CHAPTER 13

Self-regulation and Discretion

Nicolas Suzor and Rosalie Gillett

Introduction

Who should decide what content is permissible online? There is increasing pressure on platforms to do more to remove harmful speech, avoid removing legitimate speech, and ensure that their moderation systems are free from bias. Global communications platforms wield an inordinate degree of power and govern their networks with almost-absolute discretion (Suzor 2019). Clearly, there is unease about platforms making ad-hoc decisions and applying rules that they make up as they go along

This research was supported by an Australian Research Council Discovery Project grant (DP190100222) and the Australian Research Council Centre of Excellence for Automated Decision-making and Society (CE200100005).

N. Suzor (✉)
QUT Digital Media Research Centre, Kelvin Grove, QLD, Australia
e-mail: n.suzor@qut.edu.au

R. Gillett
Australian Research Council Centre of Excellence for Automated Decision-Making and Society, QUT, Kelvin Grove, QLD, Australia
e-mail: rosalie.gillett@qut.edu.au

T. Flew and F. R. Martin (eds.), *Digital Platform Regulation*,
Palgrave Global Media Policy and Business,
https://doi.org/10.1007/978-3-030-95220-4_13

(Barrett 2020; Buni and Chemaly 2016). Although platforms have been improving their content moderation processes, industry self-regulation is often thought of as far too weak to bring real accountability to platform governance (Helberger et al. 2018). There is, accordingly, an understandable desire among commentators to see more democratic oversight for digital platforms (Winseck 2020; Haggart and Keller 2021).

We share the view that democratic rule-making is increasingly important to regulate the power of digital platforms. There are strong arguments in favour of public regulation based on clear and enforceable legal standards, properly made by legitimate bodies in accordance with democratic processes and constitutional limitations (Haggart 2020). Good public regulation of platforms likely also requires adapting antitrust to the platform economy (Khan 2016; Teachout 2020) and more targeted regulation of infrastructure (Frischmann 2012), the flow of private information, and trade practices (de Streel et al. 2020).

In this Chapter, we argue that improving the self-regulation of internal governance practices of platforms is a critical component of any regulatory project. Discussions about platform governance sometimes treat regulatory approaches as a choice between apparently distinct and exclusive models: self-regulation, co-regulation, multi-stakeholderism, or democratic rule (Haggart and Keller 2021). We suggest that self-regulation does not displace the need for greater scholarly attention to democratic regulation of platforms. These are not exclusive concepts (Marsden 2011). Industry self- and co-regulation may not be sufficient to bring legitimacy to platform governance (Haggart and Keller 2021), but platforms will always exercise discretion, and convincing platforms to exercise their discretionary powers responsibly is a large part of making governance legitimate (Suzor 2018).

We make our case based on the results of a qualitative study involving a broad group of participants who actively work to influence how platforms govern their users. We understand governance in broad terms as 'organized efforts to manage the course of events in a social system' (Burris, Kempa, and Shearing 2008). In this sense, platforms govern their users (Klonick 2017) and are subject to influence through overlapping 'polycentric' (Black 2008) formal and informal regulatory regimes. We interviewed 25 participants from across business, civil society, and government to understand how they sought to influence the discretionary powers that platforms wield. We investigate how 'hard' law is often enforced informally, through pressure exerted by regulators, NGOs, and private actors. We also examine how (often conflicting) demands for platforms to address 'lawful but harmful' conduct and material play out in practice. In both cases, we highlight the importance, consistently emphasized by our participants, of strong relationships between stakeholders and individual representatives of platforms who have a degree of influence and discretion to effect change or at least to broker connections to those who can.

Our argument is that platforms must always have a role in regulating lawful speech—that platforms must influence cultures, affordances, and social norms—and that regulating ordinary, lawful speech is critical to addressing harm. Here, we make two claims: self-regulation is both necessary and good. Necessary, in that in any regulatory regime, there are always zones of discretion within which platforms will interpret and enforce the rules they create and impose on users. And we argue that this discretion is good, in that private platforms should govern in ways that are appropriate for their unique cultures (and, for the majority of platforms, their business interests). We offer a simple proof in the moral responsibilities that platforms bear to address the pressing need for cultural change in violence against women – responsibilities that cannot fully be carried out or overseen by states or other external actors.

Because platforms exercise discretion and are influenced by a wide array of social actors, we suggest that finding ways to improve the daily practice of self-regulation by technology companies is still a necessary and important goal that will persist regardless of any formal regulatory schemes that apply. Our findings show how governance in practice, whether backed by formal law or not, involves a great deal of discretionary power and external influence. We conclude that understanding how loose networks of civil society, businesses, journalists, regulators, users, and others can effectively exert pressure on platforms for prosocial ends, and the limits at which these efforts fail, continues to be a fundamentally important challenge. Understanding how platforms respond to external demands that are judged to be positive or negative by and for different societies at different points in time is, accordingly, a key pre-requisite to understanding how democratic processes could effectively promote public interests in a global pluralistic networked environment. This is the ongoing challenge of 'digital constitutionalism', (Celeste 2019) which builds on the insights from regulatory theory that 'constitutionalizing self-regulation' (Black 1996) is necessary to bring legitimacy to systems of governance that are partially autonomous while also expected to account to diverse groupings of state and private actors.

