Gillespie, Tarleton et al.

Article

Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates

Internet Policy Review

Provided in Cooperation with:
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

Suggested Citation: Gillespie, Tarleton et al. (2020) : Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 9, Iss. 4, pp. 1-29, https://doi.org/10.14763/2020.4.1512

This Version is available at:
https://hdl.handle.net/10419/225649

INTERNET POLICY REVIEW
Journal on internet regulation

Volume 9 | **Issue** 4

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Expanding the debate about content moderation: scholarly research agendas for the coming policy debates

Tarleton Gillespie *Microsoft Research* tarleton@tarletongillespie.org
Patricia Aufderheide *American University*
Elinor Carmi *University of Liverpool* Elinor.Carmi@liverpool.ac.uk
Ysabel Gerrard *University of Sheffield***Robert Gorwa** *University of Oxford*
Ariadna Matamoros-Fernández *Queensland University of Technology*
Sarah T. Roberts *University of California, Los Angeles*
Aram Sinnreich *American University***Sarah Myers West** *New York University*

DOI: https://doi.org/10.14763/2020.4.1512

Published: 21 October 2020
Received: 19 May 2020 **Accepted:** 5 August 2020

Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Gillespie, T. & Aufderheide, P. & Carmi, E. & Gerrard, Y. & Gorwa, R. & Matamoros-Fernández, A. & Roberts, S. T. & Sinnreich, A. & Myers West, S. (2020). Expanding the debate about content moderation: scholarly research agendas for the coming policy debates. *Internet Policy Review*, *9*(4). DOI: 10.14763/2020.4.1512

Keywords: Content moderation, Platforms, Internet policy, Social media, Regulation

Abstract: Content moderation has exploded as a policy, advocacy, and public concern. But these debates still tend to be driven by high-profile incidents and to focus on the largest, US based platforms. In order to contribute to informed policymaking, scholarship in this area needs to recognise that moderation is an expansive socio-technical phenomenon, which functions in many contexts and takes many forms. Expanding the discussion also changes how we assess the array of proposed policy solutions meant to improve content moderation. Here, nine content moderation scholars working in critical internet studies propose how to expand research on content moderation, with implications for policy.

This paper is part of **Trust in the system**, a special issue of *Internet Policy Review* guest-edited by Péter Mezei and Andreea Verteş-Olteanu.

Introduction

By Tarleton Gillespie and Patricia Aufderheide

Content moderation scholarship faces an urgent challenge of relevance for policy formation. Emerging policies will be limited if they do not draw on the kind of expansive understanding of content moderation that scholars can provide.

Content moderation – the detection of, assessment of, and interventions taken on content or behaviour deemed unacceptable by platforms or other information intermediaries, including the rules they impose, the human labour and technologies required, and the institutional mechanisms of adjudication, enforcement, and appeal that support it—has exploded as a public, advocacy, and policy concern: from harassment to misinformation to hate speech to self-harm, across questions of rights, labour, and collective values. Academics have explored how and why platforms moderate, what kind of publics they're producing in doing so, and what responsibilities they should hold around these interventions. This work has opened up even more fundamental questions about the enormous power of platforms.

As concern about moderation has grown, scholarly attention has grown with it—somewhat—from specific controversies to deeper, structural questions about how moderation is organised and enacted. But there remains a tendency for research to be driven by high-profile incidents and people: the 2016 election in the United States put worries about misinformation [1] front and centre; the Christchurch shooting pushed hate and domestic terrorism as the highest priority; the Covid-19 pandemic put misinformation and conspiracy back in front.

Moreover, discussion tends to focus almost exclusively on the largest, US based platforms. There are good, or at least understandable reasons, why this is so. These platforms are enormous, and their policies affect billions of users. Their size makes them desirable venues for bad faith actors eager to have an impact. Their policies and techniques set a standard for how content moderation works on other platforms, they offer the most visible examples, and they drive legislative con-

1. Here we use misinformation as an umbrella term for various kinds of unreliable information. For a more tightly defined definition, distinguishing between misinformation, disinformation, fake news, and propaganda, see Carmi et al., 2020; Jack, 2017.

cerns. But the inordinate attention is also structural. Critics talk about Facebook and YouTube as stand-ins for the entire set of platforms. Journalists hang critical reporting on high profile decisions, blunders, and leaks from the biggest players. Scholars tend to empirically study one platform at a time, and tend to choose large, well-known platforms where problems are apparent, where data will be plentiful, and that are widely used by or familiar to their research subjects.

This tendency extends to policy-making. US and European policymakers have also focused on the latest controversies and the biggest players. It is not surprising that, when the US Congress began to probe these questions, first in the hot seat was Mark Zuckerberg, followed soon after by senior leaders from Google, Twitter, and YouTube. As enormously significant and problematic as Facebook certainly is, it has become inordinately prominent in moderation debates—the preferred object of research, the go-to example in characterising the problem, the stand-in for all other platforms.

Academic scholarship on content moderation can contribute to this increasingly urgent policy discussion, and is regularly being called on to do so. But to address some of these tendencies, it must be grounded in an understanding of moderation as an expansive socio-technical phenomenon, one that functions in many contexts and takes many forms. Expanding the scope and range of research on content moderation is critical to developing sound policy. The range of contentious phenomena cannot be captured by studying just misinformation, or hate, or pornography. The largest, US-based platforms do not provide a reliable guide for the entire social media ecology; innovative moderation strategies may emerge from smaller platforms, platforms outside of the US, and platforms that imagine themselves and their communities very differently than Facebook does. New platform tactics and new laws require more scrutiny—not just so that they may be understood, criticised, and improved, but to understand how they implicitly frame the nature of the problem, positioning some approaches as on the table for discussion and sweeping others out of view.

And any policy enacted to regulate moderation or curb *online harms*, [2] while it may reasonably have Facebook or YouTube in its sights, will probably in practice apply to all platforms and user-content services. In that case, the result could further

2. While *online harms* as a term gets used more generally, it also harks to the UK's Online Harms White Paper (2019), which takes the first step in developing a new regulatory framework for online safety and "make clear companies' responsibilities to keep UK users, particularly children, safer online". https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper.

consolidate the power of the biggest tech companies, those best able to manage the regulatory burdens. This has been a concern in related areas, for example privacy (Krämer & Stüdlein, 2019) and copyright protection (Samuelson, 2019; Romero-Moreno, 2019). If power asymmetries are to be challenged, we need to understand how different values are engineered into these mechanisms across a wider array of examples.

And, so much may change in the wake of the global Covid-19 pandemic, not only about how content moderation is accomplished, but the broader position of social media platforms amid democratic life and theories of public responsibility. Our scrutiny and our efforts should look very different when we pull the lens back from the usual suspects, to understand content moderation more deeply and holistically.

In this, cross-disciplinary work is critical, since legal, engineering, communication, political and economic issues are all at stake. The set of proposals that follow comes from critical internet scholars, and we hope that it encourages a broader and cross-disciplinary conversation. As a starting point to a larger discussion, each of our authors provides a suggestion for expanding the study of content moderation, with the ultimate goal of sound policy grounded in human rights and open societies. The first group of authors considers specific applications and processes across a wider variety of platforms. The second reexamines the implications of content moderation for the health of open societies. The third considers the future of governance, particularly governmental regulation.

1. Looking beyond Facebook

Moderation everywhere

By Tarleton Gillespie

If we want to encourage a more expansive understanding of content moderation and the policies that regulate it, both the public and the scholarly debates we're having today are dangerously narrow. We need to grasp the breadth and depth of moderation, across the entire ecosystem of content provision and deep into the infrastructural stack of distribution; and we need to understand how the practices and effects of moderation are interconnected in ways that warrant greater attention.

