Gillespie

The human labor of moderation

From *Custodians of the internet* (2018)

In May 2017, the Guardian published a trove of documents it dubbed “The Facebook Files.” These documents instructed exactly what content moderators working on Facebook’s behalf should remove, approve, or escalate to Facebook for further review. The document offers a bizarre and disheartening glimpse into a process that Facebook and other social media platforms generally keep under wraps.

The most important revelation of these leaked documents is the fact that they had to be leaked. These documents were not meant ever to be seen by the public. They instruct and manage the thousands of “crowdworkers” responsible for the first line of human review of Facebook pages, posts, and photos that have been either flagged by users or identified algorithmically as possibly violating the site’s guidelines.

The labor that platforms put toward moderation, and the labor we as users are conscripted to perform as part of this project, are not just part of how platforms function, they constitute it. Platforms are made by the work that goes into content moderation, and they are not platforms without it.

Mapping out the dispersed, shifting, and interwoven labor force in more detail is difficult. Most of this work is hidden—some happens inside the corporate headquarters, some happens very, very far away from there, obscured behind the shields of third-party partners. The rest is portioned out to users in ways so woven into the mundane use of platforms that it is hard to notice at all.

Even the bits that show themselves—the flags and complaint mechanisms, the ratings and age barriers, even the occasional deletion—reveal little about how they fit into the larger project of moderating an entire platform. Users who have run up against moderation decisions may find it difficult to inquire about them, seek an audience with anyone at the platform, lodge a complaint, or request an appeal.

Each platform makes different choices about how to arrange and divide out this labor, where to put different kinds of decision processes, how to reconcile challenging cases. My aim is to make generally clear what is involved: platforms currently impose moderation at scale by turning some or all users into an identification force, employing a small group of outsourced workers to do the bulk of the review, and retaining for platform management the power to set the terms.

Internal Teams

At the top, most platforms have an internal policy team charged with overseeing moderation. The team sets the rules of that platform, oversees their enforcement, adjudicates the particularly hard cases, and crafts new policies in response. These are usually quite small, often just a handful of full- time employees. At a few platforms the team is an independent division; at others it sits under the umbrella of “trust and safety,” “community outreach,”customer service, or technical support. At others, setting policy and addressing hard cases is handled in a more ad hoc way, by the leadership with advice from legal counsel.

These teams are obscure to users, by design and policy. They are difficult for users to reach, and the statements and policies they generate are often released in the voice of the company itself. All together they are a surprisingly small community of people.

In their earliest days, many platforms did not anticipate that content moderation would be a significant problem. Some began with relatively homogenous user populations who shared values and norms with one another and with the developers—for example, back when Facebook was open only to tech- savvy Ivy League university students. Many of the social norms that first emerged were familiar from college life, and the diversity of opinions, values, and intentions would be attenuated by the narrow band of people who were even there in the first place. Other sites, modeled after blogging tools and searchable archives, subscribed to an "information wants to be free" ethos that was shared by designers and participants alike.

In fact, in the early days of a platform, it was not unusual for there to be no one in an official position to handle content moderation. Often content moderation at a platform was handled either by user support or community relations teams, generally more focused on offering users technical assistance. As these sites grew, so did the volume and variety of concerns coming from users. Platforms experienced these in waves, especially as a platform grew in cultural prominence, changed dramatically in its demographics, or expanded to an international audience. Some tried to address these growing concerns the way online discussion groups had, through collective deliberation and public rule-making.

Today the teams that oversee content moderation at these platforms remain surprisingly small, as much of the front-line work handled by these once-150- strong teams has been outsourced. Again, it is difficult to know exactly how many. Because the work of content moderation is now so intertwined with legal policy, spam, privacy, the safety of children and young users, ad policies, and community outreach, it really depends on how you count. Individual employees themselves increasingly have to obscure their identities to avoid the wrath of trolls and harassers. Most of all, platforms are not forthcoming about who does this work, or how many, or how. It is not in their interest to draw attention to content moderation, or to admit how few people do it.

But when a policy is reconsidered or a tricky case is discussed, such considerations are often being made by a very small group of people, before being imposed as a rule that potentially affects millions of users.