Background

Most major digital platforms have grown up under the wide protection of US law. The protections introduced by the Communications Decency Act, codified in 47 U.S.C. § 230, ensures that platforms are generally not legally liable for content posted by their users and are free to moderate as they see fit (Klonick 2017). Platforms have individual policies and community standards that set the rules for conduct—frequently set out in ways that are vague or unclear to users (West 2018). These rules are enforced through complicated content moderation processes, often including a mix of outsourced workers reviewing content; machine learning classifiers and hash matching tools that detect, prioritise, and remove material; and internal policy teams that set standards, oversee moderation, and make final decisions in some cases (Roberts 2019).

Platforms also operate various additional procedures to handle takedown requests from external users, in addition to internal flagging procedures. For example, any major platform will have a system for receiving large volumes of copyright takedown notices; direct connections with police and coordinating organisations for identifying child abuse material (Holt et al. 2020); other channels for receiving requests for information or content removal from law enforcement agencies; and channels to receive other requests for content removal, whether authorised under law or not. When the number or severity of incoming requests becomes high enough, platforms will usually build dedicated workflows—for example, specific processes to handle non-consensual explicit imagery (Gillespie 2017) or requests under the European Right to be Forgotten. Some incoming requests are processed wholly automatically, some are dealt with by legal teams, and others handled by other parts of the company. In some cases, decisions based on these requests are also used to train automated systems to detect similar content in the future.

At a large enough scale, the content moderation systems of major platforms quickly become extraordinarily complex. For many years, activists, academics, and journalists have criticized the bias, arbitrary rules, bad decisions, and the lack of clarity and certainty in the commercial content moderation systems of major platforms (e.g. York 2021; MacKinnon 2012; Suzor 2011; Buni and Chemaly 2016). While many platforms have improved over the years in response to heavy public and media pressure (see, for example, WAM! 2013), these are all still major problems. When measured against the norms of legitimate governance that are routinely applied to nation states, platforms fare extremely poorly (Suzor 2018). When evaluated against substantive human rights concerns, content moderation systems also fail spectacularly in many ways (Kaye 2019). And no major commercial platforms provide serious democratic processes for developing editorial rules and overseeing their enforcement (Haggart 2020).

The pressure on platforms to change their content governance processes is strong and intensifying. There is clear demand for platforms to do more to suppress harmful speech and to avoid suppressing valuable speech—even if there is less consensus about where these categories begin and end. In terms of public policy, a dizzying array of policy reports, law suits, and legislative proposals are under various stages of development and debate across the world (Puppis and Winseck 2021; Flew and Gillett 2021). These proposals vary widely; the range of public policy options for platform governance is broad (Heldt 2019a). Some are based in the familiar realm of intermediary liability, where platforms are legally responsible for facilitating harms caused by their users. Some impose new obligations on platforms to remove unlawful or prohibited content upon receiving a complaint, like the German Network Enforcement Act (NetzDG) (Heldt 2019b; Schulz 2018). Others invest public regulators with powers to require platforms to remove content—like the new Australian Online Safety Act. Some approaches include requirements for transparency reporting (Wagner et al. 2020). Other approaches focus on encouraging or facilitating industry self-regulation (Bridy 2019), like the ‘Christchurch Call’, developed in the aftermath of the live-streamed massacre in 2019 (Hoverd et al. 2020). Some seek to create new, generalized duties of care on platforms to address foreseeable harm (Woods 2019). Others look to telecommunications and competition policy to inform public accountability and structural changes to internet industries (Winseck 2020). Still more options include extending media classification standards to internet platforms (Flew et al. 2019), or developing new public–private partnerships to create co-regulatory standards for acceptable content (Haggart et al. 2021).

No matter what form they take, however, all these legal obligations on platforms will be interpreted through each platform’s priorities and implemented through their own processes. Some forms of regulation will impose greater accountability for how platforms choose to comply, but compliance is never perfect, nor is it automatic. The danger of emphasizing public regulation over private action is that it can lead to a false binary. Global platforms operate across many different legal systems, and their practices are influenced by an extremely broad range of actors—including states and their constituent components; their business partners, competitors, suppliers, and customers; their public audiences, NGOs, and media organisations. Some scholars have suggested recently that platform governance scholarship has perhaps paid insufficient attention to the work of those stakeholders in governance, particularly outside of formal multi-stakeholder regimes (Papaevangelou 2021). Developing a better understanding of how networked platform governance works in practice, and how it can be improved, is the problem to which we now turn.

Research Methods

This study relies on qualitative interviews with a broad range of stakeholders who actively work to influence how platforms govern their networks. We draw on Gorwa's (2019) 'platform governance triangle' to group interview participants into three groups of institutional actors: firms, NGOs, and government. We recruited regulators who exercised legal authority to compel compliance and regulators who worked informally; lawyers who represented platforms dealing with incoming complaints; community managers; journalists; NGOs advocating for stronger rules for removal of harmful speech and the protection of counterspeech; and firms that specialize in 'reputation defence', by scrubbing or burying negative material online. Our groupings are kept at a high level; the focus of this study on platforms does not require comparison across groups of external stakeholders. The strength of this broad methodology is that it helps to contrast how different regulatory approaches are experienced in practice; the unavoidable limitation is that our data should not be used to generalize across particular forms of regulation, social issues, or stakeholders.