While it deserves its share of scrutiny, too often we take Facebook—the software, the company, its methods, and its problems - as a proxy for platforms more broad-

ly. This can be deeply limiting. If nothing else, there are many kinds of social media platforms that configure content moderation differently (Caplan, 2019). Platforms also differ in ways that affect how content moderation works: by size, reach, and language, most obviously, but also by technical design, genre, corporate ethos, business model, and stated purpose. And moderation also happens on sites and services quite different from Facebook: on comment threads and discussion forums, in multi-player game worlds, in app stores, on dating sites, and on the many service, labour, and crowdfunding platforms of the gig economy.

We tend to think about moderation as being performed by and on individual platforms. But moderation, in both its practices and its effects, *overlaps* in important ways. While each platform comes up with its own rules and procedures separately, there can be a great deal of similarity in the ways these rules are phrased, and in the way violations are understood. There is intense exchange between the policy teams at major US platforms, both through the actual personnel moving from job to job, and through more informal points of professional contact. Major platforms keep an eye on each other, and in some moments even appear to act in concert—the *deplatforming* of Alex Jones only being the most striking example (Kraus, 2018). And while the actual work of moderation is generally conducted by each platform individually, there are important ways in which they collaborate. Separate platforms owned by the same parent company, like Instagram and Facebook, share resources. Smaller platforms outsource their moderation to a small set of third-party services, which deploy the same software and human teams across their clients' sites. And while antitrust law may still discourage large-scale collaboration, recent years have seen the formation of what Evelyn Douek (2020) calls *content cartels*—for instance, GIFCT, a database overseen by a set of platform companies, used to track and share content that at least one participating company has identified as terrorist propaganda (regardless of whether their criteria align).

Moderation also overlaps in the experience of creators and users. While platforms may function independently as institutions, from the user's perspective these services coexist; they are options to choose from, or to use in tandem. Many of those who develop a public voice online depend on multiple platforms and services: a political provocateur might be active on YouTube, Twitter, and Facebook; their sustained public presence may also depend on Paypal to handle payment services, Patreon for donations, Threadless to sell their merch, a Mailchimp newsletter, EventBrite for organising events, Google AdSense for their advertising. Each of these might decide that their content violates their standards. While none of these components may be essential by themselves—it is not as if I have been silenced if

I simply cannot process incoming donations or sell t-shirts—being banned by all of them at once does begin to approach something like censorship, more so than just being suspended from Twitter, at least in a way that our current thinking about content moderation has yet to address.

The moderation field is not only wide, it's also deep. Moderation decisions get made all up and down the infrastructural stack of services, often in ways that are much more opaque than the decisions made by Facebook and the like (Masnick, 2018; Plantin et al., 2018; van Dijck et al., 2019). Web hosting and site management services make decisions about which sites are simply beyond the pale—a fact made most apparent in 2017 when the CEO of Cloudflare announced he was no longer willing to host the white nationalist site Daily Stormer (Johnson, 2018), even as he publicly lamented that he shouldn't have the power to make such a decision (Prince, 2017).

Cloud computing services also moderate, not by removing particular bits of content, but by rejecting whole sites or entire platforms. Typically, services like Amazon Web Services or Microsoft Azure claim a position of neutrality, preferring not to be in the business of picking and choosing, drawing on the protections of Section 230 of the Communication Decency Act (*CDA 230*) and the sensibility of *net neutrality* enjoyed by ISPs. At the same time they reserve the right, in their terms of service or contractual agreements, to drop any client for a wide range of reasons. [3] The decisions they make do not look like platform moderation, in that they are not procedural, consistent, or accountable: most happen in the context of a specific business relationship, where a problematic client will be quietly released from their contract and urged to find another provider. This was apparent when Microsoft was accused of threatening to ban right-wing social media platform Gab after a complaint came in to customer service; Microsoft apologised for the confusion (Lecher, 2018), and Gab made noise about its rights being threatened. But later, Microsoft urged their client to leave, a move that had much the same effect (Ingram, 2018). This is content moderation, by other means.

Moderation at an infrastructural level is not only harder to see or hold accountable, it can create *stacks*, where one intermediary must abide by the rules of another, meaning users are regulated by both together, in ways difficult to discern. For

3. See for example, the "Acceptable Use Policy" (last updated 2016) for Amazon Web Services, their cloud computing service: "We reserve the right, but do not assume the obligation, to investigate any violation of this Policy or misuse of the Services or AWS Site. We may… remove, disable access to, or modify any content or resource that violates this Policy or any other agreement we have with you for use of the Services or the AWS Site."

instance, platforms that function largely in the mobile ecosystem find that their standards must be aligned with that of Apple, otherwise their app may be dropped from their App Store. A service looking to be more flexible, progressive, or permissive might find itself constrained by a more conservative or capricious infrastructural provider; and rules implemented further down the stack are even less evident to users, and less available for critique (Tusikov, 2019).

Policies that understand the stacked and overlapping qualities of moderation across the information ecosystem will be better suited to addressing more systemic problems. For example, if many service providers start making moderation decisions that align—because they see problems in the same ways, or because there is political cover in acting in concert - the entire information environment may tend towards ideological consistency. Speakers banned across the many stacked and overlapping services will experience deplatforming to a much deeper degree. On the other hand, if all the nodes on the network set their own policies, in ways that are different and changing and capricious and opaque, those with contentious things to say will face an unpredictable landscape, an uncertainty inhospitable to their ability to speak publicly.

Encryption poses distinct new problems: the case of WhatsApp

By Ariadna Matamoros-Fernández

The content moderation debate needs to expand to a broader range of popular platforms that are making important shifts in the development of social media. WhatsApp, launched in 2009 and acquired by Facebook in 2014, is paradigmatic of social media's shift towards more private, integrated, and encrypted services (Zuckerberg, 2019). WhatsApp can be understood as social media insofar as content sharing among small and large groups, public communication, interpersonal connection, and commercial transactions converge in key features of the app. The platform's rise and immense popularity in key markets like India, Brazil, and Indonesia opens up new challenges not only for content moderation on the app, but also for the regulation of platforms by governments.

First, WhatsApp has quickly evolved from a one-to-one chat service to a global social media platform, and has introduced a number of new technical fixes to the problem of *information disorder* (Wardle, 2018) and abuse. Despite encryption, WhatsApp moderates content both at the account and content level. At the account level, the company uses machine learning to detect abusive behaviour, disables over two million accounts per month, and scans unencrypted information such as

profile pictures, which has been instrumental in detecting child pornography activity within the app (WhatsApp, 2019).

At the content level, accusations of disinformation and mob violence (Newman et al., 2019; Rajput et al., 2018) pushed WhatsApp to implement measures to curb the virality of problematic messages. Unregulated virality on the app depends on a combination of WhatsApp's main technical affordances: encryption, groups of up to 256 people, and the forward function. Most of the groups on the platform are family and friends, local community and neighbourhood groups of less than ten people (Lu, 2019). In these intimate spaces, one might think that content moderation would not be needed at all. However, in countries like Brazil, Malaysia and India, some WhatsApp groups can be much larger and act as semi-public forums (Banaji et al., 2019; Caetano et al., 2018). The sharing of news and the discussion of politics are popular; this communication takes place among strangers, and context collapse may occur (Newman et al., 2019). The combination of large groups and users' unlimited ability to forward messages helped information on WhatsApp be easily shared "at scale, potentially encouraging the spread of misinformation" (Newman et al., 2019, p. 9).

