needed to resolve tensions in the system (again, from a holistic societal rather than a legal perspective). AI (or algorithms) could, of course, be seen as new forms of gatekeeping, though these forms are unlike those of traditional news organizations:

they are machine-based rather than editorial or journalistic, and the rules that might govern how AI shape media are yet to evolve. We can also see this in how currently, there are daily debates in the news media about the governance of AI.

THE THREE 'EFFECTS': BRUSSELS, SILICON VALLEY, AND BEIJING

We will return to the implications of this argument about a systemic perspective soon. Before we do so, we can turn to the differences in how platforms are currently regulated (or not) and more broadly than for content moderation. In this regard, it has become conventional to distinguish between three 'effects', and to explain these, we make can pairwise comparisons. We can begin by contrasting the Brussels (EU) versus the Silicon Valley (US) effects. For this pair, Tambini (2021) uses the classic distinction proposed by Isaiah Berlin between negative rights ('freedom from') and positive rights ('freedom to'). EU regulation follows traditions in European positive law (positing law which applies to a wide domain based on first principles such as the dignity of persons) while the 'Silicon Valley effect' follows US law with its strong protection of freedoms, including the First Amendment of the US constitution which protects – one might also hypostatizes - freedom of speech and of the media. US law also, as mentioned earlier, follows case precedent rather than positing first principles.

On the basis of this distinction, Tambini argues that the negative rights approach to media freedom is not enough; there must also be positive rights including public intervention to promote free expression, as well as duties of responsible journalism and responsibilities of the media (in view of the subsidies and privileges given them). He argues that the media should be defined 'in terms of the activities they perform', including for the public interest (2021: 130-31) and autonomy or freedom from the state and the market (2021: 131, see also Jungherr and Schroeder, 2021), but he admits that platforms are more likely to be regulated in these terms, as media, in Europe than in the US where such regulation could conflict with the First Amendment.

Tambini is mainly concerned to strengthen the public role of media, including their autonomy and plurality. But while his proposals would also potentially include regulation for content moderation and free speech, these ideas, as Tambini acknowledges, apply to the EU - and the UK where an Online Safety Act was passed in 2023 - but not in the US, and it is not clear how they would apply to platforms operating outside the EU. It can be added in passing that some might want to distinguish within the US between the 'Silicon Valley' as against the 'Washington DC' effect, with the former more open to the world and the latter aimed at domestic regulation and US economic interests: it is not clear how much these diverge, but this could be a further tension among the tensions identified earlier.

Thus the US and EU will remain distinct in terms of their 'effects'. What about the pair of China and the US? Erie (2023) has detailed the 'Beijing effect', whereby the Chinese party-state has laid down extensive legislation and regulations governing online content, with strong sanctions for platforms that do not comply. As Miao (2024) points out, it is precisely the vagueness of these regulations, for example that AI content moderation should be consistent with 'socialist values' that makes them effective. To give just one example, it is not at all clear what the consistency with 'socialist values' would imply for how AI should work in moderating content. This vagueness entails that, since the rules of permissible content are deliberately underspecified, platforms want to be overly cautious to avoid fines or crackdowns and so to stay within limits of what they guess is allowable. The same applies to those who post content. For the US-China pair, Erie (2023) has argued that in terms of law, the two countries now have different positions with regard to foreign law: the US has regarded its own law as 'exceptional' and treated the laws of other countries as 'others' while trying to extend its laws as part of a liberal international order which is now faltering.

The Chinese party state, on the other hand, is now also trying to extend its laws to countries that it engages with, trying to integrate them, as part of a recent and ongoing drive to extend its soft power and its economic power. Erie also notes the contrast between the US individual rights legal system, as against that of 'imperial and contemporary China', which 'have, under ideologies of Confucianism and communism, respectively, emphasized the paternalistic state as granting different privileges and rights to groups ordered within the broader political community' (2023: 743). The aspiration of 'going out into world' (China's terminology for extending its soft power) with the Chinese legal system is to position China as champion of the Global South against what China perceives as American 'racial capitalism' (Erie, 2023). The dysfunctions of American society are of course seen in China primarily through the lens of the inequalities generated by its capitalist economy, with racial inequality foremost among these.

Erie further argues that, instead of this 'other' within the international liberal order, the US should look to its own faults, such as rising populism and xenophobia, and engage with or embrace the needs of the developing countries it deals with. But China also faces tensions of its own between its domestic laws and the efforts to extend its legal order; notably the opposition to notions of human rights. In short, from a bird's eye view rather than a China-centric or US-centric one, there are tensions between the two major geopolitical powers which also host the dominant tech companies and their respective conceptions of laws and rights. These tensions cannot be resolved but must rather be seen in terms of whether there are universal norms or rights that go beyond the extensions of both American and Chinese laws abroad (which we will come back to in the Conclusion). The application of domestic laws within both countries is a separate issue, but social media uses go beyond the laws of the countries in which they originate. Hence it is necessary to invoke the larger current context, with a rising tide of economic nationalism in both countries (e.g., Suesse, 2023; Schroeder 2020), plus their attempts to shape the global order, and if there are universal norms or otherwise enforceable norms beyond the two.