Between 2017 and 2018, we conducted 25 interviews with people who are involved in seeking the preservation or removal of internet content. In 'firms' (n = 11), we include representatives of companies, groups of companies, and industry associations. Second, 'NGO' (n = 11), comprises non-government organizations, civil society, academics, and private individuals who identify as advocates or activists. Finally, we include representatives from regulatory agencies, government officials, and inter-governmental organizations in the ‘regulator’ grouping (n = 3). To protect the identities of the interview participants, we have withheld their names and the organisations that they work for. The following sections thematically represent participants’ experiences attempting to convince platforms to remove or protect online content.

We conducted this research primarily in Australia, where the laws that apply to digital platforms vary extraordinarily in their approach (Pappalardo and Suzor 2018). Australian intermediary liability regimes differ widely in the strength of the incentives they provide platforms to comply with demands of our participants. The range of legal consequences includes severe criminal sanctions, established takedown regimes, threats of civil liability, and issues that are only dealt with in the public arena, not through law. This variety of rules provides a useful opportunity to understand how people dealing with platform governance issues experience different regulatory approaches. Some participants were outside of Australia; their experiences are used particularly for the analysis of extra-legal moderation of lawful content that is not jurisdiction-specific.

Legal Rules Are Routinely Enforced Informally

The first thing to note about calls for greater public regulation is that legal regimes differ in how much discretion they expect platforms to exercise. Some regimes, like copyright takedowns, are highly standardised, requiring little or no exercise of discretion by platforms (Urban et al. 2016). Others, like defamation law, place the burden of assessing the merits of complaints on platforms (Pappalardo and Suzor 2018). Where platforms are required to exercise their judgment about whether to remove content, the decisions they make can vary to the point of appearing arbitrary or incoherent. Our interview participants, including both public regulators and private advocacy organisations, explained how they have had to develop informal relationships with platforms in order to be able to effectively request removal of unlawful content. Participants most often described developing and maintaining rapport and meaningful relationships with those who worked at large social media platforms. Several of our participants noted that these established relationships meant they were often much more successful at requesting platforms to remove unlawful content than police were.

Platforms have developed formal processes specifically for responding to requests from law enforcement agencies and for dealing with common private legal demands. These are often much more onerous and slower than informal channels can be: "[Platforms are] extremely slow to respond to law enforcement requests". To avoid the overheads of formal processes, we heard that sometimes law enforcement officers would refer material to NGOs to report to platforms, rather than take formal action under law:

> We've been in the situation where we've had police come to us with content, saying: Can you help us get a response to this? And that's happened quite often, actually […] this is the role that civil society plays, particularly in the US-focused context where distrust of government is part of the culture. And therefore civil society actually helps bridge the gap, that platforms can be notified in a manner that is voluntary, where the platforms, any response the platform takes, when it's coming from civil society, is the platform's own decision. It's not under government compulsion. (NGO)

Even our participants from regulatory agencies told us that their usual mechanisms for enforcement were informal. An official from a public regulator who asked us to paraphrase their comments explained that when dealing with social media companies, even though they have some legal enforcement powers, they had never sought to use them in court. Their main tool was 'reputational damage': they would ask platforms to remove content, and if they did not, the regulator would make a public statement that the company has not complied.

The extra formalities for legal requests exist in part because platforms have been under heavy criticism for many years for acceding too readily demands that they remove content or hand over personal information. In democracies and authoritarian states, law enforcement agencies have worked to exploit informal pressure and tacit agreements ("invisible handshakes": Birnhack and Elkin-Koren 2003) to circumvent fetters on government power and due process safeguards (Elkin-Koren and Haber 2016). Private actors too exert informal pressure on platforms and develop mutual agreements to enforce their legal rights. Intellectual property owners, for example, have a long history of working with internet infrastructure companies, banks and credit card processors, and others to shut down or financially strangle sites that traffic counterfeit (and sometimes legal) pharmaceuticals, luxury goods, media, and other goods—often with 'non-regulatory' support from public agencies (Bridy 2014). Criticisms of these practices have led platforms to develop stricter procedures for appraising incoming legal requests that allow platforms to make a considered decision about whether to comply or resist (Eichensehr 2018).

Not all legal enforcement mechanisms are slow. For child abuse material (which is universally condemned) and copyright infringement (where complaints are very high in volume), moderation by platforms is routinely automated. But the platform response time for non-automated takedown requests varies widely. Even for a participant whose role in a public agency concerned child abuse material, the effectiveness of takedown requests to platforms for clearly unlawful material often depended 'on the personal relationships that exist between investigators and key representatives of those companies.'

