Moderation's Excess

Author : Annemarie Bridy

Date : March 27, 2020

Hannah Bloch-Wehba, *Automation in Moderation,* Cornell Int'l L. J. (forthcoming), available at SSRN.

In 2012, Twitter executive Tony Wang proudly described his company as "the free-speech wing of the free-speech party."[1] Seven years later, *The New Yorker*'s Andrew Marantz declaimed in an op-ed for *The New York Times* that "free speech is killing us."[2] The intervening years saw a tidal shift in public attitudes toward Twitter and the world's other major social media services—most notably Facebook, YouTube, and Instagram. These global platforms, which were once widely celebrated for democratizing mass communication and giving voice to the voiceless, are now widely derided as cesspools of disinformation, hate speech, and harassment. How did we get to this moment in the Internet's history? In *Automation in Moderation*, Hannah Bloch-Wehba chronicles the important social, technological, and regulatory developments that have brought us here. She surveys in careful detail both how algorithms have come to be the arbiters of acceptable online speech and what we are losing in the apparently unstoppable transition from manual-reactive to automated-proactive speech regulation.

Globally, policy makers are enacting waves of new legislation requiring platform operators to scrub and sanitize their virtual premises. Regulatory regimes that once protected tech companies from liability for their users' unlawful speech are being dramatically reconfigured, creating strong incentives for platforms to not only remove offensive and illegal speech after it has been posted but to prevent it from ever appearing in the first place. To proactively manage bad speech, platforms are increasingly turning to algorithmic moderation. In place of intermediary liability, scholars of Internet law and policy now speak of intermediary accountability and responsibility.

Bloch-Wehba argues that automation in moderation has three major consequences: First, user speech and privacy are compromised due to the nature and limits of existing filtering technology. Second, new regulatory mandates conflict in unacknowledged and unresolved ways with longstanding intermediary safe harbors, creating a fragmented legal landscape in which the power to control speech is shifting (in ways that should worry us) to state actors. Third, new regulatory mandates for platforms risk entrenching rather than checking the power of mega-platforms, because regulatory mandates to deploy and maintain sophisticated filtering systems fall harder on small platforms and new entrants than on tech giants like Facebook and YouTube.

To moderate the harmful effects of auto-moderation, Bloch-Wehba proposes enhanced transparency obligations for platforms. Transparency reports began as a voluntary effort for platforms to inform users about demands for surveillance and censorship and have since been incorporated into regulatory reporting obligations in some jurisdictions. Bloch-Wehba would like to see platforms provide more information to the public about how, when, and why they deploy proactive technical measures to screen uploaded content. In addition, she calls for disaggregated and more granular reporting about material that is blocked, and she suggests mandatory audits of algorithms to make their methods of operation visible.

Transparency alone is not enough, however. Bloch-Wehba argues that greater emphasis must be placed on delivering due process for speakers whose content is negatively impacted by auto-moderation decisions. She considers existing private appeal mechanisms, including Facebook's much-publicized

"Supreme Court," and cautions against our taking comfort in mere "simulacr[a] of due process, unregulated by law and constitution and unaccountable to the democratic process."

An aspect of Bloch-Wehba's article that deserves special attention given the global resurgence of authoritarian nationalism is her treatment of the convergence of corporate and state power in the domain of automated content moderation. Building on the work of First Amendment scholars including Jack Balkin, Kate Klonick, Danielle Citron, and Daphne Keller, Bloch-Wehba describes a troubling dynamic in which platform executives seek to appease government actors—and thereby to avoid additional regulation—by suppressing speech in accordance with the prevailing political winds. As Bloch-Wehba recognizes, this is a confluence of interests that bodes ill for expressive freedom in the world's increasingly beleaguered democracies.

Automation in Moderation has much to offer for died-in-the-wool Internet policy wonks and interested bystanders alike. It's a deep and rewarding dive into the most difficult free speech challenge of our time, offered to us at a moment when public discourse is polarized and the pendulum of public opinion swings wide in the direction of casual censorship.

1. Josh Halliday, *Twitter's Tony Wang: "*We are the free speech wing of the free speech party*,"* **Guardian**, Mar. 22, 2012.
2. Andrew Marantz, *Free Speech Is Killing Us*, **NY Times**, Oct. 4, 2019.

Cite as: Annemarie Bridy, *Moderation's Excess*, JOTWELL (March 27, 2020) (reviewing Hannah Bloch-Wehba, *Automation in Moderation,* Cornell Int'l L. J. (forthcoming), available at SSRN), https://cyber.jotwell.com/moderations-excess/.