This allows us to turn to a three-way comparison: Bradford, responsible for coining 'Brussels effect', has recently argued (2023) that Brussels should dominate over the Silicon Valley and Beijing effects. She distinguishes between the American 'market-driven', China's 'state-driven', and the EU's 'rights-driven' regulatory models (2023: 11). She also distinguishes between 'horizontal' or inter-state (including the EU's superstate) battles and vertical battles between governments and tech companies. Here we can focus on her arguments not about regulatory models in general, but content moderation in particular. In this regard, her three-fold distinction among regulatory models is too simple: China, like the EU, has a strongly law-focused approach the internet, even if legal rights in China prioritize consumer more than citizen rights. The EU protects both, but it is based on fundamental ideas about human dignity. And while China's government exercises control over digital media, it must also be responsive to citizens at least in maintaining 'performance legitimacy' (Zhao, 2009). In the US, markets have more power, relative to other parts of the world, than the state. But as Philippon (2019) has shown, in some senses the US has a more state-shaped market than Europe or the EU.

Thus instead of horizontal battles between the EU, US and China and vertical battles between the government and tech companies - we must instead go back to (bounded) states versus (non-bounded) markets (Bradford's idea of 'empires', which straddles the two, she admits, is 'metaphorical' (2023: 6)). The difference is thus not between market (US), state (China) and citizens (EU) rights. Instead, the difference is also between consumers and citizens, or between civil and political as against social rights, with the EU having all three, the US having weak social rights, and China having weak civil and political rights but emphasizing social and consumer rights. Further, the Chinese state-driven regulatory model is not in the ascendant worldwide, as Bradford claims (2023: 364): it is ascending for certain purposes mainly, in more authoritarian countries and with limited reach outside of China, and some of the

TABLE 1 The 'three effects'.

	Reach of digital media companies	State-market relationship	Main 'effect' of regulation
Brussels	Negligible	Social market	Citizen and consumer protection
Silicon Valley	Global, apart from Chinese market	Laissez faire	Protecting individuals, sectoral regulation
Beijing	Domestic, plus limited overseas reach	State steerage of economy	Consumer protection with little citizen protection

purposes are not regulatory but more to do with cybersecurity and surveillance as well as economic benefits. The 'three effects' can be summarized in a simplified way in Table 1.

The three 'effects' are thus likely to continue to compete, but with platforms spanning across them, so that tensions will remain. The example often given about this are the tensions between EU regulators and American platforms (see Gorwa, 2024). But there are others that straddle all three: just one illustration here are the short video platforms owned by Bytedance, TikTok and Douyin. TikTok falls into the Western effects insofar as the app has users outside of China while Douyin, the version used in China, falls into the Beijing effect. The example is telling because TikTok has caused well-documented controversy in the US, with Congress passing a law whereby TikTok must either sell the app to a non-Chinese owner, or cease to operate in the US (Naughton, 2024). But in the EU, TikTok has not been subject to controversy as in the US, but since it is Chinese owned, it must comply with EU regulations over content moderation. In short, TikTok is a prime example of tensions over which 'effect' applies.

Further, these effects apply not only to content moderation in the three parts of the world where the 'effects' originate, but also in what Gorwa (2024) calls 'the majority world' (or also the Global South). To illustrate this, we can turn briefly to Africa. As Garbe, Selvik and Lemaire point out, the platforms 'appear comparatively inactive on the African continent' (2023: 87) in terms of content moderation. They give the examples of Meta and Twitter, which removed hardly any content in Africa compared with large countries outside that continent (2023: 88). At the same time, in these often less than democratic and sometimes authoritarian countries, the state uses mostly technological means (shutdowns, restricting access, or blockages) but also legal approaches to regulate content that challenges the state's legitimacy. Thus we need to consider the 'effects' in contexts where it will be primarily regulated for the sake of keeping governments in power by eliminating critical voices. But while discussion in the Global North centres on content that conflicts with civility, there is much more of a need for platforms to prevent content regulation which is effectively censorship. Similar considerations apply to cases of borderline content which may offend cultural sensibilities, and which are illustrated by TikTok and how it operates in yet other parts of the world such as South and Southeast Asia (Liu, 2024).