Effective corporate regulation frequently requires a long-term relationship between the corporation and the regulator that is sensitive to the internal processes and culture of the firm (Black and Baldwin 2010). Informal enforcement can be effective in securing compliance, but it relies on the threat of potential penalties and on the moral force of the law (Parker 2006). Several regulators in our study explained how they were able to escalate serious issues to internal contacts within major platforms, and at least where they could be dealt with locally, the platform's response time would often be within a few hours. One regulator told us that they found platforms were 'pretty responsive' to requests 'where there's a real direct threat of harm to a person, whether a child or an adult'. But regulators struggled with 'grey area' content that was less clearly unlawful or harmful, noting that takedown requests 'are dealt with inconsistently, I think, and sometimes perhaps not in a way that we would say accords with our reasonable expectations' (Regulator).

Some regulators expressed discomfort about the potential legitimacy problems that arise from the informal use of their powers. For example, when legislation imposes penalties on platforms for failing to remove image-based abuse, but the main channel for enforcement is informal, one of our participants worried that their role might be 'playing judge and jury'—a challenge to due process that they recognised was at odds with the need to act quickly:

> That's a lot for a government body […] when time is of the essence. If a naked picture of me is on the internet and I haven't consented to that […] You can't wait for a court process, you need that taken down, my mental

> health is at stake, my life could very much be at stake, and depending on what community I come from, so could the lives of my loved ones. (Regulator)

Despite these legitimacy concerns, however, broader legal scholarship suggests that legal rules are frequently enforced informally across many different areas of regulation. Law is always experienced differently in practice than it is written—and it is informal practices, not the courts, that govern most interactions (Ellickson 1991). Regulators often seek to procure compliance through light-touch informal channels before escalating to more formal rules and penalties (Ayres and Braithwaite 1992). ven to the extent that laws set minimum standards for content or processes for determining and enforcing the rules, platforms still exercise a great deal of discretion in applying those rules (Douek 2020). This is true for private legal demands too; our private sector participants, including lawyers and representatives from reputation management firms, noted that the platforms they dealt with or represented would often take a risk-management approach to demands for content removal based on formal law.

In a practical sense, then, a substantial zone of discretion is inevitable. At any reasonable scale, the full due process of state institutions becomes unworkable in terms of time, cost, and complexity. Routine enforcement of speech law online will likely continue to be done largely by platforms, who will continue to exercise discretion in deciding whether and how to fulfil their various legal obligations. In the past, platforms have structured their businesses to concentrate their people, assets, and income in jurisdictions that provide them more legal protections (and frequently, lower tax). Given the complex geopolitical struggles between states and regional authorities that underpin different approaches to platform regulation (Gray 2021), these zones of jurisdictional conflict and associated discretion are unlikely to disappear in the foreseeable future. It is also important to note that we should not aim for perfect compliance; there are many cases where we expect technology companies not to defer to legal demands from states (often for personal information or censorship) in order to protect the rights of users worldwide (Svantesson 2014).

Regulating 'Lawful but Harmful' Content

The discretion of platforms to enforce their own rules is strongly subject to influence from stakeholders. Many of our participants told us about how, in the absence of binding legal obligations on platforms, they leveraged their relationships to draw attention to material that contravened the platform's rules or legitimate content that had been wrongfully removed. Some of our participants from regulatory agencies explained how they used informal channels to request that platforms remove content that the regulator was not legally empowered to compel the platform to remove. One regulator, for example, explained how they were able to ask an imageboard provider to remove sexual material that violated the privacy of a local complainant, even though the imageboard was known for its limited rules and was well outside of the territorial jurisdiction. They explained their work in terms of providing reasons and evidence to the platform to regulate themselves—noting that 'the rules that they establish can, in fact, be enforced…'. They went on, however, to explain that this approach primarily worked for material that was obviously already prohibited under the platform's own rules.

What content is, and ought to be, prohibited by platforms is deeply contested. Platforms are frequently criticized for not sufficiently understanding local contexts and cultures when they enforce their rules—which means they often misunderstand hateful content or wrongly remove counterspeech, particularly speech by, or targeted at, marginalized groups (Matamoros-Fernández 2017). Our participants reported that they often struggled to convince platforms to take action where the content was ambiguous, the harm was less visible, or additional context was required. Both regulators and NGOs noted that platforms are less responsive to take down requests that fall within the "grey area of determining whether or not some kind of protection attaches to that speech" (Regulator). An NGO representative who tackles hate speech said: "the threshold for what's considered offensive is incredibly high, both legally and quite often from the members of the public. So casual racist comments, although they may be grossly offensive, are unlikely to get removed." At the same time, participants were often concerned about platforms applying their rules in an overly restrictive way that silenced the voices of marginalized users:

> Facebook in particular has a history of ignoring what is flat-out violence, pages devoted to violence against women, and meanwhile taking down

> pages where women are owning their own sexuality or showing post-surgery breast cancer photos and those kinds of things. (NGO)

Some NGO representatives who advocated for marginalized groups explained that without a deep understanding of diverse cultures and languages, platforms cannot adequately moderate their users' content. These participants leveraged their organization's profile and expertise to show platforms how they could better address the needs of their users. One NGO representative observed the important role they played in using their organization's experience and expertise 'teaching' platforms—and their machine learning classifiers—'to recognize the subtleties of hate speech.'