Growing demands from users, activists, and policymakers that WhatsApp take greater responsibility for this glut of misinformation on its network have pushed the platform to provide more content moderation mechanisms to users. In 2018, the company limited the number of times a user could forward a specific message, from twenty to five, and it began labelling forwarded messages to help users distinguish content shared by friends and family from content forwarded from elsewhere (Hern, 2018). WhatsApp has also begun labelling messages forwarded more than five times as "Frequently forwarded". In August 2020, the company launched a pilot feature in seven countries including Brazil, Mexico, and the US: a magnifying glass icon next to messages that have been shared more than five times allows users to upload these links via their browsers. The feature is meant to facilitate users' ability to fact-check viral messages outside the app (WhatsApp, 2020). And, amidst the Covid-19 *infodemic* (World Health Organization, 2020) and illustrating a classic *governance by shock* move (Ananny & Gillespie, 2016), WhatsApp imposed a new limit so that messages shared many times could only be forwarded to one chat at a time (WhatsApp Blog, 2020).

Forwarding messages decreased by 25% globally as a direct result of the implementation of some of these measures (WhatsApp, 2019). These claims, though, cannot be validated by external research, since only WhatsApp has access to behavioural patterns on the app—from group size to viral media. More research on

users' experiences with problematic messages on the app, and their responses to WhatsApp's new features to curb the spread of abuse and misleading content, is needed in order to evaluate the efficiency of these measures and explore alternatives to the challenges of content moderation on this platform.

Second, encryption also complicates the debates around how to regulate platforms in order to tackle the circulation of problematic content in digitally mediated spaces. And again, WhatsApp offers the most striking example. Serious offences on WhatsApp, such as terrorist and criminal activity, have led governments to pressure Facebook to provide 'backdoor access' to encrypted messages on the service. But breaking encryption in this way would compromise activists using the app to circumvent state censorship and surveillance (Johns, 2020; Treré, 2020). Australia alone has succeeded in passing a law that gives the government new powers with regard to snooping encrypted data. The Telecommunications and Other Legislation Amendment (Assistance and Access) Bill 2018 (Parliament of the Commonwealth of Australia, 2018) requires digital intermediaries to provide federal police and intelligence agencies with "technical assistance" to access encrypted conversations from criminal suspects. Other countries may follow. India is drafting a new law on data protection that contemplates different exceptions for the processing and hosting of user data by the Indian government (Sharma, 2019). These examples put WhatsApp and other encrypted platforms, such as Telegram and Signal, at the centre of global debates around content moderation, user privacy, security, and freedom of expression. The pervasiveness of encrypted platforms in mediating everyday life in some parts of the world is a reminder that viable content moderation measures without breaking encryption are needed. Some argue that automated content moderation for encrypted services could be done locally on users' devices without requiring 'backdoor' access and compromising security (Mayer, 2016; Reis et al., 2020). But the development and implementation of these measures will require joint efforts between platforms, civil society, and governments.

'Too good to be true': the challenges of regulating social media startups

By Ysabel Gerrard

In the content moderation debate, we need to carefully consider how best to regulate social apps that are either new or have become too big too quickly. New apps that lack substantive planning or an institutional apparatus for content moderation often become so overwhelmed by problems like cyberbullying and unsolicited explicit image sharing that they collapse, either because they are quickly removed

from app stores or because their founders shut them down. On the other hand, viral success, without internal processes that adjust to legal and regulatory requirements, raises the risk of real harm to social media users.

Failed social media app Fling can help highlight this problem. Founded in 2014, Fling invited users to send photos and videos to complete strangers around the world. Fling was an instant hit, but the app was quickly overrun with pornographic images designed to harass female users, leading to its removal from Apple's App Store (Nardone, 2015). Despite employing "a full-time team devoted to keeping dicks at bay" (Nardone, 2015) and spending "all hours of the day and night trying to build a new version of the app that met Apple's guidelines" (Shead, 2017), Fling could not recover. As Fling's founder and former CEO put it, the app's rapid success "seemed too good to be true…and it was" (Nardone, 2015). Content moderation challenges were not the only problem at Fling (see Shead, 2017), but the app's failures highlight the many uncertain, expensive, and risky steps that startups like Fling feel they must make in their early days, and which might make heavy regulatory obligations hard to meet.

The anonymous app Secret offers another example. Founded in 2017, Secret allowed users to anonymously share a "secret" with friends who also used the app. Secret was extremely popular with young people and was at one point the most-downloaded app in eight countries (Katz, 2017), but users' bad behaviour simply became too much for the team to manage. Secret's former CEO David Byttow said his team "couldn't contain it, could not control it" (cited in Katz, 2017). Byttow shut the app down in 2015.

This is a common origin and downfall story for many social media startups: platforms like Fling and Secret do not collapse because they are unpopular, but because they are *too* popular, leaving their founders unprepared for the scale of content moderation necessary. Apps like Fling and Secret thus pose distinct challenges to content moderation policy because they are often *popular-by-surprise*; that is, their founders do not expect them to reach such dizzying heights of success so quickly. At the height of their popularity, some new apps rival their more "traditional" competitor platforms in number of users. But at new companies like Fling, start-up sized workforces—especially content moderation teams - might be woefully inadequate.

Effective content moderation requires a great deal of knowledge and expertise, and the spectacular failures of some social media startups suggests that this knowledge is often gained too late, or not at all. Moreover, not all tech startups as-

sume they will make money early on (Crain, 2014), which means the moderation knowledge and staffing problems that come with popularity-by-surprise are baked into current business and growth models. As Crain explains, growth—"investing in technology, and buying competitors" (2014, p. 379)—often comes first, and profit follows. And startup apps are also very much at the mercy of Apple and Google. As Suzor explains, given the legal reality that the companies behind social media platforms "have almost absolute power over how they are run" (2019, p. 11), it is digital intermediaries like app store owners that often decide when a social media startup becomes too unsafe.

This presents anyone who is invested in content moderation with some questions: what would be appropriate and effective content moderation processes for social media startups? Do startups warrant looser obligations than the more established players, or tighter ones? And would one-size-fits-all regulation, intended to apply to platforms like YouTube and Facebook, stifle startups before they could start, or produce a minimum expectation that would prevent unprepared apps like Fling and Secret from launching?

At present, no countries have (to my knowledge) laws requiring social media startups to have content moderation workforces at all, or for them to take a particular shape. But attention to and regulation of social media content moderation is growing. For example, the United Kingdom's Online Harms legislation proposes holding online businesses to account for harmful content: for failing in their *duty of care* (Department for Digital, Culture, Media and Sport, 2020). Meanwhile, Germany's Network Enforcement Act (*NetzDG*) forces "platforms to ensure that 'obviously unlawful content' is deleted within 24 hours" (Heldt, 2019, p. 1).

Regulatory obligations like these create two issues for social media startups. First, any law requiring a social media company to be responsive, especially within a certain timeframe, surely more or less requires that there be someone present to *do* the responding. The stories of app failures told above, and the increasingly global push for greater regulation, suggest a need to regulate startups differently to the big players. Second, fining companies for failing in their duty of care would likely kill off some *popular-by-surprise* apps (for a similar argument, see Hern, 2019).

The history of online regulation is repeatedly told with an eye to the big players. As Gorwa (2019a, p. 10) notes, it might have seemed 'bizarre' that Ask.fm—a site designed for asking questions anonymously, and which got the blame for a number of teen suicides (Henley, 2013) - was part of the European Commission's *EU In-*

ternet Forum in 2014, alongside Facebook, Google, Microsoft, and Twitter. But smaller, explosively popular apps can be useful regulatory cases too, even if they have already failed.