HOW CONTENT MODERATION WORKS

Against this background of country comparisons, we can turn to how content moderation actually works on different platforms, 'on the ground'. Since we don't know much about how and why content moderation changes, we can take the example of Twitter, where de Keulenaar et al. (2023) have traced these workings and changes. The authors make a

distinction between 'ugly' content which is illegal and where moderation policies of removing this content have not changed very much, as against 'bad' content, which either meets certain 'levels of harm' or violates shifting norms of acceptability: 'bad content' is 'content is defined as objectionable by its stakeholders and users' (2023: 280). In respect to 'bad content', Twitter has increasingly adopted a policy of downgrading or issuing warnings against this type of content. de Keulenaar et al argue that this 'modulated moderation' has made moderation more differentiated but also contradictory, since the responsiveness to 'stakeholders and users' is bound to be 'contextual', shifting with time but also with the priorities of the platform not to lose users - who might go elsewhere because Twitter is too 'strict' but also because it might be perceived as too 'lax' and unwelcoming or having negative social impact. It needs to be mentioned that the study was done before Twitter (now called X) changed with Elon Musk's ownership, with further complications, as we shall see. This is also a good place to mention that the transparency reports that platform companies provide may or may not contain accurate enough information about content takedown, which also depend on the metrics that they use.

Apart from taking a platform-centric approach, it is possible to take a user-centric approach to content moderation. One such user-centric approach is to glean the public's views about content moderation with experimental evidence. Thus the survey experiment by Kozyreva et al. (2023) asked Americans about the rules they want for content and whether they prefer free speech or removal of harmful misinformation. The survey found, among other things, partisanship between Democrats who wanted more removal and Republicans who wanted less. Perhaps more interesting here is the common ground between them: of the four scenarios that those surveyed were presented with - election denial, anti-vaccination, Holocaust denial, and climate change denial—there was most agreement for the removal of Holocaust denial. Furthermore, it is also interesting to note that the party stances have historically been reversed: it can be imagined that conservative Republicans used to be more in favor of curbing speech (we can think here of the McCarthy era in the 1950s) than Democrats, who were champions of free speech (we can think here of the Berkeley free speech movement in the 1960s). And it is also important to note that currently, support for free speech and regarding censorship is selective among both parties, as when some in the Republican party and their supporters favor banning content in line with the 'culture wars' (which go back at least to the 1990s, so before online content moderation) while many Democratic politicians and supporters oppose restrictions on progressive cultural expression.

THE QUESTION OF THE LEGITIMACY OF CONTENT MODERATION

The changing politics of what is regarded as legitimate expression brings us to the question of the legitimacy of content moderation. de Keulenaar et al's (2023) analysis of Twitter comes to the conclusion, given the shifting sands that they identify, that platforms have a 'lack of legitimacy' in relation to speech regulation (2023: 281). But legitimacy can subject to various interpretations: there is the legitimacy of rulers and of rules or of political authority, as conceptualized by Max Weber. Then there is the legitimacy of scientific authority, which might, for example, apply in the case of misinformation about climate change or vaccines. Another more common sense understanding of legitimacy is where it would be more appropriate to talk about the acceptability of norms among the public. Legitimacy could also be 'procedural', which applies to state authority, but it can also apply outside the state to the norms of institutions such as international legal bodies and then depend on the legitimacy of those institutions or their authority outside of the state.

This question of the perception of different types of legitimacy has been mapped by Haggart and Keller (2021), who compare several cases in terms of 'input', 'throughput' and 'output' legitimacy. They argue that the UK's Online Harms Bill (the predecessor to the Online Safety Act mentioned earlier) is most grounded in democratic state-based legitimacy which other proposals lack. Their other cases, Kaye's (2019) UN universal human rights based principles for freedom of speech which is aimed against states which curb it, or Facebook's Oversight Board, or Judicial Adjudication as with the Manila Principles developed by civil society groups—lack such democratic legitimacy. But Haggart and Keller's analysis merely points to how different types of legitimacy have different sources. The obvious solution here, a differentiated approach, which allows universal principles to supersede where appropriate and national (state-based) legitimacy to govern those aspects which do not conflict with universal principles but impose standards on subordinate ones—is too simple for a world in which platforms based in one nation with its rules extend their rules to other nations.

THE TRANSFORMATIONS OF GATEKEEPING

Content moderation is often seen through the lens of platforms and their dysfunctions and obligations. And while the functions and obligations of traditional media and of the public arena are well-understood, platforms operate for profit and within the law—without the functions and obligations of traditional media and the public arena. This difference deserves spelling out: The gatekeepers for traditional media have been based on the norms of public speech and behavior plus journalistic standards. Gatekeeping (Shoemaker & Vos, 2009) is well-established as a theory of how traditional media have established norms, such as newsworthiness, or for filtering news and information, and so also keeping out other information (e.g. non-newsworthy, or partisan), for the public about events and especially about politics. Here we can therefore see how content moderation is a new problem arising with new social media technology: in the era of mass media (and even partly the internet before social media), content moderation was not a problem. Mass media had gatekeepers which would not have allowed, say, beheadings or posting hate towards public figures. Social media have terms of service which can be stricter that national laws, but they do not have established or failsafe safeguards against users posting content such as this. Furthermore, this newness also applies to emerging technology like automated (algorithmic or AI) content moderation which is bound to be part of the solution for how content is moderated.