Platforms rely heavily on the labour of external organisations to help them identify and prioritize harmful content. Content moderation requires users to report ('flag') content they find objectionable (Crawford and Gillespie 2014). But the accuracy of user reports, measured against the platform's rules, is typically quite bad; users frequently flag content that is not prohibited (Matias et al. 2015). Our NGO participants explained how they provide platforms with a trusted source for vetted flags. They undertake the work of investigating and triaging complaints, understanding context, and identifying those that are most serious. Some of our participants also told us how they do the additional painstaking work of 'translating' the complaints of users into the rather technical categories of rules that platforms use—without which, they felt, user concerns were much more likely to be ignored.

There are major limits to the influence of civil society actors on the policies of platforms. Even though platforms were often responsive to specific removal requests where there was clear harm, participants described the game of "whack-a-mole" they played with social media companies to keep content down. One NGO representative described their efforts to get image-based abuse removed from YouTube: "it would pop up again and we would have to intervene again because YouTube was not responding the way that they were supposed to" (NGO). Some participants thought that their takedown requests were unsuccessful because they competed with platforms' business interests: "And so, I think, whether or not they're receptive has a direct correlation to whether or not what we're asking for goes directly to their business model or to their bottom line." (NGO).

One of the major challenges of informal content regulation is that it is difficult to drive longer-term policy change. Participants complained that even where platforms acted on the individual reports they made, ‘there’s been very little concrete action beyond that … at a policy level and not just individual case levels, we’ve seen that there hasn’t been much apart from rhetoric at the moment.’ (NGO) Some platforms have explicit ‘trusted flagger’ programs that are designed to prioritise complaints from experienced NGOs and regulators, and some of our participants found these programs to be quite effective in terms of receiving quick responses from platforms. Other NGO actors in our study, however, thought that their relationships with platforms were tokenistic. An NGO representative who advocated for women online believed that this tokenism meant platforms did not fully understand the concerns of their users:

> This is part of why I think they aren’t really listening to the stakeholders that they invite to the table or really asking them the right questions, because if they did, some of these things that they roll out and then roll back they wouldn’t be doing. (NGO)

Another NGO representative explained how the organization they worked for was a member of a social media platform’s safety board, but that they doubted the meaningfulness of this partnership and understood it as a public relations stunt: “we knew that they were just using us to look good.”

Platforms are perhaps most responsive when faced with public crises. Policy changes and promises are frequently made in response to ‘public shocks’ (Ananny and Gillespie 2017), but lasting change is more challenging. One of our participants explained their experience as an editor of a major news publication featuring Indigenous writers discussing discrimination and abuse. The editor and the writers repeatedly had their articles removed and their personal accounts suspended from major social media platforms for sharing links to their published articles about racism. The editor explained that their complaints to these platforms had been repeatedly ignored, and it was only after they were able to turn one incident into a major news story that the platform concerned was willing to engage. Even then, the editor characterized the platform’s response as a public relations exercise by people ‘who are not genuine, they just genuinely want the problem to go away.’ When the editor re-shared the same content a year later, the platform again suspended their account, suggesting the platform's initial response was an isolated reaction to a crisis, not an attempt to address underlying problems: 'if you were sensitive to Aboriginal customs, then you would work out a way to fix it, but they haven't.' (Firm).

Platforms play an important but fraught role in setting and enforcing the boundaries of acceptable speech. Informal pressure on platforms to regulate lawful speech is common, but some public regulators or law enforcement officials respondents expressed concern about the legitimacy of asking platforms to enforce rules that are not provided by law. One of the regulators we spoke to explained that they routinely approach platforms with complaints under their terms of service, but were concerned about the implications:

> When it comes to adults, where do you draw the line between robust discussion and disagreement, such as vile disagreement and conduct that should be regarded as worthy of regulation… (Regulator)

The regulator continued, articulating a concern that is core to the rule of law: that rules ought to be clear, validly made, and fairly enforced: 'if it's worthy of regulation, why aren't the police properly granted that role?'.

The Necessity of Private Discretion

There is clearly something deeply troubling about relying on the extra-legal enforcement of non-democratic prohibitions on speech by unaccountable private platforms. But the set of rules that platforms enforce—and are frequently expected to enforce—is necessarily much broader than what laws require. Platforms are not 'common carriers': they are legally entitled to determine their rules and enforcement procedures, and with limited exceptions, they are not prohibited from discriminating for or against certain types of content or groups of speakers. This, we suggest, is a Good Thing. Policy that would limit the ability of platforms to discriminate against different types of lawful speech and different speakers would not only flatten competitive differences between platforms but likely drown us all in cesspits of spam, abuse, disinformation, and irrelevance.