The focus on established social media companies means that both the debate and the regulations it produces are likely to be ill-suited to the growing pains of flash-in-the-pan apps. Do startups need looser regulations so they can grow? Or do they need stricter rules, because a lack of regulation might make their users more vulnerable to harm? It depends on what you are asking, and regulation that is attentive to this category of companies has to wrestle with both possibilities. But policies that only imagine established platforms like YouTube and Facebook are never going to accommodate the unique challenges popular-by-surprise apps present. To address this, policymakers could commit to supporting startups when they grow faster than planned and do not have enough staff for content moderation, or could provide new social apps with the knowledge and expertise to develop better, safer systems.

2. Moderation, community, and democracy

Democracy cannot survive algorithmic content moderation

By Aram Sinnreich

Any policy meant to improve social media content moderation must attend to how moderation is actually done. Other scholars (Carmi, 2019; Caplan, 2019; Roberts, 2019) have highlighted the complexities of affective and immaterial aspects of the human labour involved in content moderation. I would like to address the flip side of this concern: namely, the social and political challenges emerging from algorithmic content moderation practices, and the potential risks involved when we delegate these kinds of decisions to artificial intelligence and other nonhuman processes. I believe it is essential to include these kinds of questions in any debate about content moderation. To ignore them would be a tacit acceptance of the legitimacy of these systems, without adequate critical investigation. Furthermore, laws or regulations that fail to anticipate and address proactively the broader implications of automated content moderation will—like all policies that overlook second-order and longer-term consequences—likely create as many problems as they solve.

For years, algorithms have been promoted as essential to content moderation practices, from the automatic facial anonymisation algorithms on Google Street

View (Ruchaud & Dugelay, 2016) to the proposed (but yet-to-be-implemented) automated copyright policing outlined in Article 17 of the EU Copyright Directive (Bridy, 2020; Quintais, 2020). Additionally, algorithms are often treated as a failsafe for times and situations in which human content moderators are unavailable or prohibitively expensive. For instance, during the Covid-19 pandemic of 2020, Facebook opted not to delegate some of its sensitive moderation practices to American workers confined to their homes, and instead chose to police certain violations of its terms of service (e.g., pornography, terrorism, and hate speech) using automated systems (Dwoskin & Tiku, 2020; Thomas, 2020). At the time of writing, it is unclear whether and when these content moderation labour forces will return to their pre-pandemic employment levels.

The most obvious deficiency of automated content moderation is the greater risk of false positives and negatives: an educational video about breastfeeding may be mislabeled as pornography, while weaponised disinformation may be promoted in the same way as a news article from a reputable source. But the greater risk of algorithmic content moderation comes in its tacit threats to democratic norms and institutions (Sinnreich, 2018). There are several mechanisms by which this may happen.

Quantisation of culture. It is the inevitable result of delegating nuanced and contextualised cultural decision-making (and meaning-making) processes to an algorithm. The fair use doctrine of American copyright law, establishing whether infringement has occurred, requires a judgment call about the context of use, and is thus famously resistant to algorithmic assessment (Aufderheide & Jaszi, 2018). Similarly, US Supreme Court Justice Potter Stewart famously declined in 1964 to define "obscenity", arguing that, in the absence of a hard and fast rubric, "I know it when I see it" (*Jacobellis v. Ohio*, 1964). In other words, our definitions of permissible and impermissible expression depend upon human judgment, rooted in cultural context. They were designed to be so. If we delegate these decisions wholly to algorithms, not only will they make *wrong* decisions, out of keeping with organic cultural values, they will have the effect of reifying algorithmic logic as the arbiter of meaning and legitimacy, displacing these organic processes epistemologically and impoverishing our cultural spheres. Some, such as Shoshana Zuboff (2018), even propose we are seeing an emergent *surveillance capitalism* that will create our futures, with *predictive products*.

Institutional convergence. A cornerstone of functional democracy is the separation of powers (Vibert, 2007). At every level of governance, this means that the responsibilities for making laws, evaluating lawfulness, executing laws, and meting out

penalties are delegated to different institutions, operating independently of one another. Platform content moderation, however it is performed, violates this principle by delegating virtually all legislative, judicial, and executive functions to a single, unaccountable, privately-controlled entity. That entity itself is of course subject to judicial oversight and statutory limits, and the algorithm is frequently developed in response to the parameters set by such legal requirements. In some cases (for instance, Facebook's turn to the non-profit news outlet Correctiv in Germany) third parties might be invoked for particular services. But the corporate entity still gets to set the terms.

When this moderation is performed via algorithm, even the fig leaf of ethical human oversight and accountability is removed from the process (Gorwa et al., 2020). Not only can this produce a discriminatory public sphere in which open debate cannot possibly flourish, but like the quantisation of culture, this process actually threatens to undermine democratic epistemologies. In other words, when consumers accept automatic oversight of platforms as an appropriate form of governance, it paves the way for citizens to accept autocracy as an appropriate form of government. And of course, as Julie Cohen meticulously shows, platforms are eagerly seizing governance by exploiting neoliberal loopholes in today's governing institutions (Cohen, 2019).

Expansion of scale. Laws, and the cultural values that shape both statute and jurisprudence, are local, regional, and national in scale. Platforms, however, are global. When we delegate important decisions about public speech to algorithms operated by these platforms, we undermine national sovereignty and self-determination, creating the conditions for corporatocracy and monoculture. Unlike constitutional governments, corporations have no duty to uphold democratic values; in fact, their fiduciary duties often run counter to these values. Some (e.g. Kaye, 2019; Jørgensen, 2019; Suzor, 2019), are looking to global values such as human rights to structure a counter-movement to this threat.

In short, whatever challenges are posed by human labour being exploited for the purposes of platform content moderation, we must not lose sight of the equal or greater challenges we face when those decisions are delegated to opaque, unaccountable, privately owned algorithms. Nothing less than the future of democratic society is at stake.

Thinking beyond content in the debate about moderation

By Sarah Myers West

Content moderation may be too limited a term to describe the kinds of actions platforms take in response to user behaviour. As this issue rises on the policy agenda, alongside emerging concerns about hate speech, extremist content, the spread of misinformation, and online abuse, the content moderation discussion has largely centred on its speech dimensions—as being primarily about *content*. Given the central role platforms play in mediating our social, economic, and political lives, we are in need of a more expansive way to account for how platforms enact their policies, and how users struggle with them.

In research I conducted with users whose social media accounts had been suspended for community guidelines violations, the impacts that stood out most had little to do with their speech (West, 2018). Users felt cut off from their loved ones by the suspension of their account, because Facebook was their primary means of communication with far-flung relatives. Disabled users worried that if their health deteriorated, nobody in their network would know if they didn't have access to their account. Artists reported their follower counts were wiped out entirely after access to their Instagram accounts was restored, losing the primary means through which they sold their work. Alternative media outlets reported being unable to function when their admins received automatic 30-day bans from the platform they published on. And users found they were cut off from not only the platform they were banned from, but also the third-party login infrastructures that rely on Facebook and Google logins. Ysabel Gerrard (Gerrard 2018; McCosker & Gerrard, 2020) similarly points to how companies using commercial content moderation struggle when they encounter users' experiences with their own behavioural and mental health challenges.

