Shifting the Content Moderation Paradigm

Author : Ari Waldman

Date : March 1, 2022

evelyn douek, *Content Moderation as Administration* (Jan. 12, 2022), available on SSRN.

As law-and-technology scholar evelyn douek explains in her eye-opening, scholarly, and well-written *Content Moderation as Administration*, the conventional account of content moderation is wrong and its policy implications are off the mark. douek argues that we should toss aside the assumption that content moderation is a series of individual decisions made by people and computers acting as judges. The better way to think about it is as a process of ex ante rights administration and institutional design. Instead of learning lessons from judicial process, we need to learn from administrative law.

A system of immeasurable scale purportedly designed to reflect liberal First Amendment principles, content moderation now includes algorithms and artificial intelligence, armies of third-party moderators from the Global South paid very little to make decisions in seconds, and, a lot of money for Silicon Valley executives. Of course, this has led to repeated and repeatedly horrible results. Content moderation rules and practices facilitated genocide, helped swing elections toward fascists, and routinely and systematically censored queer and nonnormative sexual content. Right wing politicians got in on the act, as well, claiming designed-in and as-applied anti-conservative bias when the evidence proved the opposite. Facebook responded by creating an oversight board with a lot of fanfare, but very little power.

Through it all, the vision of content moderation has remained roughly the same: *ex* ante automated filtering and ex *post* judicial*ish* review of whether user-generated content violated platform policies. If this "first wave" of content moderation scholarship is right, then presumably, the best way to protect speech and social media users is to demand procedural due process*ish* protections: transparency and rights to appeal. And that's precisely what those members of Congress who are legitimately concerned about content moderation have proposed.

The standard picture of content moderation is like an old Roman emperor whose thumbs up or thumbs down decides the fate of a gladiator: some all-powerful person or all-powerful thing is deciding whether a post stays up or comes down. Content moderation, then, happens post-by-post.

douek explains that almost none of that is helpful or correct. As many scholars have argued, content moderation involves an assemblage of people and things. Platforms do more than just decide to keep content up or take it down. And, most importantly, these misguided assumptions contribute to misguided policy.

Case-by-case *ex post* review misses systemic failures. It also provides inadequate remedies: a moderator could take something down or put something back up, leaving the problems of training and institutional design untouched. And the cycle will continue as long as the structural problem remains. Case-by-case review also lends itself to privacy theatre like the Facebook Oversight Board. By the nature of its design, it may eventually address a few takedown decisions, but has little to no impact on how the whole system works.

In place of this misguided vision, douek proposes a "second wave" of content moderation scholarship, discourse, and solutions. douek deftly argues that content moderation is a product of *ex ante* system

design. It is one result of a larger institutional structure that frames the flow of all sorts of information. Content moderation is also the product of multiple corporate goals, not just the ostensible desire to reflect and perpetuate a liberal vision of free speech. Policy reform should reflect that.

douek suggests that one way to do that is to learn from the literature in collaborative governance, an approach to administrative regulation of corporations that involves public and private entities working together to achieve mutual goals. It benefits from private expertise while using a wide toolkit—audits, impact assessments, transparency reports, ongoing monitoring, and internal organizational structures, among others—cabining private discretionary decision-making by making firms accountable to the public and to regulatory agencies. Proponents see the multi-stakeholder model of governance as a more effective way of governing fast-changing and technologically complex systems, an argument made in profound and powerful detail by Margot Kaminski.

Collaborative governance is meant to help regulators supervise vast organizational systems ex ante before they do something wrong. Its *ex ante* approach and process toolkit are supposed to instantiate public values into every phase of organizational function. In that way, it is supposed to influence everyone, create systems up front, and foster the development of organizations more attuned to popular needs and values.

douek makes a compelling argument that collaborative governance is the better way to approach content moderation, both conceptually and as a matter of policy. Instead of an *ex post* appeal process, the collaborative governance approach means integrating officers whose entire jobs are to advocate for fair content moderation. It means giving those employees the safety and separation they need from departments with contrary motivations in order to do their work. It means transparency about underlying data and systemic audits of how the system works.

What's so compelling about *Content Moderation as Administration* is that it changes the paradigm and pushes us to respond. douek has described a new and more compelling way of looking at content moderation. We all have to learn from their work, especially those of us writing or interested in writing about content moderation, collaborative governance, or both. The challenge, of course, will be guarding against managerialism and performative theatre in the content moderation space. Compliance models are at best risky when not subordinated to the rule of law and, in particular, a vision of the rule of law attuned to the unique pressures of informational capitalism. But those questions come next. *Content Moderation as Administration* does an article does an outstanding job of challenging the conventional account that has been at the core of content moderation scholarship for so long.

Cite as: Ari Waldman, *Shifting the Content Moderation Paradigm*, JOTWELL (March 1, 2022) (reviewing evelyn douek, *Content Moderation as Administration* (Jan. 12, 2022), available on SSRN), https://cyber.jotwell.com/shifting-the-content-moderation-paradigm/.