From Status Update to Social Media Contract

Author : Rebecca Tushnet

Date : November 29, 2017

Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 **Harvard L. Rev.** (forthcoming 2017), available at SSRN.

Under current US First Amendment jurisprudence, the government can do very little to regulate speech online. It can penalize fraud and certain other kinds of false or potentially misleading speech; direct true threats; and infringement of intellectual property rights and related speech. But it cannot penalize most harassment, hate speech, falsity, and other speech that does immediate harm. Nor can the government generally bar particular *speakers*. Last Term, the Supreme Court struck down a provision of state law that tried to prevent convicted sex offenders from participating in "social media" where minors might also be participating.

There are good reasons for most of the limits the courts have imposed on the government's speech-regulating powers—yet those limits have left a regulatory vacuum into which powerful private entities have stepped to regulate the speech of US social media users, suppressing a lot of speech that the government can't, and protecting other speech despite their power to suppress it. The limits these intermediaries impose, with some important exceptions, look very similar whether the speech comes from the US or from a country that imposes heavier burdens on intermediaries to control the speech of their users. Klonick's fascinating paper explores the evolution of speech regulation policies at major social media companies, particularly Twitter and Facebook, along with Alphabet's (Google's) YouTube.

Klonick finds "marked similarities to legal or governance systems with the creation of a detailed list of rules, trained human decision-making to apply those rules, and reliance on a system of external influence to update and amend those rules." One lesson from her story may be the free speech version of ontogeny recapitulating phylogeny: regardless of what the underlying legal structure is, or whether an institution is essentially inventing a structure from scratch, speech regulations pose standard issues of definition (defamation and hate speech are endlessly flexible, not to mention intellectual property infringements), enforcement (who will catch the violators?), and equity/fairness (who will watch the watchmen?).

Klonick's research also provides important insights on the relative roles of algorithms and human review in detecting and deterring unwanted content. While her article focuses on the guidelines followed by human decision-makers, those fit into a larger context of partially automated screening. Automated screening for child pornography seems to be a relative success story, as she explains. However, as many interested parties have pointed out in response to the Copyright Office's inquiry on §512's safe harbors and private content protection mechanisms, even with automated enforcement and "claiming" by putative copyright owners via Content ID, algorithms cannot avoid problems of judgment and equitable treatment, especially when some copyright owners have negotiated special rights to override the DMCA process, and keep contested content down regardless of its fair use status, once it's been identified by Content ID.

Klonick's account can also usefully be read alongside Zeynep Tufekci's Twitter and Tear Gas: The Power and Fragility of Networked Protest. Tufekci covers some aspects of speech policies that are particularly troubling, including the misuse of Facebook's "real name" policy to suppress activists in countries where using a formal name could potentially be deadly; targeted, state-supported attacks on activists that involve reporting them for "abuse" and hate speech; and content moderation that can be politically ignorant, or worse: "in almost any country with deep internal conflict, the types of people who are most likely to be employed by Facebook are often from one side of the conflict—the side with more power and privileges." Facebook's team overseeing Turkish content, for example, is in Dublin,

disadvantaging non-English speakers and women (whose families are less likely to be willing to relocate for their jobs). Similarly, Facebook's response to the real-name problem is to allow use of another name when it's in common use by the speaker, but that usually requires people to provide documents such as school IDs. As Tufekci points out, documents using an alternate identity are most likely to be available to people in relatively privileged positions in developed countries, thus muting their protest but leaving similar people without such forms of ID exposed.

These details of implementation are far more than trivial. And Tufekci's warning that governments quickly learn how to use, and misuse, platform mechanisms for their own benefit is a vital one. The extent to which an abuse team can be manipulated will, I expect, soon become a separate challenge for the content policy teams Klonick documents—if they decide to resist that manipulation, which is not guaranteed. Some of these techniques, moreover, resist handling by an abuse team even when identified. When government-backed teams overwhelm social media with trivialities in order to distract from a potentially important political event, as is apparently common in China, what policies and algorithms could identify the pattern, much less sort the wheat from the chaff?