The content moderation debate should expand beyond treating platforms as primarily venues for public speech, or as silos that exist in isolation from one another. Instead, we might think of them as a web of private infrastructures that we traverse in our digitally mediated lives. For years, they have done much more than offer us platforms for communication and the sharing of user generated content (Steinberg, 2020; Nieborg & Helmond, 2018): they are also marketplaces, payment systems, advertisers, gaming sites, and media distributors (Swartz, 2020; Acker & Murthy, 2018; Napoli & Caplan, 2017; Leaver & Wilson, 2016; Nieborg, 2015). In so doing, they have enmeshed themselves with the wider web and with deeper layers of existing non-digital infrastructures, acquiring a scale and indispensability that requires us to shift the kinds of questions we ask about them (Helmond, 2015; Plantin et al., 2018; Plantin & Punathambekar, 2018). Because the business models of the companies that run these platforms rely on the acquisition and integra-

tion of multiple service offerings, our activities in any one node can have reverberating effects across a wide web of private infrastructures. Perhaps the debate around moderation would be more robust if we thought about platform decision-making in a way that better acknowledges this impact. Rather than thinking about content moderation in terms of its effects on speech alone, we should instead consider the ripple effects that moderation can have on the social fabric of our communities, by influencing our access to platforms that are increasingly central to our ability to work, live, and thrive together.

Out with the deviant, in with the social

By Elinor Carmi

What kind of world do you want to live in? This is the question that guides platforms when they produce the (commercial) spaces we use every day. The biggest platforms (e.g., Youtube, Facebook, TikTok, Instagram) use commercial content moderators (Roberts, 2019) to shape this world every day, according to ever-changing guidelines. But content moderation is certainly not the only way platforms shape our information worlds. Examining what content moderators do and what it means more broadly to the field of media and communication requires tuning into a larger question about how our mediated experience is shaped. In my book *Media Distortions* (Carmi, 2020a), I show how content moderators are part of a longer history of filtering between the noise and the signal—a decision-making process discerning what is deviant and what is the norm. While *deviant* may be interpreted differently depending on the medium, company, country etc., it stands for the way media companies want to decide for us how we should experience and perform a certain kind of sociality. I call this practice *rhythmedia* (Carmi 2020b)—the way media companies render people, objects, and their relations as rhythms and (re)order them for economic purposes, to produce a certain kind of sociality. Consider two examples.

Going back to the early 20th century, what was considered deviant was often termed "noise". In 1929, Bell Telephone Company joined the Noise Abatement Commission (NAC) to measure New York City and create a noise map (Thompson, 2004). The people who formed this Commission came from various interest groups, like real estate and insulation companies, who wanted to reorder the city to become more commercially oriented. Bell measured spaces and people across the city with its own tools and units to establish what is sound (normal/healthy) and what is noise (abnormal/sick), to create a dynamic database (I call this practice *processed listening* [Carmi, 2019]). The people who were categorised as noisy in-

cluded African-Americans, street pushcarts (mostly "foreigners"), and union protesters—in other words, people whose behaviours harmed the business models of the NAC/Bell. The NAC conducted rhythmedia by orchestrating the rhythm of the city - the way all the components (people and roads, buildings and automobiles) were tempo-spatially ordered (Carmi, 2020).

Almost a century later, what is considered deviant is termed "antisocial". Social media companies use various tools, both human (commercial content moderation) and nonhuman (algorithms, interface/default design), to shape our experiences in their platforms. Recently we learned that TikTok's guidelines instructed their content moderators to filter out people and spaces who are deviant (Biddle et al., 2019). The company argued that this approach was meant to prevent bullying, but the leaked documents suggest it was to attract new users, because deviant posts "decrease the short-term new user retention rate" (ibid). The platform conducts *processed listening* for everything that is happening in its space, and if you are old, ugly, disabled, fat or activist, then content moderators would algorithmically punish you by suspending or even permanently banning your account, or by narrowing the distribution of your post (Kuo, 2019)—muting your voice. If you live in poor areas, like slums, you would also be silenced from what other people can listen to in order to "retain new users and grow the app" (ibid). In other words, you would be excluded from what the platform considers as social, because your appearance, condition, opinion, or environment do not yield more profit for the company. But filtering important information and people from social media has serious consequences on our health, accommodation, and politics. It shapes our experience and how we understand the world around us.

What does it mean for us? As media scholars we need to expand our research beyond platforms and understand what these actions mean for the wider politics of commercialising our public spaces, especially with the *Internet of Things* and *smart cities* being heralded as the future. Moreover, the myth of these systems being completely automated is a lie - what Astra Taylor (2018) terms *fauxtomation*. There is always a decision-making process behind categorising behaviours and people as deviant, but so far these have been hidden and unaccountable. What does amplifying it mean for our autonomy, understanding, and imagination? Finally, to establish what is considered deviant, media companies have normalised tracking, measuring, categorising, and recording our mediated lives. We do not know how these archives can and will be used to categorise some of us as deviant, but we need to demand and create a different setting for our mediated lives.

3. The future of regulating content moderation

Content moderation has a regulatory politics

By Robert Gorwa

In the past few years, scholars, activists, and journalists have done invaluable work that has helped us further understand the complex infrastructures of content moderation (Gillespie, 2018; Roberts, 2019; Suzor, 2019). But even as conversations about governing content on platforms have captured the public conversation, rapidly becoming one of the most talked-about aspects of information policy in Europe and North America today, our understanding of the political and regulatory dynamics around content moderation remains limited. How are the content policy processes of companies like Facebook affected by pressure from policymakers, and shaped by regulatory commitments made in various jurisdictions? What are the strategies that policymakers use to get firms to change their rules, either regionally or globally? What are the factors that determine the success of these efforts? These are critical questions that will require interdisciplinary policy and legal work as content moderation continues to become a hotly contested global public policy issue.

A more explicitly political research programme that foregrounds public policy and regulatory studies should help us better understand content moderation as political relationship between what I have previously outlined as a *platform governance triangle* of political actors (Gorwa, 2019a): individual firms and industry associations; non-governmental civil society groups, individuals, journalists, and researchers; regulators, policymakers, and various political institutions. To what extent does a government's ability to shape private content policies hinge on market power, as opposed to their regulatory capacity or their ability to effectively mobilise coalitions of consumers (Drezner, 2008; Culpepper & Thelen, 2019)? Similarly, there is so much that we still don't know about the dynamics of how platform companies exert influence politically — how they lobby, build advocacy campaigns, and leverage other tactics to push back against policymakers, and how these strategies might mirror or differ from what large multinationals in other industries have done for over a century (Hofferberth, 2019; Mikler, 2018; Fuchs, 2009). These are not merely academic points for debate: they have major policy implications for citizens, civil society, and policymakers. What are the civil society strategies that work best as they fight to push content regulation in a more rights-preserving direction (Jørgensen, 2019)? What strategies could countries in the Global South pursue to ensure adequate justice for their citizens within processes shaped predomi-

nantly in the US and the EU?

These complex questions fall outside the scope of what has thus been the mainstream of content moderation research. They will likely require a consideration of domestic political factors, transnational linkages among policymakers, changes in international governance institutions, and other trends in modern, global, regulatory governance. Such questions far exceed the traditional wheelhouses of scholars trained in communication, digital media, or platform studies. Nevertheless, as more governments implement regulatory measures—or steer voluntary and informal standards setting organisations that are operated by industry (see Douek, 2019; Gorwa, 2019b) - the policy ecosystem that affects how companies design, implement, and enforce their content standards has become increasingly complex. If we wish to expand the scholarly debate around content moderation to better keep up with these developments, we should seek to grow this interdisciplinary conversation further, recruiting political scientists, regulatory politics researchers, and public policy academics and practitioners to join the digital media researchers and legal scholars that have driven the discussion thus far.

Commercial content moderation is a soft economic and political tool

By Sarah T. Roberts

Content moderation as regulated under CDA 230—the US law that both exempts ISPs from legal responsibility for their users' actions and also permits them to moderate on their platforms without losing that exemption—is a tool of soft power for US-based global firms intent on proliferating that standard worldwide.