But platforms strain against the established norms and the lack of institutions that uphold them when expression pushes the boundaries of communication outside of gatekept ones. Thus non-gatekept content on platforms exists in a kind of limbo where it is not clear which rules apply (as with foreign election interference) and borderline cases which are used to test if or when new rules can be applied. Content moderation could be governed by the norms of the public sphere (or 'public arena' a term which points to the contestedness of this 'sphere'; see Jungherr & Schroeder, 2021), which was gatekept in the era of traditional media, but goes beyond it because social media also consists of non-gatekept user-generated content (UGC). The norms of the public arena could thus include that there should be maximal scope for citizens and civil society within the bounds of Mill's no harm principle. Put differently, there could be maximal toleration within the bounds of no harm, but one problem is that citizens and publics may not favor this principle, and another is that this may make parts of the online sphere inhospitable to or harmful for certain groups.

Gatekeeping - but also the non-gatekept part of media content - is shaped by the 'three effects' discussed earlier. In China, the non-gatekept part of the public sphere is regulated towards political stability. In the EU, the non-gatekept part of the public sphere is regulated

procedurally, with the aim of creating an order that serves the public transparently. The US is optimized for free speech, with the quest for regulation politicized along partisan lines. But commercially, the platforms also vary: The platforms in China maximize active engagement rather than passive attention via advertising; they must maintain order online or face formal or informal government sanctions (Gorwa, 2024: 149-53). American platforms want to maximize attention for advertising revenue and there are debates about whether this stirs up political enmity. In any event, American platforms need to regulate themselves or face sanctions and they also need to fit with the constraints imposed on them outside the US such as the Digital Services Act in the EU. In other words, for gatekeeping too, looking at regulation and the 'three effects' alone does not suffice: it is also necessary to examine the political and economic environment of media; in other words, the role of media in society.

Gatekeeping for traditional media and the public arena is about the acceptability of content to which the state and regulators but also about how journalistic norms set bounds. But it is also the case that the non-gatekept part of media must be able to challenge the state and to challenge other types of authority such as scientific expertise in the case of climate change. The bigger picture here is that it is not clear yet how content moderation—as opposed to platform governance generally (which includes other areas such as cracking down on scams, bots, online commerce, etc—the border with content moderation might be blurry here, as with 'fake attention') – should take place: platforms only partly fall within the gatekept public arena. For example, content from the gatekept public arena like TV news is put onto non-gatekept platforms like YouTube—but YouTube does not follow other norms of gatekeeping or of the public arena like providing 'balance' or 'inclusiveness' in or among news programmes—see Figure 1 below and the area of overlap between the two media circles; they are also part of a non-gatekept broader public venue for expression. So – gatekept versus non-gatekept are different, while platforms are under increasing pressure to gatekeep what was previously non-gatekept.

Nielsen & Ganter (2022) have noted the shift to platform power and away from traditional media, and platforms have the capability to target and tailor people with information. This calls for regulation about how harmful content is delivered outside of traditional 'delivery systems' but also in terms of how free speech is secured by platforms (even if, as private

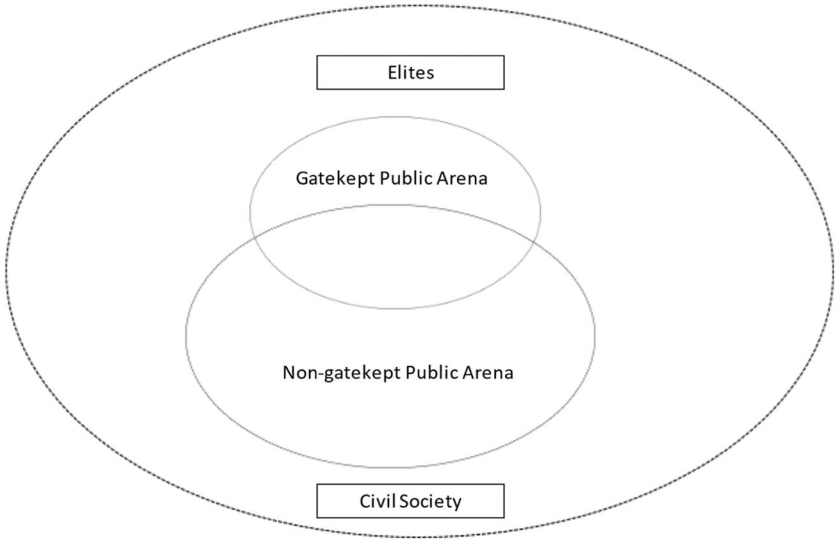

FIGURE 1 Gatekept and non-gatekept public arenas.