At any rate, pragmatically, we are not heading towards a future where platforms are required to moderate less. Platforms are under increasing pressure to do much more to regulate 'lawful but harmful' speech online. Take, for example, the demands on platforms to address toxic and hateful content on their networks. Ordinary hateful speech that does not rise to the level of explicit hate speech is generally not prohibited, but it nevertheless creates and reinforces the foundations for violence and discrimination. Harmful criminal acts that we view as aberrant are made possible by the normalization of ordinary abusive behavior (Kelly 1988). Part of the link is explicit; malicious actors use covert and coordinated hate campaigns (Lewis et al. 2020; Marwick and Caplan 2018) to spread and reinforce harmful attitudes toward marginalized groups (Shifman and Lemish 2011; Matamoros-Fernández 2020). Users learn to deliberately skirt legal rules and develop strategies to avoid content moderation systems (Matamoros-Fernández 2020; Bhat and Klein 2020). A great deal of discrimination is propagated and normalized through everyday sexism and misogynistic views (Jones et al. 2019) and sexist humour (Shifman and Lemish 2011). But the perpetuation of oppression is also implicit in ordinary expressions of prejudice and acts of discrimination that enable widespread abuse and harassment to become normalized online (Gillett 2019). Much of this harmful speech is not and should not be regulated by law—the abilities of states to create laws that make content unlawful to distribute are necessarily restricted in scope and subject matter. This does not mean that hateful speech should not be regulated; rather, that it should be regulated through social norms and private approbation (Matsuda 1989). This likely includes rules set by platforms which, we have suggested elsewhere, have a responsibility to address systemic inequalities that are perpetuated, at least in part, by these types of speech on their networks (Suzor et al. 2018).

Strong government regulation of digital platforms is more democratic (Haggart 2020) and better aligned with the rule of law and constitutionality (Winseck 2020) than private ordering, self-regulation, and discretionary power. But legal rules cannot cover the entire field of decisions that platforms make. The interpretation of rules is always imprecise – rules expressed in natural language are necessarily open to interpretation (Hart 1994). Even where they are clear, rules are never perfectly enforced; there is a great deal of content on major platforms that might be prohibited but has never been reported. Users are less likely to report prohibited content that they do not perceive to be highly harmful or routine, and platforms often choose not to enforce their rules strictly.

The answer is not to try to remove discretion. The limits that societies impose on the ability of states to exercise coercive power do not translate directly to digital platforms. Discretionary power is fundamentally necessary to platforms as we know them. In a world of information abundance, content moderation and curation is the commodity that platforms offer to their users (Gillespie 2018). Digital platforms implement extensive rules designed to protect their business interests, meet the expectations of their users, and shape their own distinct cultures (Burgess and Baym 2020). They need a degree of discretion to align their rules, affordances, and processes to their distinct cultures and priorities (Klonick 2017). Platforms also need discretion to create and enforce timely rules that respond to harmful lawful content and reinforce prosocial norms on the limits of socially acceptable speech.

We suggest instead that one of the critical tasks ahead for scholars of platform governance is to better understand how discretionary power can and should be appropriately limited and made accountable – what regulatory scholars call ‘throughput legitimacy’ (Haggart and Keller 2021). Discretion is legitimate where it is constrained within a zone of autonomy; generally speaking, platforms currently enjoy ‘broad’ discretion: power without effective oversight (Suzor 2011). The development of new mechanisms to limit – or ‘constitutionalize’ – the discretionary power of platforms is critical to improving platform governance (Celeste 2021; Suzor 2019). But for global platforms enmeshed in many varied controversies with a great many stakeholders over the governance of their networks, this is no easy task. From the little we know so far about the rapidly changing decision-making of platforms, whatever legal limits we might seek to impose on platforms, internal commitment, effective self-regulation, and extra-legal pressure will have a major impact on compliance. As with so much else, cultural change is key.

References

Ananny, M., & Gillespie, T. (2017). Public Platforms: Beyond the Cycle of Shocks and Exceptions. In *Interventions: Communication Research and Practice*. San Diego, CA.

Ayres, I., & Braithwaite, J. (1992). *Responsive Regulation: Transcending the Deregulation Debate*. Oxford University Press, USA.

Barrett, P. (2020). *Who Moderates the Social Media Giants? A Call to End Outsourcing*. NYU Stern Center for Business and Human Rights. https://bhr.stern.nyu.edu/blogs/2020/6/4/who-moderates-the-social-media-giants.

Bhat, P., & Klein, O. (2020). "Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter." In *Twitter, the Public Sphere, and the Chaos of Online Deliberation*, edited by G. Bouvier and J. E. Rosenbaum, 151–172. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-41421-4_7.

Birnhack, M. D., & Elkin-Koren, N. (2003). The Invisible Handshake: The Reemergence of the State in the Digital Environment. *Virginia Journal of Law and Technology*, *8*, 6–13.

Black, J. (1996). Constitutionalising Self-regulation. *The Modern Law Review*, *59*(1), 24–55.

Black, J. (2008). Constructing and Contesting Legitimacy and Accountability in Polycentric Regulatory Regimes. *Regulation & Governance*, *2*(2), 137–164.

Black, J., & Baldwin, R. (2010). Really Responsive Risk-Based Regulation. *Law and Policy*, *32*(2), 181–213.

Bridy, A. (2014). Internet Payment Blockades. *Florida Law Review*. http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2494019.

Bridy, A. (2019). Leveraging CDA 230 to Counter Online Extremism (September 1). https://papers.ssrn.com/abstract=3538919.

Buni, C., & Chemaly, S. (2016). The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech. *The Verge*. http://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech.