Along with this comparison, Klonick's piece offers the opportunity to revisit some relatively recent techno-optimists—West Coast code has started to look in places more like outsourced Filipino or Indian area codes, so what does that mean for internet governance? Consider Clay Shirky's *Cognitive Surplus: Creativity and Generosity in a Connected Age*, a witty book whose examples of user-generated activism now seem dated, only seven years later, with the rise of "fake news" disseminated by foreign content farms, GamerGate, and revenge porn. It's still true that, as Joi Ito wrote, "you should never underestimate the power of peer-to-peer social communication and the bonding force of popular culture. Although so much of what kids are doing online may look trivial and frivolous, what they are doing is building the capacity to connect, to communicate, and ultimately, to mobilize." Because of this power, a legal system that discourages you from commenting on and remixing the first things you love, in communities who love the same thing you do, also discourages you from commenting on and remixing everything else. But what Klonick's account makes clear is that discouragement can come from platforms as well as directly from governments, whether because of over-active filters such as Content ID that suppress remixes or because of more directly politicized interventions such as those Tufekci discusses.

Shirky's book, like many of its era, was relatively silent about the role of government in enacting (or suppressing) the changes promoted by people taking advantage of new technological affordances. Consider one of Shirky's prominent examples of the power of (women) organizing online: a Facebook group organized to fight back against anti-woman violence perpetrated in the Indian city of Mangalore by the religious fundamentalist group Sri Ram Sene. As Shirky tells it, "[p]articipation in the Pink Chaddi [underwear] campaign demonstrated publicly that a constituency of women were willing to counter Sene and wanted politicians and the police to do the same…. *[T]he state of Mangalore arrested Muthali and several key members of Sene* … as a way of preventing a repeat of the January attacks." (Emphasis mine.) The story has a happy ending because actual government, not "governance" structures, intervened. How would the content teams at Facebook react if today's Indian government decided that similar protests were incitements to violence?

The fact that internet intermediaries have governance aspirations without formal government power (or participatory democracy) also directs our attention to the influences on the use of that power. Klonick states that "platforms moderate content because of a foundation in First Amendment norms, corporate responsibility, and at the core, the economic necessity of creating an environment that reflects the expectations of its users. Thus, platforms are motivated to moderate by both the Good Samaritan purpose of § 230, as well as its concerns for free speech." But note what drops out of that second sentence—explicit acknowledgement of the profit motive, which becomes both a driver of some speech protections and a reason, or an excuse, for some speech suppression. Pressure from advertisers, for example, led YouTube to crack down on "pro-terrorism" speech on the platform. Klonick also argues that "platforms are economically responsive to the expectations and norms of their users," which leads them "to both take down content their users don't want to see and keep up as much content as possible," including by pushing back against government takedown requests. But this seems to me to equivocate about who the relevant "users" are—after all, if you're not paying for a service, you're the product it's selling, and content that advertisers or large copyright

owners don't want to see may be far more vulnerable than content that individual participants don't want to see.

One question Klonick's story raised for me, then, was what a different system might look like. What if platforms were run the way public libraries are? Libraries are the real "sharing" economies, and in the US have resisted government surveillance and content filtering as a matter of mission. Similarly, the Archive of Our Own, with which I am involved, has user-centric rules that don't need to prioritize the preservation of ad revenue. Although these rules are hotly debated within fandom, because what is welcoming to some users can be exclusionary to others, they are distinctively mission-oriented. (I should also concede that size, too, makes a difference—eventually, a large enough community that includes political content will attract government attention; Twitter hasn't made a profit, but it has received numerous subpoenas and national security letters.)

Klonick suggests that the key to optimal speech regulation for platforms is some sort of participatory reform, perhaps involving both procedural and substantive protections for individual users. In other words, we need to reinvent the democratic state, embedding the user/citizen in a context that she has some realistic chance to affect, at least if she knows her rights and acts in concert with other users. The obvious problem is the one of transition: how will we get from here to there? Klonick understandably doesn't take up that question in any detail. Absent the coercive power of real law, backed by guns and taxes, it's hard for me to imagine the transition to participatory platform governance. Moreover, the same dynamics that brought us *Citizens United* make it hard to imagine that corporate interests—both platform and advertiser—would accede to any such mandates, likely raising First Amendment objections of their own.

Klonick's article helps to identify how individual speech online is embedded in structures that guide and constrain speakers; its descriptive account will be very useful to understanding these structures. I worry, however, that understanding won't be enough to save us. We want to think well of our governors; we don't want to be living in 1984, or Brave New World. But the development of intermediary speech policies tells us, among other things, that we might end up looking from man to pig, and pig to man, and finding it hard to tell the difference.

Disclosure: Kate Klonick is a former student of mine, though this paper comes from her work years later.

Cite as: Rebecca Tushnet, *From Status Update to Social Media Contract*, JOTWELL (November 29, 2017) (reviewing Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 **Harvard L. Rev.** (forthcoming 2017), available at SSRN), https://cyber.jotwell.com/from-status-update-to-social-media-contract/.