companies, securing free speech is not within their remit). Finally, it should be remembered that, while much has been made of the situation whereby private sector platform companies benefit from how 'lawful but awful' content increases engagement and so profits, it is also the case that these companies must retain engagement and thus not put off their user base. There is a related issue, which is that platforms can try to counteract organizations that point out how much 'awful' content they are permitting to gain more revenue, as with Twitter (or X) https://www.theguardian.com/technology/2023/aug/08/techscape-elon-musk-x-twitter-lawsuit-ccdh-hate-speech.

Musk's takeover of Twitter is also an interesting case because unlike platforms like Facebook which have obligations towards shareholders, which includes their reputations, a volatile CEO like Musk who sees himself as unconstrained by obligations to shareholders is more like the media tycoons of an earlier era who could shape content according to their whims. Thus Twitter (X) has been documented to take contradictory stances on speech that are attributed to Musk (PBS, 2023) and under increasing scrutiny after Musk's takeover for whether it is devoting enough resources to content moderation (New York Times, 2022). A further issue, however, is that a vibrant non-state commercial or a not-for-profit sector of platforms - vibrant in the sense of providing alternative sources of information (Dahl, 1998) - is necessary for a healthy democratic environment for free expression and provision of information.

It is true that the distinction between gatekept and non-gatekept content and forms (media) is blurry, as with political podcasts that are not gatekept on social media or citizen journalists that constitute a 'fifth estate', as Dutton has called it (2023). But this blurriness in practice is not an argument against analytical distinctions or distinguishing between how norms apply in gatekept media and outside them. Further, content moderation can also take place via downranking or filtering or similar techniques, which can be seen as a kind of gatekeeping (for the example of TikTok, see Zeng & Kaye, 2022). Finally, the use of AI in journalism (see, e.g., Simon, 2022) is bound to blur the distinction in the future still further, though similar arguments about maintaining analytical distinctions apply.

Trust in the traditional gatekept media and perhaps in their gatekeeping function among the public has declined (Reuters Institute, n.d), as has the prominence of gatekept media. There may also be upsides to this declining prominence, as with a greater diversity of voices, but we should also bear in mind the argument that the public arena should constantly allow the contestation of the bounds of gatekeeping in light of changing norms. Still, the function of the gatekept media is to set the political agenda, and the lack of trust and lesser prominence of gatekept media can of course provide strong arguments to defend these media further.

FREE SPEECH AND HARM

Content moderation is about expression and harm. But there are ongoing debates about the boundaries of free speech and hateful speech (see the overview by Howard, 2019), with agreement that curbing speech should not be abused by governments to curb criticisms, as is sometimes done in less-than-democratic states, since that is self-serving censorship. In relation to censorship, another way to foreground the bounds of free expression are debates about how it can be used to silence critics within academia. Srinivasan (2023) distinguishes between academic freedom and freedom of expression to make the point that freedom of expression does not necessarily obtain in the academy: Academic freedom entails open and reliable knowledge whereas free expression is an input into politics. The two can overlap when countries try to restrict universities for political reasons, as in China and India, and this also happens within the US when expression is politicized in the academy over certain

issues. So, for example, with the recent controversies of the Israel-Hamas war and the resignation of presidents of Ivy League universities like Harvard (BBC, 2024; for the quite different case of Germany, see Revers and Traunmüller, 2020). Even in China, universities, and especially foreign universities, enjoy a certain amount of academic freedom, though within strict limits. In India, to take a different example, prime minister Modi has tried to curtail academic freedom by appointing university administrators who support his political agenda. Universities should be bastions of open and critical political debate.

These issues go beyond platforms, but what they highlight is that media, including in nontraditional media, require some degree of media autonomy from states and other institutions. They also illustrate that harm pertains not only to individuals but also to the harm that can be done when institutions that should be autonomous, like universities or traditional media, are threatened. Social media require autonomy for the somewhat different reason that they should not be beholden to states that try to censor them. Finally, the choices or trade-offs are not between free expression and content moderation is not a zero-sum game since both extremes of completely free expression and completely moderated content are untenable and there is bound to be borderline content which is subject not only to trade-offs or tensions but also legitimate contestation.