In addition, the group of people doing this kind of work is not only small, it is socially and professionally interconnected. Many of these platforms have their corporate headquarters in a tight constellation around San Francisco, which means people with this particular set of skills often move between companies professionally and circulate socially. Given its unique set of challenges, it is still a tiny group of people who have become expert in the task of large-scale platform moderation; members of this group have quickly become familiar with one another through professional conferences, changes in employment, and informal contact.

As they increasingly face legal, political, and public relations implications, some platforms have begun drawing employees from outside the company to fi t moderation teams with experts in sexual assault, antiterrorism, child exploitation, and so on. Still, the policies of these enormous, global platforms, and the labor crucial to their moderation efforts, are overseen by a few hundred, largely white, largely young, tech-savvy Californians who occupy a small and tight social and professional circle.

Crowdworkers
As recently as 2014, Twitter was still claiming, “Every report by a user is reviewed by a member of Twitter’s Trust and Safety team.” Even for Twitter, which has leaner rules than similar- sized platforms, this statement is hard to believe—if what it meant was its handful of permanent employees. But Twitter, like many social media platforms, now employs a substantially larger group of people to provide a first wave of review, beneath the internal moderation team.

They might be employed by the platform, at the home office, or in satellite offices located around the world in places like Dublin and Hyderabad. But more and more commonly they are hired on a contract basis: as independent contractors through third- party “temp” companies, or as on-demand labor employed through crowd work services such as Amazon’s Mechanical Turk, Upwork, Accenture, or TaskUs—or both, in a two- tiered system. The leaked 2017 documents discussed at the start of this chapter were the removal instructions provided by Facebook to its crowdworkers, to guide, harmonize, and speed their review of flagged content.

Crowdworkers are now used as a first- response team, looking at flagged posts and images from users and making quick decisions about how to respond. The moderators following those leaked instructions would largely be looking at content already reported or “flagged” by Facebook users. They would either “confirm” the report (and delete the content), “unconfirm” it (the content stays), or “escalate” it, passing it up to Facebook for further review.

This is a significant workforce, with an unseemly task: as one moderator put it, “Think like that there is a sewer channel and all of the mess/dirt/waste/shit of the world flow towards you and you have to clean it.”

Facebook currently promises some kind of response within twenty- four hours, and in 2016 all of the major platforms promised European lawmakers to ensure review of possible terrorist or extremist content within a one- day window. To meet such a requirement, human review must be handled fast. “Fast” here can mean mere seconds per complaint—approve, reject, approve—and moderators are often evaluated on their speed as well as their accuracy, meaning there is reward and pressure to keep up this pace. Each complaint is thus getting just a sliver of human attention, under great pressure to be responsive not just to this complaint, but to the queue of complaints behind it.

These crowdworkers are obscured intentionally and by circumstance.
Many are in parts of the world where labor is cheap, especially the Philippines and India, far from both the platform they work for and the users they are moderating; they are also distanced from the company through contract labor arrangements and the intervening interfaces of the crowdwork platforms that employ them and organize their labor. Work conditions can be grim. For a widely read 2014 Wired article, Adrian Chen investigated workers in the Philippines employed by crowdworker platform TaskUs to do content moderation for U.S.- based platforms, under very different working conditions from those one might find in Silicon Valley.

The pay is meager: a quarter- cent per image, according to a 2012 report—or between 1 and 4 dollars per hour, according to Chen’s report.

Kristy Milland, an outspoken activist in the Mechanical Turk community, noted,

“People say to me ‘Oh my god, you work at home? You’re so lucky. . . . You can’t tell them ‘I was tagging images today—it was all ISIS screen grabs. There was a basket full of heads.’ That’s

what I saw just a few months ago. The guy on fire, I had to tag that video. It was like 10 cents a photo."

The work of this globally-disperse force is distanced, literally and figuratively, from the rest of the company's workers. For many, this is full-time work, even if cobbled together from piecemeal assignments.

month—the precariousness of the labor, more than the impact of the content.
Content moderation is currently one of the most common of the "human computation" tasks being handled by this growing but largely concealed workforce. And because this work is both necessary and invisible, the rights and expectations around this kind of contingent information work are just now being established, outside of public view.