Issues in content moderation include labour conditions (Roberts, 2019), user rights (Langvardt, 2018), and globalisation (Uy-Tioco, 2019). But there is also a fight underway for control between firms and governments. Regulatory debates follow on the recognition that content moderation has significant human rights and geopolitical implications (Banchik, 2020). American firms' indemnity for the digital material flowing through their branded services under CDA 230 is now central to this regulatory debate. It gives them incredible discretion, because they are free from the liability experienced by most media firms (Sylvain, 2019). Recent Congressional hearings have revealed that the firms and their allies (such as advocacy organisation Electronic Frontier Foundation), have lobbied for or supported the inclusion of CDA 230-like language in transnational trade agreements. This could greatly expand the scope of the policy, at a time when it is publicly being questioned in the

United States (Nix & Kern, 2019). This expansion of indemnity represents an exponential leap in the power and reach of US-based social media firms.

Meanwhile, just as the public has become aware of content moderation, it has also grown in importance for the digital services that require it. Just a few years ago, the major Silicon Valley social media firms still relegated large-scale moderation of user-generated content (UGC) to an afterthought, at best (Chen, 2014). One issue that C-suite denizens were most likely to avoid was the way the practice of moderation, both mission-critical (from a brand protection and advertiser relations perspective) and also one of the most stigmatised parts of their media production chain, puts the lie to the claim that these global social media firms were mere engines of free speech. Firms avoided conversations about content moderation whenever possible, choosing instead to wow a largely co-opted tech media sector with the latest advancement in functionality (Stewart, 2017).

Yet it became increasingly clear, through the work of journalists, academics and whistleblowers, that moderation decision-making actually constituted something so potent that law professor Kate Klonick described the platforms as the *new governors* (2018). At the same time, by keeping content moderation largely invisible to most users, these platforms constructed a social media ecosystem that most users had no idea was mediated. Ultimately, low-status, low-wage, and easily-replaced content moderators, working in the heart of Silicon Valley, were some of the first to make connections between their work and its implications (Roberts, 2017). Without the push from civil society, academics, and reporters to take up their revelations and make content moderation more transparent (Crawford & Gillespie, 2016), the impact of such decisions would have never been rendered so clearly.

The politics of platforms are now subject to much greater public debate. Regulation is no longer so unfathomable, with some regions of the world (e.g., the EU and its member states) much more aggressive toward social media than others (Knight, 2018). In late 2019, the US Congress began to revisit CDA 230. At a 2019 hearing, attorney Katherine Oyama of Google told legislators that Google and other firms were adding CDA 230-like clauses to antidemocratic and typically secret covenants such as multinational trade agreements. Rather than see CDA 230 fade into irrelevance, it seemed that the firms had found a way to expand its reach (House of Representatives, 2019).

This constellation of internal policies, commercial practices, and US law has allowed American social media firms to proliferate to a global scale, with consequences for the rest of the world. American mores, jurisprudence and norms, con-

flated with what is best for American *firms*, have been seamlessly packaged into technological and policy affordances, bolstered by content moderation favourable to the social media companies themselves. A uniquely American, even Silicon Valley outlook has primacy in this arrangement, supported by a Washington, D.C. establishment with a revolving door in and out of the tech industry's lobbying cadre. That cadre is gearing up not only for a battle against antitrust (Ryan, 2019), but also to stave off attacks by other major economic forces such as Disney and IBM, who consider CDA 230 an unfair carve-out for the social media industry (McCabe, 2020). This status quo is unlikely to withstand the next few years, but until change comes, content moderation and all its parts must be understood in the frame of the soft power, hegemonic force that it is. Long supported by the discretion and indemnity provided by CDA 230, its force and impact will only become more powerful and simultaneously less accountable to public scrutiny if it moves into secret US trade agreements.

A regulatory turning point? The power of protest

By Patricia Aufderheide

We can learn from past movements intended to regulate savage capitalism, as we work toward new policies regulating today's digital platforms.

The power of platform moderation to act effectively as a regulator of daily life is now evident. This lands us firmly in the realm of policy—the question of what steps we take at a societal level to address the societal problems caused, inadvertently or not, by private actors. Corporate actors are already there, grappling daily with the opportunities and constraints of the political forces facing them, given the challenges their runaway inventions have created for them, and for us. We can also see the venues where action is and can be taken. There are so many: states, international agencies, multistakeholder organisations, standards bodies, public interest organisations and corporations themselves all want and have a role. As yet, as Suzor (2019) has noted, pressing for a human rights values framework, they largely have not expressed, much less agreed upon, any common values that should drive these regulatory processes—other than enthusiasm for innovation and (as Cohen [2019] has analysed) a neoliberal regulatory bias.

Is this a moment of regulatory pivot—possibly even, to invoke a term Karl Polanyi (1944) gave us, the beginning of a *double movement*? Polanyi described a *great transformation* in English society, as "rational" market principles came to infuse culture, with the state combining with economic institutions to fundamentally change

the terms of daily life. The *double movement* was a regulatory pushback against its excesses, triggered by combined corporate and governmental concerns to maintain industrial capitalism. Movements to strengthen workers' rights, control pollution, govern growth, and protect public health all threatened the future of industrial growth, without such regulations. None of the policies that resulted from these movements went unassailed by corporate forces eager to retain their autonomy, and all of them were sadly less-than. Indeed, some argue that the *double movement* is ultimately nothing more than capitalist forces saving themselves (Cooper, 2017). But these efforts did fundamentally change both how industrial capitalism developed and what people expected of businesses and of government.

Perhaps we are approaching the *double movement* for this era, as Cohen (2019) suggests. We may be arriving at that point, she argues, in part through the tight relationship between neoliberal politics and corporate fecklessness, creating enough social dysfunction to trigger demands (however inchoate) for pushback. We have already begun to see the first regulatory steps to address what Srnicek (2016) calls *platform capitalism*, but they are piecemeal. And, as is typical of political processes generally, they respond to crisis and have pushback of their own. Thus, Snowden's revelations have driven corporate decisions to invest more in security/encryption, which then created new privacy challenges, which aroused further legislative scrutiny. So far the debate about whether to rewrite or delete CDA 230 has largely revealed the confusion and ignorance of legislators and pundits about the workings of both internet business and internet governance, but it will become more nuanced.

If we are approaching such a moment, we can also look to the past for clues. The changes spurred by the Great Transformation's *double movement* (Polanyi, 1944) were not the acts of a wise, caring, and forward-looking government. Rather, they resulted from social movements, public scandals, and political demands that society not be subsumed into an economy. They were multifarious acts combining to become unignorable counterpressure. They were not coherent ideologically. Catastrophic economic failures and man-made disasters served as triggers. Smaller versions of this phenomenon have surfaced again and again in history—for instance, in the US with an early 20th century antitrust movement and policies that reined in the most egregious of the *robber barons* (Wu, 2018) and with the New Deal. During the 1930s, the US federal government responded not only to the economic catastrophe of the Great Depression but also to political mobilisation and public outrage demanding fundamental governmental and social change. In the end, society settled for regulatory measures that tempered savage capitalism in the US.

If we seek values-driven regulation that tempers the fecklessness that communications platforms now freely exhibit, then academic researchers, writers, journalists, and public interest organisations have their work cut out for them. Efforts such as the Santa Clara Principles, [4] currently under revision, are a small but welcome first step. The Internet Engineering Task Force's internal struggles about how to interpret human rights values in standards-setting for internet protocols are important, and deserve more attention—see, for instance, its creation of an HTTP status code, called 451 (named for Ray Bradbury's novel about censorship), indicating that material searched for was unavailable for censorship reasons (Cath & Floridi, p. 460). Facebook's Oversight Board, already under close scrutiny, could be an important test case in establishing privately-ordered norms in content moderation.