FROM FACTS TO NORMS

It is useful to consider for a moment the different ways that social science perspectives can be brought to bear on content moderation: anthropology can tell us that harms and free speech are contextual. This brings into focus the scope and limits of cultural relativism: such knowledge must also contextualize law and ethics within different political systems, so that cultural relativism can be overcome by objective and reliable knowledge. An international relations perspective concerning the relations between states - in this case, as we have seen, the three effects – can distinguish three competing zones which are also in contention for influence and reach across the rest of the world. There is currently much concern about the role of social media, and especially content moderation, due to the recent and ongoing political malaise and especially right-wing populism. The impression that one has, especially with academic research efforts including from computer scientists about content moderation, is that the aim is to improve the non-gatekept public arena with stronger content moderation; a laudable goal. Further, there is much concern with mis- and disinformation, even though many would agree that the main problem is often misleading information rather than outright falsehoods. But content moderation cannot and arguably should not be tasked with shaping politics top-down since non-gatekept content can both be an expression of bottom-up discontent as well as content that comes top-down from challengers to the status quo. The discipline of law can clarify the applicable norms and principles involved, but as we saw earlier, a conceptualization of the media as a system and its role in society is also required. Finally, the arguments presented here could no doubt be expanded by adding the lobbyists, policymakers, and NGOs and other actors (such as creators) and their strategies; but these elements are part of civil society and can be incorporated as structural elements of the role of media in society (see Figure 1) and their transformations. Put differently, long-term and structural and de-politicized principles are needed for an interdisciplinary analysis.

All platforms have faced similar dilemmas in terms of content moderation but treat them in quite different ways. There is a difference, for example, between how, in the US and Europe, the debate has been to push for more moderation via regulation to minimize harms, while in authoritarian countries like China, content moderation is about control that is aimed at social stability which, as mentioned, relies on the state sanctioning media and on the vagueness of regulation (Gorwa, 2024; Miao 2024). To achieve this for the non-gatekept

part of the public sphere should not only be subject to platforms maintaining their popularity among users (and so also not subject to users' whims), nor to the whims of states and regulators or media owners, nor to public acceptability generally (apart from users' legitimate inputs) - but rather, universally, subject to a mix of regulated and self-regulated principled rules for a plural sphere of (online) expression.

This sphere of online expression partly overlaps with but it is also different from the gatekept part of the public arena which needs to uphold the additional principle of providing a forum suitable for issues of common concern (Jungherr & Schroeder, 2021). The non-gatekept part, in contrast, adds issues of concern that cannot be found in traditional media and but that also pose a challenge to the very definition of existing common concerns and add private expression. Essentially, the non-gatekept part is now being shaped by various forces – states and how they impose regulation (or not), platforms, and civil societies – all with their different norms around 'objectionability' (de Keulenaar et al., 2023) over and above harm – for which there may be 'global' standards in theory if not in practice. Apart from these forces, as we have seen, there are several tensions. For non-gatekept platforms, as long as there are different types with competition among them, and as long they do not promote harm and so follow the injunction against not impinging on negative liberty, it is not clear what the rules should be except, of course, that they compete for users, who may be put off by certain content or by its removal. Social science can supply the analytical concepts and empirical findings for defining the boundaries of harm or other content that needs moderation – for example by having concepts about which groups are harmed and the like - and thus pave the way for normative debates about options.

One way towards an objective and holistic account and so to make progress is to avoid tendentious language, which brings us back to the earlier discussion about systems. Let us think for a moment about Google and Facebook and Twitter and the like simply as sociotechnical systems; or rather, as parts of a larger online media system. That way, we avoid 'platforms', which is a business term, but also infrastructures, which is a term from science and technology studies but which is only partly appropriate: partly, because infrastructures, whether in the private or public sector, serve some kind of essential societal function - and how 'essential' platforms are is subject to debate. These systems then, if we limit ourselves to media rather than the provision of services like Amazon, are not infrastructures in as much as they are not governed by the same rules as the public arena; they are partly outside its gatekept part. But they should be subject to governance when they serve essential functions like health information or when they are essential to socializing with others. This duality, being both private and public and both essential for some things but not for others, is not unique to these systems: here we can think of transport or energy supplies. 'Platforms' - or these systems - are unique only in that their function is to serve needs for information and communication rather than energy or transport and the like, and so part of the public arena of media, gatekept or not. And again, the public arena of traditional media like news organizations had a - however imperfect - gatekeeping function in the media system; 'platforms', which have been added to the media system but are un-gatekept in this way, have yet to establish settled norms.

The rules for information and communication, or concerning speech and harm, have always been imperfect. But we know where we are coming from – a history of oppression and intolerance – and where we are trying to go – the opposite. Platforms (or sociotechnical media systems) are one new vehicle for expression, and the US has section 230 of the Telecommunications Act of 1996 (which means they are not responsible for content, unlike other media) while China has the opposite (platforms are legally responsible for content). Germany's NetzDG is more restrictive than the US and the EU's Digital Services Act of 2023 is set to steer a different course. These old and new rules are for the private sector, and here (again) we are only interested in the rules that apply to communication and information (media) and not to services like e-commerce and ridesharing and the like.