When moderation is made as invisible as possible, the content we do see
seems like it is simply there, a natural phenomenon.

The little press coverage available about this hidden workforce has focused primarily on the psychological toll of having to look at the worst the Internet has to offer, all day long, day in and day out. In 2010, the New York Times worried that content moderators, having to look at this "sewer channel" of content reported by users, were akin to "combat veterans, completely desensitized," and noted a growing call for these platforms to provide therapeutic services. Chen's exposé makes this point powerfully: this is not just about deciding whether a nipple is or is not showing, it's about being compelled to look at the most gruesome, the most cruel, the most hateful that platforms users have to offer.

Community Managers
Platforms that are designed to allow smaller subgroups to exist within the broader population of users can empower moderators to be responsible for the activity of a subgroup. These roles differ depending on the shape of the platform in question; I am thinking here of the policing power given to the admins of Facebook Groups, the "redditors" who moderate subreddits at Reddit, the administrators with enhanced editing and policing powers on Wikipedia and Quora.

Community moderators occupy a structurally different position from the moderation teams employed by the platforms, a difference that can be both advantageous and not. As participants in or even founders of the group, sometimes with a lengthy tenure, they may enjoy an established standing in the community. They may be enforcing rules that were generated by and consented to by the community, making enforcement easier. They are often deeply invested in the group's success. On the other hand, they are typically volunteers, usually uncompensated and often underappreciated.

Communities that regularly run afoul of the sitewide guidelines can become a liability for the platform, something Reddit seems to face on a regular basis. In recent years Reddit has had to ban active subreddits that were flourishing under moderators who were unconcerned about, or actively supportive of, the goings on within: subreddits dedicated to sharing stolen nude images of female celebrities, affirming white supremacist ideals and denigrating minorities, or circulating revenge porn. Reddit was slow to act in many of these cases—some charged the company with studiously looking the other way until pressured to respond. But its hesitation may have also been a product of their commitment to deferring as much of the moderation as possible to redditors. Still, when a group grows too extreme or too visible, policies of the platform supersede the will of the community manager.

Flaggers
Enormous platforms face an enormous moderation project, but they also have an enormous resource close at hand: users themselves. Most platforms now invite users to “fl ag” problematic content and behavior, generating a queue of complaints that can be fed to the platform moderators—typically, to its army of crowdworkers first—to adjudicate. Flagging is now widespread across social media platforms, and has settled in as a norm in the logic of the social media interface, alongside “favoriting” and reposting.

Enlisting the crowd to police itself is now commonplace across social media platforms and, more broadly, the management of public information resources. It is increasingly seen as a necessary element of platforms, both by regulators who want platforms to be more responsive and by platform managers hoping to avoid stricter regulations.

Flagging has expanded as part of the vocabulary of online interfaces, beyond alerting a platform to offense: platforms let you fl ag users who you fear are suicidal, or flag news or commentary that peddles falsehoods. What users are being asked to police, and the responsibility attached, is expanding.

Platforms are empty software shells; the creative and social labor of users fills those shells. Somehave argued that the economic value of social media platforms is overwhelmingly built upon this labor that users give away without financial compensation, though there are arguably other forms of compensation users experience: social connection, reputation, and so forth.

Platforms typically describe the flagging mechanism as giving an offended viewer or victim of harassment the means to alert the platform. But this is just one reason to flag, and the easiest to celebrate.

One representative of YouTube told me that the content moderation team often saw surges of flags when a new country got access to the platform. “When we actually launch in other countries, we’ll see the flagging rates just completely spike, because, I think it’s a combination of sort of users in those new countries don’t know the rules of the road and so they’re uploading anything and everything, and the community that is YouTube is actually, they can recognize the outliers pretty quickly and they sort of know what is cool for the community and what isn’t, and so it’s interesting to see sort of the new country absorbed into the community norms and very quickly you see that the flagging rates stabilize.

Superflaggers, Peer Support, and External Efforts
Shifting the work of moderation to users is meant to address the problem of scale, but it gives platforms a fluid and unreliable labor force making judgments that are uneven in quality and sometimes self- interested and deceptive.