The evolving academic discussion of content moderation, figured not as a bad-actor problem about big platforms doing wrong but as a complex issue deserving informed analysis, is important in the mix of activities that could drive toward a *double movement*. Legislators and regulators deserve to get well-informed research, translated into usable, even actionable bullet points, along with the routine deluge of corporate lobbying.

Conclusion

By Tarleton Gillespie and Patricia Aufderheide

Together, these observations suggest some important aims for moving this debate forward.

It is vital that we expand the debate about content moderation. First, this means drawing lessons, insights, and suggestions from platforms other than Facebook; conversely, it means designing policy architectures that are suited to more than just Facebook. If moderation is endemic to the provision of information and happens everywhere across the network, then any solutions social media companies, scholars, or lawmakers propose—both to pernicious online harms and to the missteps of platforms in addressing them—must appreciate their effect on the broader ecosystem. This certainly means tailoring different obligations to services of different scales, purposes, and designs. But it also means recognising the peculiar dynamics of moderation across the whole of the information ecosystem.

We need more thorough study of the impact of content moderation on different

4. https://santaclaraprinciples.org/

geographical, political and cultural communities. Research that focuses on *users* and takes either a platform-centric view of who those users are, or leans on the convenient research subject populations of university undergraduates or Mechanical Turk workers, will fail to apprehend how differently site policies land for different subcultures, linguistic communities, political tribes, and professions. It also means moving beyond a *speech* framework, to think about impact in terms of opportunities, values, ideologies, representations, norms, and cultural flourishing.

Innovations in machine learning (ML) and automated content moderation have focused overwhelmingly on identification techniques: can software spot pornography, harassment, or hate more accurately or more quickly than a human? Not only are there problems with these ambitions (Gillespie, 2020), but this work has crowded out other possible uses of ML and software techniques to support other dimensions of content moderation. Research should prioritise tools that might *support* human moderators, community managers, and individual users, to better apprehend the contours of existing norms or the risk of certain patterns of behaviour, so that moderators and volunteers can make more informed decisions for themselves and their community. Data-scientific techniques might also help users and community managers better grasp how differently other communities experience similar content or behaviour, giving empathy and civic responsibility the support of data.

As much as platform moderation could improve, it may also be a perennially impossible task to do in such a way that no one encounters harm, friction, or restriction. Users of social media may have unreasonably high hopes for what their experience should be, largely because of the endless promises made by social media platforms that it would be so. We need to educate and adjust the expectations of users, to both understand what a difficult and vital process this is, to demand it be transparent and accountable, to recognise how they are implicated in it, and to prod their sense of agency and ownership of these sometimes unavoidable dilemmas.

If regulators, researchers, journalists, and policy analysts are to address current and potential challenges in content moderation, they will need to build on empirical knowledge about actual corporate behaviour. Getting reliable empirical data from companies is a perpetual regulatory battle, in any realm. Both corporate imperatives and companies' very real need to stay ahead of bad actors militate against transparency. Today's transparency reports are an impoverished first step, the smallest of gestures in the right direction. In the same way that universities' ethics boards and institutional review boards protect human subjects, it might be

possible to design protective mechanisms for information sharing with certain categories of researchers; generations of cybersecurity research have shown that. Developing skills in this area would also strengthen platforms' ability to protect its users. As well, a regulatory requirement for some transparency across all platforms would perhaps in a salutary way change business conditions for all.

Any regulatory proposal needs to acknowledge, at some level, the public-utility-like nature of platforms (Rahman, 2018). Platforms now act, for most people, as some kind of essential service. They provide, usually with monopoly power, a function people have come to depend on for myriad daily needs. Regulators are the representatives of the public, defending their interests as a social entity rather than just as individual consumers. The public is not only the sum total of people who individually are jointly affected by governmental and corporate actions and who act on their representatives to address problems, but also the public of tomorrow, whose present will be shaped by ours (Dewey, 1927). Platforms typically provide what most of their users think of as public access or a public space of discourse, while they retreat from responsibility for that role in their guise as private actors. If they are providing, across jurisdictions, what comes down to essential information services for people, they need to be treated as such. This would be a gigantic step away from today's regulatory practices; even obvious public utilities, such as electricity and telephone services, have been decoupled from many public utility responsibilities. But it only takes one crisis, such as Covid-19, to awaken the public to their profound dependence on private actors to connect them. Finally, any regulatory structures need to be normatively designed around values for just and open societies, and for the human rights values that enable them.

ACKNOWLEDGEMENTS

Our thanks to the organisers of AoIR 2019, and to all those who attended our roundtable sessions there. The vibrant discussion deeply informed the work presented here. Our thanks also to Nicolas Suzor, who offered superb contributions to the roundtable, but was unable, because of contractual conflicts, to contribute to the written version. As well, we thank our editors and reviewers at *Internet Policy Review*, for their thoughtful, constructive, and encouraging feedback.

References

Aufderheide, P., & Jaszi, P. (2018). *Reclaiming fair use: How to put balance back in copyright*.

University of Chicago Press.

Banaji, S., Bhat, R., Agarwal, A., Passanha, N., & Pravin, M. S. (2019). *WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India* [Report]. London School of Economics and Political Science. http://www.lse.ac.uk/media-and-communications/assets/documents/research/projects/WhatsApp-Misinformation-Report.pdf

Bridy, A. (2020). The Price of closing the 'value gap': How the music industry hacked EU copyright reform. *Vanderbilt Journal of Entertainment & Technology Law*, *22*, 323–358.

Caetano, J. A., de Oliveira, J. F., Lima, H. S., Marques-Neto, H. T., Magno, G., Meira, W. Jr., & Almeida, V. A. F. (2018). Analyzing and characterizing political discussions in WhatsApp public groups. *arXiv*. http://arxiv.org/abs/1804.00397

Caplan, R. (2019). *Content or context moderation? Artisanal, community-reliant, and industrial approaches* [Report]. Data & Society. https://datasociety.net/library/content-or-context-moderation/

Carmi, E. (2019). The Hidden listeners: Regulating the line from telephone operators to content moderators. *International Journal of Communication*, *13*, 440–458. https://ijoc.org/index.php/ijoc/article/view/8588/0

Carmi, E. (2020a). *Media distortions: Understanding the power behind spam, noise and other deviant media*. Peter Lang.

Carmi, E. (2020b). Rhythmedia: A Study of Facebook immune system. *Theory, Culture and Society*, *37*(5), 119–138. https://doi.org/10.1177/0263276420917466

Cath, C., & Floridi, L. (2017). The Design of the Internet's architecture by the Internet Engineering Task Force (IETF) and human rights. *Science and Engineering Ethics*, *23*(2), 449–468. https://doi.org/10.1007/s11948-016-9793-y

Cohen, J. E. (2019). *Between Truth and Power: The Legal Constructions of Informational Capitalism*. Oxford University Press. https://doi.org/10.1093/oso/9780190246693.001.0001

Cooper, M. (2017). *Family values: Between neoliberalism and the new social conservatism*. Zone Books.

Crain, M. (2014). Financial markets and online advertising: Reevaluating the dotcom investment bubble. *Information, Communication and Society*, *17*(3), 371–384. https://doi.org/10.1080/1369118X.2013.869615

Crawford, K., & Gillespie, T. (2016). What Is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Dewey, J. (1927). *The public and its problems*. Henry Holt.

Dwoskin, E., & Tiku, N. (2020, March 24). Facebook sent home thousands of human moderators due to the coronavirus. Now the algorithms are in charge. *The Washington Post*. https://www.washingtonpost.com/technology/2020/03/23/facebook-moderators-coronavirus/

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12), 4492–4511. https://doi.org/10.1177/1461444818776611

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Gorwa, R. (2019). The platform governance triangle: Conceptualising the informal regulation of

online content. *Internet Policy Review*, *8*(2), 1–22. https://doi.org/10.14763/2019.2.1407

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Hern, A. (2018, July 20). WhatsApp to Restrict Message forwarding After India Mob Lynchings. *The Guardian*. https://www.theguardian.com/technology/2018/jul/20/whatsapp-to-limit-message-forwarding-after-india-mob-lynchings.