What we can agree on is that political elites and governments should not be allowed to use curbing free speech via content moderation as a tool to keep themselves in power. We can also agree that universities and scientific institutions should be autonomous and producers of open knowledge, and that gatekept media systems should be autonomous not just from states but also from undue market or commercial interference. We can further agree that transnational considerations of security sometimes override these general principles (see Garton Ash, 2016) but this should be kept to a minimum required for this purpose. Further, the no harm principle can be agreed as a baseline, before we get to content moderation for other reasons. The tensions – in practice – can therefore be deduced from this combination of agreed baselines. Still, there are bound to be trade-offs beyond these baselines: the tools for content moderation will inevitably apply negative utilitarianism – minimizing harm - but these tools may have too many 'false negatives' (and so miss harmful content) or too many 'false positives' (and so over-moderate content). But the danger of curtailing free speech will also come from governments and not just from platforms and their imperfect tools.

Garton Ash has argued for 'a more universal universalism' in relation to free expression and that 'we must strive to create conditions in which we agree on how to disagree' even if we can 'never all agree' (2016: 380-81). There is an ongoing process by UNESCO to develop guidelines for 'safeguarding freedom of expression and access to information…in the context of regulating digital platforms' with a multistakeholder approach https://www.unesco.org/en/internet-conference/guidelines (but recall the discussion of Haggart and Keller). It can also be mentioned that the UN's article 19 and the subsequent covenant have been ratified by most countries except China. Going forward is thus likely to be similar to the past; muddling through in the foreseeable future and living with rules catching up and leaving a vacuum where agreement is still subject to contestation. It can be mentioned, finally, that decentralized solutions like Mastodon have been added to avoid the problems of 'one-size-fits-all' or the problem of having rules imposed from above as with Musk's Twitter (now X). But these solutions, even if they were to become popular, create new ones, such as how expression works across the 'islands' governed by the different decentralized 'communities' that affiliate with those with like-minded ideas about regulation, and how to keep one's personal voice across different such 'communities'.

CONCLUSION

There may come to be some convergence between the Silicon Valley and Brussels effects, but the Beijing effect will remain distinct in terms of its legitimacy and China's platforms will remain essentially nation-bound. But there is also a question about how the gatekept public arena relates to the different and non-gatekept online realm, even if both share certain features, such as that they should be autonomous from the state and from undue market influence. But the online realm outside the gatekept media realm is also differently gatekept; at a minimum negatively rather than positively as with journalistic and news media norms. Thus it should exclude certain – harmful – content, but it pays little heed to diversity or reliability and the like, as the gatekept public arena does. So, this broader online realm has added differentiation to the media system, and how its content is much more lightly moderated reshapes a mediated society, even if its rules – and rules about the absence of rules – are still emerging.

One set of rules for this new part of the online realm applies to harms related to expression, and here, the negative injunction is a baseline. Other online harms not related to expression – election interference, climate or vaccine misinformation – may need more than negative injunctions. They are all part of a media system, though with two partially

overlapping parts (the gatekept public arena, and beyond it). Platforms need to be bound by rules from outside rather than just being allowed to self-regulate if they violate the negative injunction, and perhaps more. However, they also have other forces which impel them towards applying rules because they need to create a more positive online environment for expression and for information because their reputation determines their reach. Yet their reputation and reach cannot be the sole determinant as it may come at the expense of the interests of minorities, which need to be protected. Furthermore, the nongatekept part of the public arena part of the online or media environment also subject to norms associated with its functions: creating a vibrant civil society that makes for a society that can make demands on its elites towards more plural and diverse inputs (including minorities) in a peaceful society that deepens and broadens equal freedoms. Since this non-gatekept part is the domain of private sector companies, they are under no formal obligation to further these norms. Here I have dealt mainly with harm in relation to free expression. This approach could also be extended to other types of online harms. Social science can lay out the conditions – dominant forces, main patterns and what shapes them, and tensions – that determine the severity and consequences of harms, and then identifies the main options and the trade-offs in addressing them.

ACKNOWLEDGMENTS

I am grateful to comments from Manuel Tonneau and Diyi Liu on earlier drafts. Also helpful were comments at the workshop on Content Moderation and Free Speech at the Technical University of Munich in October 2023 and very useful sets of comments from three reviewers of this journal.

ORCID

Ralph Schroeder http://orcid.org/0000-0002-4229-1585

REFERENCES

BBC (2024). Claudine Gay resigns as Harvard University president. https://www.bbc.com/news/world-us-canada-67662871

Bradford, A. (2023). *Digital Empires*. Oxford University Press.

