The Constant Trash Collector: Platforms and the Paradoxes of Content Moderation

Author : Rebecca Tushnet

Date : July 25, 2019

Tarleton Gillespie, **Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media** (2018).

Tarleton Gillespie's important book *Custodians of the Internet* unpacks the simultaneous impossibility and necessity of content moderation, highlighting nuance rather than answering questions. Within big companies, content moderation is treated like custodial work, like sweeping the floors—and recent revelations reinforce that the abjectness of this work seems to contaminate those who do it. The rules are made by people in positions of relative power, while their enforcement is traumatic, poorly-paid, outsourced scutwork. But for major platforms, taking out the trash—making sure the site isn't a cesspool—is in fact their central function.

Gillespie urges us to pay attention to the differences between a content policy—which is a document that both tries to shape the reactions of various stakeholders and is shaped by them—and actual content moderation; both are vitally important. (Facebook's newly announced "Supreme Court" is on the former side: it will make important decisions, but make them at a level of generality that will leave much day-to-day work to be done by the custodial staff.) Every provision of a content policy represents a horror story *and also something that will definitely be repeated*. Gillespie is heartbreakingly clear on the banality of evil at scale: "a moderator looking at hundreds of pieces of Facebook content every hour needs more specific instructions on what exactly counts as 'sexual violence,' so these documents provide examples like 'To snap a bitch's neck, make sure to apply all your pressure to the middle of her throat'—which Facebook gives a green checkmark, meaning posts like that should stay." (P. 112.)

Given the scale of decision-making, what else could be done? The Apple App Store preapproves apps, but constantly struggles because of the need for speed and the inevitable cries of censorship when Apple decides an app is too political. And Apple can get away with preapproval only because the flood of apps is orders of magnitude less than the flood of Instagram photos or similar content, and because app developers are relatively more likely to share ideas about reasonable expectations of a platform than are web users in general.

Most other platforms have relied on community flagging in the first instance. Relying on the users to report the need for clean-up is practically convenient, but also grants "legitimacy and cover" by signaling that the platform is listening to and trying to help users who are being harmed, while still leaving ultimate control in the platform's hands. Mechanisms to report content can also be used by harassers, such as the person who reported hundreds of drag queens for violating Facebook's real-name policy. YouTube sees flags surge when a new country gets access to the platform; Gillespie's informant attributed this to users' lack of knowledge of the YouTube community's values. Gillespie sees in this reaction an "astounding assumption: a different logic would be to treat this new wave of flags as some kind of expression of values held by this new user population." (P. 130.) Flagging will remain contested not just because values vary, but also because flagging relies on a devoted but small group of users to voluntarily police bad behavior, in conflict with another small group that deliberately seeks to cause mayhem.

There's another approach, self-labeling, which the nonprofit Archive of Our Own (with which I volunteer) tries: ask users to evaluate their own content as they post it, and provide filters so users can avoid what they don't want. This distributes the work more equitably. But tagging is time-consuming and can deter use, so commercial platforms make self-tagging limited, either relying on defaults or on rating entire users' profiles, as on Tumblr. But self-tagging raises problems with consistency, since Tumblr users don't always agree on what's "safe," not to mention what happens

when Tumblr itself decides that "gay" is unsafe by definition. I'm obviously invested in the AO3; precisely because his analysis is so incisive, I wish Gillespie had spent a little time on what noncommercial platforms decide to do differently here and why.

Gillespie also has a great discussion of automated filtering. It's relatively easy to compare images to hashes that screen out known child pornography images. But that relative ease is a product of law, technology, and policy that is hard to replicate for other issues. The database that allows screening is a high priority for platforms because of the blanket illegality of the content, and hashing known images is a far simpler task than identifying new images or their meaning in context. Microsoft developed the software, but recognized it as good PR to donate it for public use rather than keeping it as a trade secret or licensing it for profit, which isn't true of other filtering algorithms like YouTube's Content ID. Machine learning is trying to take on more complex tasks, but it's just not very good yet. (After the book came out, we learned that Amazon's machine learning tool for assessing resumes learned to discriminate against women—a machine learning algorithm aimed at harassment or hate speech will likewise replicate the biases of the arbiters who train the computer, and who tell it that "choke a bitch" is fine.) Also, we don't really know what constitutes a "good" detection rate. Is 95% success identifying nude images good? What about a false positive rate of 1%? Are the false positives/negatives biased, the way facial recognition software tends to perform worse with darker-skinned faces?