Burgess, J., & Baym, N. K. (2020). *Twitter: A Biography*. New York: NYU Press.

Burris, S., Kempa, M., & Shearing, C. (2008). Changes in Governance: A Cross-Disciplinary Review of Current Scholarship. *Akron Law Review*, *41*, 1–66.

Celeste, E. (2019). Digital Constitutionalism: A New Systematic Theorisation. *International Review of Law, Computers & Technology*, *33*(1) (January 3), 76–99. World.

Celeste, E. (2021). Digital Punishment: Social Media Exclusion and the Constitutionalising Role of National Courts. *International Review of Law, Computers & Technology*, *35*(2) (May 4), 162–184.

Crawford, K., & Gillespie, T. (2014). What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint. *New Media & Society*, *18*(3), 410–428.

de Streel, A., Feasey, R., Monti, G., Krämer, J., & Cave, M. (2020). *Digital Markets Act: Making Economic Regulation of Platforms Fit for the Digital Age*. Centre on Regulation in Europe. https://cerre.eu/wp-content/uploads/2020/11/CERRE_DMA_Making-economic-regulation-of-platforms-fit-for-the-digital-age_Full-report_December2020.pdf.

Douek, E. (2020). Australia's "Abhorrent Violent Material" Law: Shouting "Nerd Harder" and Drowning Out Speech. *Australian Law Journal*. https://papers.ssrn.com/abstract=3443220.

Eichensehr, K. E. (2018). Digital Switzerlands. *University of Pennsylvania Law Review*, *167*(3), 665–732.

Elkin-Koren, N., & Haber, E. (2016). Governance by Proxy: Cyber Challenges to Civil Liberties. *Brooklyn Law Review*, *82*(1) (February 28), 105–162.

Ellickson, R. C. (1991). *Order Without Law: How Neighbors Settle Disputes*. Harvard University Press.

Flew, T., & Gillett, R. (2021). Platform Policy: Evaluating Different Reponses to the Challenges of Platform Power. *Journal of Digital Media & Policy*, *12*(2).

Flew, T., Martin, F., & Suzor, N. (2019). Internet Regulation as Media Policy: Rethinking the Question of Digital Communication Platform Governance. *Journal of Digital Media & Policy*, *10*(1) (March 1), 33–50.

Frischmann, B. M. (2012). *Infrastructure: The Social Value of Shared Resources*. New York: Oxford University Press.

Gillespie, T. (2017). Governance of and by Platforms. In *SAGE Handbook of Social Media*, edited by J. Burgess, T. Poell, and A. Marwick, 254–278. London: SAGE Publications. http://culturedigitally.org/wp-content/uploads/2016/06/Gillespie-Governance-ofby-Platforms-PREPRINT.pdf.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. 1st edition. New Haven, CT: Yale University Press.

Gillett, R. (2019). *Everyday Violence: Women's Experiences of Intimate Intrusions on Tinder. PhD*. Queensland University of Technology. https://eprints.qut.edu.au/131121/.

Gorwa, R. (2019). The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content. *Internet Policy Review*, *8*(2) (June 30). https://policyreview.info/articles/analysis/platform-governance-triangle-conceptualising-informal-regulation-online-content.

Gray, J. (2021). The Geopolitics of "Platforms": The TikTok Challenge. *Internet Policy, Review*, *10*(2) (May 11). https://policyreview.info/articles/analysis/geopolitics-platforms-tiktok-challenge.

Haggart, B. (2020). Global Platform Governance and the Internet-Governance Impossibility Theorem. *Journal of Digital Media & Policy*, *11*(3) (November 1), 321–339.

Haggart, B., & Keller, C. I. (2021). Democratic Legitimacy in Global Platform Governance. *Telecommunications Policy*, *45*(6). Norm Entrepreneurship in Internet Governance (July 1), 102152.

Haggart, B., Scholte, J. A., & Tusikov, N. (2021). Introduction: Return of the State? In *Power and Authority in Internet Governance: Return of the State?* 1–12. London: Routledge.

Hart, H. L. A. (1994). *The Concept of Law*. 2nd ed. Oxford: Clarendon Press.

Helberger, N., Pierson, J., & Poell, T. (2018). Governing Online Platforms: From Contested to Cooperative Responsibility. *The Information Society*, *34*(1) (January 1), 1–14.

Heldt, A. P. (2019a). Let's Meet Halfway: Sharing New Responsibilities in a Digital Age. *Journal of Information Policy*, *9*, 336–369.

Heldt, A. P. (2019b). Reading Between the Lines and the Numbers: An Analysis of the First NetzDG Reports. *Internet Policy Review*, *8*(2) (June 12). https://policyreview.info/articles/analysis/reading-between-lines-and-numbers-analysis-first-netzdg-reports.

Holt, T. J., Cale, J., Leclerc, B., & Drew, J. (2020). Assessing the Challenges Affecting the Investigative Methods to Combat Online Child Exploitation Material Offenses. *Aggression and Violent Behavior*, *55* (November 1), 101464.

Hoverd, W. J., Salter, L., & Veale, K. (2020). The Christchurch Call: Insecurity, Democracy and Digital Media—Can It Really Counter Online Hate and Extremism? *SN Social Sciences*, *1*(1) (November 9), 2.