Dahl, R. (1998). *On Democracy*. Yale University Press.

Douek, E. (2022). Content moderation as systems thinking. *Harvard Law Review*, *136*, 526.

Dutton, W. (2023). *The fifth estate: The power shift of the digital age*. Oxford University Press.

Erie, M. (2023). Legal systems inside out: American legal exceptionalism and China's dream of legal cosmopolitanism. *University of Pennsylvania Journal of International Law*, *44*, 731–813.

Esposito, E. (2022). *Artificial communication*. Cambridge: MIT Press.

Garbe, L., Selvik, L. M., & Lemaire, P. (2023). How African countries respond to fake news and hate speech. *Information, Communication & Society*, *26*(1), 86–103.

Garton Ash, T. (2016). *Free speech: Ten principles for a connected world*. Yale University Press.

Gorwa, R. (2024). *The Politics of Platform Regulation: How Governments Shape Online Content Moderation*. Oxford University Press.

Haggart, B., & Keller, C. I. (2021). Democratic legitimacy in global platform governance. *Telecommunications Policy*, *45*(6), 102152.

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

Jungherr, A., & Schroeder, R. (2021). *Digital Transformations of the Public Arena*. Cambridge University Press.

Kaye, D. (2019). *Speech Police: The Global Struggle to Govern the Internet*. Columbia Global Reports.

de Keulenaar, E., Magalhães, J. C., & Ganesh, B. (2023). Modulating moderation: A history of objectionability in twitter moderation practices. *Journal of Communication*, *73*(3), 273–287.

Klonick, K. (2022). Of systems thinking and straw men. *Harvard Law Review*, *136*, 339.

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, *120*(7), e2210666120.

Liu, D. (2024). Borderline content and platformised speech governance: Mapping TikTok's moderation controversies in South and Southeast Asia. *Policy & Internet*, 1–24.

Matassi, M., & Boczkowski, P. (2023). *To Know Is to Compare: Studying Social Media across Nations, Media, and Platforms*. Cambridge MA: MIT Press.
Miao, M. (2024). Regulating Digital Platforms through Sanctionsns *Washington International Law Journal*, *33*(2).
Naughton, J. (2024). TikTok may be on borrowed time in the US, but it still holds a Trump card. https://www.theguardian.com/commentisfree/2024/mar/16/tiktok-may-be-on-borrowed-time-in-the-us-but-it-still-holds-a-trump-card
New York Times (2022) Hate Speech's Rise on Twitter Is Unprecedented, Researchers Find https://www.nytimes.com/2022/12/02/technology/twitter-hate-speech.html
Nielsen, R. K., & Ganter, S. A. (2022). *The power of platforms: Shaping media and society*. Oxford University Press.
PBS (2023). Former Twitter Insider Describes Elon Musk's Mixed Signals on Free Speech, https://www.pbs.org/wgbh/frontline/article/twitter-elon-musk-free-speech-x-documentary-excerpt/
Philippon, T. (2019). *The Great Reversal: How America Gave Up on Free Markets*. Harvard University Press.
Reuters Institute (n.d). *Trust in News Project*. https://reutersinstitute.politics.ox.ac.uk/trust-news-project
Revers, M., & Traunmüller, R. (2020). Is free speech in danger on university campus? some preliminary evidence from a most likely case. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *72*(3), 471–497.
Schroeder, R. (2020). Political power and the globalizing spread of populist politics. *Journal of Political Power*, *13*(1), 22–40.
Shoemaker, P. J., & Vos, T. (2009). *Gatekeeping theory*. Routledge.
Simon, F. M. (2022). Uneasy bedfellows: AI in the news, platform companies and the issue of journalistic autonomy. *Digital Journalism*, *10*(10), 1832–1854.
Slaughter, A. M. (2005). *A new world order*. Princeton University Press.
Srinivasan, A. (2023). 'Cancelled', *London Review of Books* https://www.lrb.co.uk/the-paper/v45/n13/amia-srinivasan/cancelled
Suesse, M. (2023). *The Nationalist Dilemma: A Global History of Economic Nationalism, 1776-Present*. Cambridge University Press.
Tambini, D. (2021). *Media Freedom*. Polity Press.
Zeng, J., & Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, *14*(1), 79–95. https://doi.org/10.1002/poi3.287
Zhao, D. (2009). The mandate of heaven and performance legitimation in historical and contemporary China. *American Behavioral Scientist*, *53*(3), 416–433.

How to cite this article: Schroeder, R. (2024). Content moderation and the digital transformations of gatekeeping. *Policy & Internet*, 1–16. https://doi.org/10.1002/poi3.425