Nor are the choices limited to removal; algorithmic demotion or screening allows people who know that the content exists to find it, while making it harder for others to stumble across it. But this makes platforms' promise of sharing much more complicated: to whom are you visible, and when? These decisions can be hard for affected users to discover, much less understand, and they're particularly important for marginalized groups. One good example is Tumblr's shadow-banning of search terms like "porn" or "gay" on its app, with political consequences; similarly, it turns out that TripAdvisor reviewers may discover that they can't use "feminism" or "misogyny" in their reviews (highlighting that algorithmic demotion always interacts with other policy choices). Meanwhile, YouTube and Twitter curate their trending pages to avoid sexual or otherwise undesired content, so it's not really what's trending but only an undeclared subset, which curation nonetheless never quite manages to avoid controversy or harm, as platforms rediscover every few months. Amazon does similar things with best-sellers to make sure that shapeshifter porn doesn't get recommended to people who haven't already expressed an interest in it. Users can manipulate this differential visibility, too, as we learned with targeted Facebook ads from Russians and others in the 2016 US presidential campaign.

Again, it might be worthwhile to consider the potential alternatives: the noncommercial AO3, like Wikipedia, has done very little to shape users' searches, unlike YouTube's radicalization machine.

In concluding, Gillespie judges content moderation to be so difficult that "all things considered, it's amazing that it works at all, and as well as it does." (P. 197.) Still, handing it over to private, for-profit companies, with very few accountability or transparency mechanisms, isn't a great idea. At a minimum, he argues, platforms should have to be able to explain "why a post is there and how I should assess." (P. 199.)

Gillespie suggests that Section 230 of the Communications Decency Act may warrant modification. He argues that it should still provide immunity for pure conduits and good faith moderation. "But the moment that a platform begins to select some content over others, based not on a judgment of relevance to a search query but in the spirit of enhancing the value of the experience and keeping users on the site," (P. 43) it should be more accountable for the content of others. (I doubt this distinction would work at all.) Or, the safe harbor could be conditioned on having obligations such as meeting minimum standards for moderation, perhaps some degree of transparency or specific structures for due process/appeal of decisions. It is perhaps unsurprising that Facebook's splashiest endeavor in this area, its "Supreme Court," won't provide these kinds of protections. It's not in Facebook's interest to limit its own flexibility in that way even if it is in Facebook's interest to publicly perform adherence to certain general principles. This performance may well reflect sincere substantive commitments, but that also has advantages in fending off further regulation and in making bigness look good, because only big platforms like Facebook can sustain a "Supreme Court." Although I have

serious concerns about implementation of any procedural obligations (related to my belief that antitrust law would be a better source of regulation than blanket rules whose expense will ensure that no competitors to Facebook can arise), for purposes of this brief review it is probably more useful to note that the procedural turn is Gillespie's own version of neutrality on the content of content policies. Authoritarian constitutionalism—where the sovereign adheres to principles that it announces but does not concede to its subjects the right to choose those principles—seems to be an easier ask than democratic constitutionalism for platforms.

Gillespie also suggests structural changes that wouldn't directly change code or terms of service. He argues for greater diversity in the ranks of engineers, managers, and entrepreneurs, who currently tend to be from groups that are already winners and don't see the downsides of their libertarian designs. This would be a kind of virtual representation that wouldn't involve actual voting, whether for representatives or for policies, but would likely improve platform policymakers' ability to notice certain kinds of harms and needs.

Innovation policy has focused on presenting and organizing information and content creation by users, not on innovation in governance and design or implementation of shared values; it could do otherwise. Gillespie's most provocative question might be: What if we built platforms that tried to elicit civic preferences instead of consumer preferences? If only we knew how.

Cite as: Rebecca Tushnet, *The Constant Trash Collector: Platforms and the Paradoxes of Content Moderation*, JOTWELL (July 25, 2019) (reviewing Tarleton Gillespie, **Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media** (2018)), https://cyber.jotwell.com/the-constant-trash-collector-platforms-and-the-paradoxes-of-content-moderation/.