Jones, C., V. Trott, & Wright, S. (2019). Sluts and Soyboys: MGTOW and the Production of Misogynistic Online Harassment. *New Media & Society* (November 8), 1461444819887141.

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

Kelly, L. (1988). *Surviving Sexual Violence*. Oxford: Polity Press.

Khan, L. M. (2016). Amazon's Antitrust Paradox. *Yale Law Journal*, *126*, 710.

Klonick, K. (2017). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review*, *131* (March 20), 1598–1670.

Lewis, R., Marwick, A., & Partin, W.C. (2020). "We Dissect Stupidity and Respond to It": Response Videos and Networked Harassment on YouTube. *American Behavioral Scientist*. https://osf.io/preprints/socarxiv/veqyj/.

MacKinnon. (2012). *Consent of the Networked: The Worldwide Struggle for Internet Freedom*.

Marsden, C. T. (2011). *Internet Co-Regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace*. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/internet-coregulation/7179CDF556745BA2313666AEE0A60E70.

Marwick, A. E., & Caplan, R. (2018). Drinking Male Tears: Language, the Manosphere, and Networked Harassment. *Feminist Media Studies*, *18*(4) (July 4), 543–559.

Matamoros-Fernández, A. (2017). Platformed Racism: The Mediation and Circulation of an Australian Race-Based Controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, *20*(6) (June 3), 930–946.

Matamoros-Fernández, A. (2020). "El Negro de WhatsApp" Meme, Digital Blackface, and Racism on Social Media. *First Monday*, *25*(1) (January 5). https://firstmonday.org/ojs/index.php/fm/article/view/10420.

Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015). Reporting, Reviewing, and Responding to Harassment on Twitter. *Women, Action & the Media!* https://ssrn.com/abstract=2602018.

Matsuda, M. J. (1989). Public Response to Racist Speech: Considering the Victim's Story. *Michigan Law Review*, *87*(8), 2320–2381.

Papaevangelou, C. (2021). The Existential Stakes of Platform Governance: A Critical Literature Review. *Open Research Europe*, *1* (July 1), 31.

Pappalardo, K. M., & Suzor, N. P. (2018). The Liability of Australian Online Intermediaries. *Sydney Law Review*, *40*, 469–498.

Parker, C. (2006). The "Compliance" Trap: The Moral Message in Responsive Regulatory Enforcement. *Law & Society Review*, *40*(3) (September 1), 591–622.

Puppis, M., & Winseck, D. (2021). Platform Regulation and Inquiries. *Google Docs*. https://docs.google.com/document/d/1AZdh9sECGfTQEROQjo5fYeiY_gezdf_11B8mQFsuMfs/edit?usp=embed_facebook.

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, Conn.: Yale University Press.

Schulz, W. (2018). Regulating Intermediaries to Protect Privacy Online—The Case of the German NetzDG. In *Personality and Data Protection Rights on the Internet*, edited by M. Albers and I. Sarlet. Rochester. NY: Social Science Research Network. https://papers.ssrn.com/abstract=3216572.

Shifman, L., & Lemish, D. (2011). "Mars and Venus" in Virtual Space: Post-feminist Humor and the Internet. *Critical Studies in Media Communication*, *28*(3) (August 1), 253–273.

Suzor, N. P. (2011). The Role of the Rule of Law in Virtual Communities. *Berkeley Technology Law Journal*, *25*(4), 1817–1886.

Suzor, N. P. (2018). Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms. *Social Media + Society*, *4*(3) (July 1), 1–11.

Suzor, N. P. (2019). *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge: Cambridge University Press.

Suzor, N. P., Dragiewicz, M., Harris, B., Gillett, R., Burgess, J., & Van Geelen, T. (2018). Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online. *Policy & Internet*. https://doi.org/10.1002/poi3.185.

Svantesson, D. (2014). Between a Rock and a Hard Place: An International Law Perspective of the Difficult Position of Globally Active Intermediaries. *Computer Law & Security Review*, *30*, 348–356.

Teachout, Z. (2020). *Break 'Em Up: Recovering Our Freedom from Big Ag, Big Tech, and Big Money*. New York: All Points Books.

Urban, J. M., Karaganis, J., & Schofield, B. L. (2016). *Notice and Takedown in Everyday Practice*. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2755628.

Wagner, B., Rozgonyi, K., Sekwenz, M.-T., Cobbe, J., & Singh, J. (2020). Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 261–271. FAT* '20. Barcelona, Spain. https://doi.org/10.1145/3351095.3372856.

West, S. M. (2018). *Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms*. New Media & Society (May 8). https://doi.org/10.1177/1461444818773059.

Winseck, D. (2020). Vampire Squids, "the Broken Internet" and Platform Regulation. *Journal of Digital Media & Policy*, *11*(3) (November 1), 241–282.

Woods, L. (2019). The Duty of Care in the Online Harms White Paper. *Journal of Media Law*, *11*(1) (January 2), 6–17.

York, J. C. (2021). *Silicon Values*. Verso. https://www.penguinrandomhouse.com/books/667400/silicon-values-by-jillian-york/.