Hofferberth, M. (Ed.). (2019). *Corporate actors in global governance: Business as usual or new deal?* Lynne Rienner.

House of Representatives. (2019). *Hearing on 'fostering a healthier internet to protect consumers'*. House Committee on Energy and Commerce. https://energycommerce.house.gov/committee-activity/hearings/hearing-on-fostering-a-healthier-internet-to-protect-consumers

Jacobellis v. Ohio, 378 U.S 184, (United States Supreme Court 1964).

Johns, A. (2020). This will be the WhatsApp election': Crypto-publics and digital citizenship in Malaysia's GE14 Election. *First Monday*, *25*(1–6). https://doi.org/10.5210/fm.v25i12.10381

Johnson, S. (2018, January 16). Why Cloudflare let an extremist stronghold burn. *Wired*. https://www.wired.com/story/free-speech-issue-cloudflare/

Katz, M. (2017, October 17). These failed apps discovered a hidden rule of the Web. *Wired*. https://www.wired.com/2017/03/these-failed-apps-discovered-a-hidden-rule-of-the-web/.

Klonick, K. (2018). The New governors: The People, rules, and processes governing online speech. *Harvard Law Review*, *131*, 1598–1670. https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/

Knight, B. (2018, January 1). Germany Implements New Internet Hate Speech Crackdown. *Deutsche Welle*. https://www.dw.com/en/germany-implements-new-internet-hate-speech-crackdown/a-41991590

Krämer, J., & Stüdlein, N. (2019). Data portability, data disclosure and data-induced switching costs: Some unintended consequences of the General Data Protection Regulation. *Economics Letters*, *181*, 99–103. https://doi.org/10.1016/j.econlet.2019.05.015

Kuo, L. (2019, November 27). TikTok 'makeup tutorial' goes viral with call to action on China's treatment of Uighurs. *The Guardian*. https://www.theguardian.com/technology/2019/nov/27/tiktok-makeup-tutorial-conceals-call-to-action-on-chinas-treatment-of-uighurs

Lu, D. (2019, September 27). WhatsApp Restrictions Slow the Spread of Fake News—But Don't Stop It. *New Scientist*. https://www.newscientist.com/article/2217937-whatsapp-restrictions-slow-the-spread-of-fake-news-but-dont-stop-it/

McCosker, A., & Gerrard, Y. (2020). Hashtagging depression on Instagram: Towards a more inclusive mental health research methodology. *New Media & Society*. https://doi.org/10.1177/1461444820921349

Nardone, M. (2015, August 30). *Getting banned from the App Store was the best thing that happened to us*. https://techcrunch.com/2015/08/30/getting-banned-from-the-app-store-was-the-best-thing-that-happened-to-us/.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D., & Nielsen, R. K. (2019). *Reuters Institute Digital News Report 2019* [Report]. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-06/DNR_2019_FINAL_0.pdf

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2016). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *20*(1), 293–310. https://doi.org/10.1177/1461444816661553

Polanyi, K. (1944). *The Great Transformation*. Farrar and Rinehart.

Prince, M. (2017, August 16). Why we terminated Daily Stormer [Blog post]. *The Cloudfare Blog*. https://blog.cloudflare.com/why-we-terminated-daily-stormer/

Rahman, K. S. (2018). The new utilities: Private power, social infrastructure, and the revival of the public utility concept. *Cardozo Law Review*, *39*(5), 1621–1689. http://cardozolawreview.com/wp-content/uploads/2018/07/RAHMAN.39.5.2.pdf

Rajput, R., Saha, A., Kumari, S., Janyala, S., TA, J., Janardhanan, A., Pandey, P., & Ghose, D. (2017, August 16). Murderous mob—9 states, 27 killings, one year: And a pattern to the lynchings. *The Indian Express*. https://indianexpress.com/article/india/murderous-mob-lynching-incidents-in-india-dhule-whatsapp-rumour-5247741/

Roberts, S. T. (2017, March 8). Social Media's Silent Filter. *The Atlantic*. https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/

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

Ruchaud, N., & Dugelay, J. L. (2016). Automatic face anonymization in visual data: Are we really well protected? *Electronic Imaging*, *2016*(15), 1–7. https://doi.org/10.2352/ISSN.2470-1173.2016.15.IPAS-181

Samuelson, P. (2019). Europe's controversial digital copyright directive finalized. *Communications of the ACM*, *62*(11), 24–27. https://doi.org/10.1145/3363179

Sharma, M. (2019, December 5). Exclusive: What Personal Data Protection Bill 2019 is likely to propose. *India Today*. https://www.indiatoday.in/india/story/exclusive-what-personal-data-protection-bill-2019-is-likely-to-propose-1625472-2019-12-05

Shead, S. (2017, February 8). Inside the crash of Fling, the startup whose founder partied on an island while his company burned through $21 million. *Business Insider*. https://www.businessinsider.com/how-fling-social-media-app-died-2016-11

Sinnreich, A. (2018). Four crises in algorithmic governance. In J. C. Joerden, R. Schmücker, & E. Ortland (Eds.), *Annual Review of Law and Ethics* (Vol. 26, pp. 190–199). Duncker & Humblot.

Srnicek, N. (2016). *Platform capitalism*. Polity Press.

Stewart, R. (2017, February 28). Facebook Tweaks Its Algorithm to Give More Prominence to Posts with 'Reactions.' *Business Insider*. https://www.businessinsider.com/facebook-tweaks-algorithm-to-give-more-value-to-posts-with-reactions-2017-2

Taylor, A. (2018, August 1). The Automation Charade. *Logic Magazine*, *5*. https://logicmag.io/failure/the-automation-charade/

Treré, E. (2020). The banality of WhatsApp: On the everyday politics of backstage activism in Mexico

and Spain. *First Monday*, *25*(12). https://doi.org/10.5210/fm.v25i12.10404

Tusikov, N. (2019). Defunding hate: PayPal's regulation of hate groups. *Surveillance & Society*, *17*(1/2), 46–53. https://doi.org/10.24908/ss.v17i1/2.12908

van Dijck, J., Nieborg, D., & Poell, T. (2019). Reframing platform power. *Internet Policy Review*, *8*(2). https://doi.org/10.14763/2019.2.1414

Vibert, F. (2007). *The rise of the unelected: Democracy and the new separation of powers*. Cambridge University Press.

WhatsApp. (2019). *Stopping Abuse: How WhatsApp fights bulk messaging and automated behavior* [White Paper]. WhatsApp FAQ. https://web.archive.org/web/20190618205325/https://www.whatsapp.com/safety/WA_StoppingAbuse_Whitepaper_020418_Update.pdf

WhatsApp. (2020, April 7). Keeping WhatsApp personal and private [Blog post]. *WhatsApp Blog*. https://blog.whatsapp.com/Keeping-WhatsApp-Personal-and-Private?lang=en

Wu, T. (2018). *The Curse of Bigness: Antitrust in the New Gilded Age*. Columbia Global Reports.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.

Zuckerberg, M. (2019, March 6). *A privacy-focused vision for social networking* [Post]. https://www.facebook.com/notes/mark-zuckerberg/a-privacy-focused-vision-for-social-networking/10156700570096634/

Published by

Alexander von Humboldt Institute for Internet and Society

in cooperation with

CREATe

centre internet et société

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya