Yale Journal of Law & Technology, Volume 23

&

The Justice Collaboratory at Yale Law School

Present:

IN A NEW LIGHT:
SOCIAL MEDIA GOVERNANCE RECONSIDERED

Sudhir Venkatesh, Tom Tyler, Tracey Meares & Farzaneh Badiei[*]

The ubiquity with which platforms for online interaction have arisen and spread across the world has kept private companies, governments and the people using these platforms playing continual catch-up, trying to both utilize the new possibilities created by internet-based communication and protect users from both traditional and newly emerging harms that occur when interacting with others. Many of the problems emerging in online platforms mirror long-term issues associated with governing interactions in real world communities, while some are unique to the new internet world. Three types of governance are important. One is self-governance, the ability of users to cooperate with others to manage their own online interactions. A second is platform governance, the capacity of private vendors to effectively manage what occurs on their platforms. Finally, online communities may cross political boundaries but they exist within a complex matrix of local, national and international regulatory communities. These all play some role in governing the form and content of online platforms.

[*] Sudhir Venkatesh, Williams B. Ransford Professor of Sociology, Columbia University; Tom Tyler, Macklin Fleming Professor of Law and Professor of Psychology and Founding Director of The Justice Collaboratory, Yale Law School; Tracey Meares, Walton Hale Hamilton Professor of Law and Founding Director of The Justice Collaboratory, Yale Law School; Farzaneh Badiei, Director of Social Media Governance Initiative, The Justice Collaboratory, Yale Law School.

Our particular concern is with platform governance of these spaces for online interaction. Most platforms originated by conceptualizing themselves as pass-through architecture for interpersonal communications. Their creators no more imagined the prospect of regularly reading people's messages than the post office workers would imagine reading people's paper letters. Moreover, platform creators viewed their role as facilitating positive social communications among willing participants. And the rise of platforms for online interaction has facilitated traditional social communications, enabled people to make new connections and helped to maintain connections in better ways. Our social world has moved from the letter to the telephone to the Tweet or post. These new forms offer an unparalleled capacity for rapid and personalized connections across broad distances. Platforms *have* facilitated positive social communications among willing participants.

Of course, as more of our social world occurs online, the problems that plague the off-line social world follow. People can use online communications to threaten, bully and embarrass others in particularly effective ways. They can use internet platforms to push out negative messages about social and political issues, messages ranging from racism to hate speech and even advocating support for terrorism. The same tools that help people make new friends and form communities around a shared interest in gardening also enable extremists to recruit new members. The proliferation of negative content has forced platforms to become content regulators, whether or not they want to take on that role. In some cases, existing problems in the off-line world are not just perpetuated but, rather, intensified online. Algorithms, core to the technical infrastructure and scalability of these platforms, are prime examples of this phenomena. Widely used, many algorithms are aimed at mimicking human decision-making for efficiency and scalability's sake. Most often, these algorithms reinforce systematic biases of the individuals and organizations training, building, and deploying them. At worst, feedback loops in algorithms can inadvertently magnify these biases further marginalizing individuals or groups.

Many platforms have looked to the deterrence model common in legal settings as an initial framework through which to regulate content. Policy teams create rules and platforms create technical and operational mechanisms to evaluate user content against those rules. Those who violate rules by posting violating content get sanctioned in some way, typically with a graduated series of sanctions. Users' posts are removed, their accounts might be suspended for some period of time, or users might even be banned from a platform. In adopting this approach online platforms have inherited both the strengths and weaknesses of traditional law.

Studies show that in democratic societies like the United States, deterrence models work to change behavior, although not particularly well. Low-level offending presents an especially challenging environment for such models, a situation typical of online platforms. On the other hand, online platforms have notable advantages over real world legal authorities because they can more readily scan user platform behavior for rule conformity and have much greater control over when and how users can utilize the platforms. Still, problems like those faced by legal authorities arise. Some are related to defining and implementing rules for content moderation, which involves turning abstract ideas into practical and operational review guidelines used by a global workforce of agents reviewing vast amounts of content for violations of these rules. Since users imagine that their communications move more or less immediately to their intended audience, platforms have sought rapid algorithms to detect harmful content, moving the initial problem of flagging problematic content evaluation from human to machine.

Human review often follows flagging by machine algorithms, but that process takes time. Human review also occurs in response to people's complaints, so harmful content may be viewed by many users prior to any platform action. Platform owners, since they control access to their platform, can also more successfully sanction offenders than can real-world legal authorities. Here too, however, users can seek to evade sanctions or bans by using multiple accounts or moving to private sites.

Platform content, especially content that violates content moderation rules, is continually in the news, reflecting limitations in the existing governance models for content moderation. On the other hand, the newsworthiness of apparent content moderation failure may simply reflect the centrality that social media has assumed in people's social interactions.

These newsworthy moderation challenges also reflect a lack of consensus about what problematic content is and how to address it. On the one hand, there are calls for flagging or taking down material that some groups feel is problematic. At the same time, others complain about the suppression or exclusion of that same content they regard as valuable. What is desirable and what should be flagged or even banned depends upon underlying values and is an active debate. While this issue conflicts particularly with political speech, even efforts to limit nudity encounter differences in people's values about what forms of nudity are and are not offensive.

Regardless of their reasons, many people are dedicated to thinking through better governance models of online platforms. Here, a multidisciplinary group of researchers reconsider the issues involved in this rapidly evolving space and consider new ideas and alternative possibilities for social media governance. This issue brings together a group of prominent scholars using a broad array methods and theoretical perspectives to address platform governance in a new light and in an evidence-informed fashion.

Our aim for this special issue is to bring a few novel approaches to platform governance which can be applicable to social media and other online platforms. The different scholars included in this issue approach social media governance through different lenses, and sometimes use different terminology (e.g., "platforms" vs. "technology firms" vs. "social media companies"). Yet the common thread is the importance of exploring new ideas for managing the social impact, good and bad, that these large players have in our society. Our hope is that this issue will spur as lively a conversation about these topics as we had at the mini conference at which each of these papers was presented. These papers reflect not only the ideas of their authors but also the feedback from the distinguished group of scholars convened to comment upon them. To make progress upon these ideas we will need a dedicated cohort of people willing to think about these problems in a different way. This issue represents our effort to create such a group.

Rethinking Models of Social Media Governance

As noted, many platforms have reacted to the problems of negative content by trying to engage in some form of content moderation. This involves identifying problematic content ranging from nudity to hate speech. A review of both rules and strategies to enforce them reveals that platforms use the legal model of suppressing bad behavior through the threat or use of sanctions. Badiei, Meares & Tyler argue that this is a mistake. Platforms should encourage users to voluntarily internalize the rules and willingly follow rules and engage in positive behavior and healthy interactions. The key to this model is to change what users want to do and thereby discourage the emergence of bad behavior in the first place. This argument has two parts. The first mirrors recent reform efforts in criminal justice in recognizing that when people view rules and authorities as legitimate, they feel a responsibility to follow those rules and authorities. This strategy promotes rule adherence in a way that lessens the need for surveillance and sanctioning. It is especially important in an arena like online platforms in which most users are well intentioned and many rule violations come through a lack of awareness of the rules.

A legitimacy-based model has the second advantage of building identification with other people in the community, leading users to want to make their online communications positive, facilitating healthy interactions and vital online communities. Evidence demonstrates that it is possible to create online platforms that promote user identification with their communities and which enhance the legitimacy of platforms and of their regulatory efforts.

In a similar vein, Schoenebeck and Blackwell argue that social media platforms have often followed the traditional legal system in focusing on punishing offenders, without paying attention to how to mitigate conflicts or repair harm to victims. Social media platforms are punitive rather than reparative and focus on removing harmful content or users. They neglect the task of helping the victims of the abuse. These authors argue that platforms would benefit from adopting reparative approaches centered on global values such as dignity, accountability, and community. Although negative content may not be illegal, it still harms others, and platforms should adopt a broader perspective which recognizes the desirability of focusing on the well-being of those who have experienced negative online interactions.

Policies and Practices for Content Moderation

Although several contributors argue that platforms for online interaction overemphasize content moderation, content moderation still is necessary, so one must ask how can moderation best be achieved? Companies struggle to find ways to implement their desired goal of lessening or even eliminating exposure to "bad" content. They are trying to find ways to identify content that would be generally viewed as bad content. One of the more challenging examples of this struggle is found in the arena of politically or socially controversial content. Here there is often disagreement about what type of messaging is inappropriate and who should make such decisions. One approach that some platforms have used is not to remove content but to give it less priority in user feeds. Another approach, discussed by Wihbey, et al, is to post content but provide some type of warning or explanation, a practice called labelling. Such labelling can take different forms. It might involve an effort to correct factual errors and aim against misinformation. It can also be motivated by a desire to help users recognize alternative perspectives on a particular issue, including perspectives that are the opposite of their own or that are held by "experts" on a topic. Labels can encourage readers to read supplementary material that the platform believes clarifies or even contradicts a particular message.

Wihbey et al. analyze this specific governance method of "labeling." They do so with an epistemological approach. Their argument is that, despite the promise of labeling as a strategy, it has thus far been mostly tactical, reactive, and without strategic underpinnings. Wihbey et al. argue that social media companies have been struggling to devise and implement policies on handling misinformation that the public finds generally palatable. In place of consistently-enforced policies that are transparent to all parties, large platforms such as Twitter and Facebook have been responding to different instances of misinformation in a seemingly piecemeal fashion: downranking some posts, removing others, and labeling or "fact-checking" still others. This approach has led to social blowback, especially in those cases where algorithms are involved. They therefore argue against defining success as merely curbing misinformation spread. The healthy way of labeling is to consider it from an epistemic perspective and to take the "social" dimension of online social networks as a starting point. The strategy in this article emphasizes how the moderation system needs to improve the epistemic position and relationships of platform users—i.e., their ability to make good judgments about the sources and quality of the information with which users interact on the platform—while also respecting sources, seekers, and subjects of information.

Obviously, in order to govern online platforms by moderating content it is necessary to have criteria that define good and bad content. Often people feel that bad content is self-evident. For example, Supreme Court Justice Potter Stewart defined the Court's standards for obscenity by saying "I know it when I see it." Online platforms, in contrast, have developed elaborate codebooks for their human reviewers and have tried to develop computer programs which embody the same rules. This requires a two-step process. First, identifying principles (e.g., "no nudity"). Those rules then have to be elaborated into guidelines that are specific enough that they can be utilized by either a human coder or an algorithm.

Pineda is concerned with the origin of the principles and, in particular, with the question of whether there are any universal principles that can rise above the values of any particular society or culture. Social media platforms began in America and they sometimes employ general principles derived from America to determine their rules. Even if this were reasonable, the rise of alternative platforms in other societies makes this approach unrealistic. So where will standards come from in the future?

Pineda argues that we can best analyze the challenges of content governance by understanding the debates and conversations that take place about culture, cultural relativism, and the universality of human rights. In particular is the West imposing its values on everyone in the guise of "universal values"? How can we resolve this through anthropological means? The ongoing work of formulating "universal" content moderation policies will benefit from understanding the histories and debates in anthropology about cultural relativism and human rights universalism in order to avoid some of the pitfalls that are inherent in this kind of global governance. Anthropology can help us distinguish between values that are universal amid the difference in the expression of values across the world. Just like the universality of human rights has been scrutinized in global governance, the general standards that social media platforms have asserted have been contested.

Credibility Online: Who Do We Trust?

Social scientists have long argued that the willingness to trust other people is central to engaging in exchanges with others. Such exchanges frequently require people to take risks based upon the belief that the other people involved in an interaction have benevolent and sincere motivations and are not seeking to take advantage of them. People have developed strategies for evaluating the trustworthiness of others in real-world interactions. However, there are questions about the degree to which a similar level of trust can be established and maintained remotely, an issue central to rapidly emerging online platforms. The core question is whether a participant in an online market like eBay is willing to trust another in the same way that people have trusted others in their community in the past, and, as in real world interactions, what mechanisms can be identified to facilitate such trust and make online markets viable.

Parigi and Lainer-Vos argue that the rise of two-sided online markets and the centrality of reputation systems have undermined trust. Instead of trust being a byproduct of interpersonal interaction, thin trust in online markets demands methodical cultivation of trust in a mostly impersonal and domain-specific fashion.

Trust is central to exchange and cooperation. In offline situations people continually struggle to decide whom to trust and when to take risks by being vulnerable to others. If people never take risks, they gain little from being in markets. If people trust too uncritically, they may rely on others who do not keep their promises. Traditional discussions of trust emphasize the role of reputations in enabling trust. Someone who might break another's trust in one situation recognizes that if they acquire a reputation for being untrustworthy no one will exchange with them in the future. Reputations in traditional communities were a shared property, and people sought out and interacted with trustworthy others. Parigi and Lainer-Vos argue that the online world poses challenges for people trying to determine whether to trust someone else. Consequently, the nature of trust is changing in this new domain.

Future Research in Online Governance

This special issue represents our attempt to contribute to this growing need for rethinking online platform governance. Undoubtedly, we will continue this work through a network of interdisciplinary scholars within the Justice Collaboratory's Social Media Governance Initiative. As we think through what future research could contribute to this conversation, it is important to highlight some areas we are particularly concerned about, like shifting the focus of scholars and policy-makers towards the design, architecture, and infrastructure decisions that shape governance.

If the prevailing model of content moderation is not the most desirable way to manage platforms, a key question is why this model exists and how it was built. At the center of the organizational culture of most online platforms is the product group. This is the group that manages the architecture of the platform: many hundreds of engineers, designers, and product managers. Because this group dominates these companies, the issue of content moderation within these organizations has been generally viewed as a technical one, something amenable to management through simple screening algorithms that can detect and remove nudity or hate speech.

The insight that content moderation is viewed as a technical problem within the purview of product teams helps to illuminate why external regulation efforts have been problematic. External constituencies typically interface with the legal and managerial elements of online platform companies—typically policy teams rather than these product teams. This means that both scholars and those seeking platform changes rarely look at product design culture and how it shapes content moderation in technology firms. The fact that content governance is housed in product units reflects the history of the evolution of platforms, which was focused on solving technical problems, not addressing issues complex social issues of content acceptability across the globe. Some of the recent efforts to share data with scholars through Transparency Reports or create Oversight Boards are examples in which the corporate leadership draws energy away from product divisions that have more substantial impact on governance of platform users.

As more public attention is paid to the impact of social media and other internet companies, it would be worthwhile for outsiders to redirect some their efforts toward the technology creation efforts of product teams. As we look towards furthering the conversation over platform governance, we need to spend more time thinking about platform architecture and the design of infrastructure in addition to the current focus on the rules themselves. Safety in automobiles can be a very helpful analogy in this regard. While speed limits and other rules of the roads are important to ensure public safety, far more critical in saving lives are the design of the cars we drive—airbags, crumple zones, or seatbelts—and infrastructure of the roads we drive on—rumble strips, clear signage, or banked turns.

This discussion also highlights the issue of platform motivations. Newspapers struggle with the problem that sensational news sells papers. In the same way, online platforms are for-profit entities. Their profits flow from putting ads in front of their users, selling knowledge harvested about its users to advertisers allowing vendors to target likely candidates for their products. This means that if extreme or salacious content attracts attention, it is to the benefit of the company to highlight such content in order to attract and retain the attention of their users. Content moderation is in conflict with this business model. As a consequence, it is sometimes difficult to discern whether companies are actually interested in effectively moderating such content or are interested in presenting an image of civil responsibility that can fend of government regulation, oversight and organized consumer push back. Discerning the internal dynamics of organizations running online platforms is also important in future study of online governance.

CONCLUSION

We are hopeful that this material contributes to the debate about how humanity might govern itself online. These papers demonstrate how to apply interdisciplinary approaches to social platform governance and go beyond the currently dominant governance mechanisms which this group collectively argues have not so far been effective. We believe the papers provide an important contribution to the technology governance landscape and we thank the editorial board at the *Yale Journal of Law and Technology* for their collaboration in publishing this special issue.