Many platform managers would like to have more high- quality flags and fewer accidental and tactical ones; there are ways to ensure this. Some platforms internally assign reputation scores to flaggers, so that subsequent flags from users who have a proven record of flagging accurately will be taken more seriously or acted upon more quickly. Others have introduced programs to encourage and reward users who flag judiciously: Microsoft’s Xbox offers the Enforcement United program; YouTube began a Trusted Flagger program in 2012, then expanded it in 2016 to the YouTube Heroes program, gamifying flagging by offering points and rewards.

Yelp oversees the Yelp Elite Squad in which high- quality reviewers are, by invitation only, made members of a semisecret group, get invited to local parties, and are fêted as tastemakers; Yelp has since hired its community managers from this group

Everyone
Finally, some platforms ask users to rate their own content when it is first posted, and then provide filtering mechanisms so that users can avoid content they want to avoid.

Unlike flagging, this enlists every user, distributing the work more equitably and diminishing the problem that those doing the flagging do not represent the whole community. The challenge in this approach is how to get users to rate their own content fully and consistently.

Platforms don't want to introduce too many steps at the moment a user posts, worried that an unwieldy and multiclick interface will discourage participation. So any user-rating process either must be lean and depend heavily on defaults or it must happen less often.

On Tumblr, for example, each user is asked to rate her entire blog, rather than each post, and the default rating is "safe."

In 2007, Flickr shifted to a user-assigned rating system, where users were expected to rate the photos they post as "safe," "moderate," or "restricted.

Just as with flagging, user ratings are invariably subjective. Ratings may vary because of differences in how the same content is assessed by different users, or because of differences in how users interpret the vocabulary of the ratings system, or both. What I might rate as "moderate," you might very much not. And when I rate my own contributions, I may also be subject to a perceptual bias: my own content is likely to seem acceptable to me.

Conclusion

Each social media platform has cobbled together a content moderation process that draws on the labor of some combination of company employees, temporary crowdworkers, outsourced review teams, legal and expert consultants, community managers, flaggers, admins, mods, superflaggers, nonprofits, activist organizations, and the entire user population.

Not all platforms depend on all of these different labor forces, and no two do so in exactly the same way. Given the scope and variety of this work, it should give us pause when platforms continue to present themselves as open and frictionless flows of user participation.

Moderation arrangements are arrangements of expediency, even exploitation. Social media platforms have successfully cemented the idea that users pay for their services not with dollars but with effort: posting, commenting, liking, reviewing, forwarding. Users have accepted the notion that their micro-contribution, their labor, is the fair price for whatever value and satisfaction they get from the platform. Adding rating and flagging to the user's already long list of jobs is a small ask.

Crowdworkers are a cheap and flexible labor force who can perform the kind of detection and classification that machine-learning techniques can't quite do, can be flexibly moved (or hired and fired) as different kinds of user behavior wax and wane, and can absorb the psychological toll of wading through the mind- numbingly mundane and occasionally traumatic. The scope and character of that labor force remain hidden.

Anyone, at any level of moderation, can make a mistake, be ill-suited to the task, or implement a partial or skewed set of values that does not match those of the users. But this intricate division of labor raises two additional challenges: the coordination of these efforts and the translation of both concerns and criteria between the different layers.

The challenge of content moderation, then, is as much about the coordination of work as it is about making judgments. First, decisions the press or disgruntled users see as mistaken or hypocritical might be the result of slippage between these divisions of labor—between what is allowed and what is flagged, between how a policy is set and how it is conveyed to a fluctuating population of crowdworkers, between how a violation is understood in different cultural climates, between what does get objected to and what should.

A platform might instantiate a reasonable rule with good intentions and thoughtful criteria, but that rule may be interpreted very differently by the internal policy team, the temporary reviewers in another part of the world, the community moderator on the ground, and the offended user clicking the flag. This is a problem for platforms that seek consistency: as one representative of Facebook put it, “The central goal of all enforcement is repeatability at the kinds of scale we’re talking about, because if Facebook or Twitter or whoever can’t do the policy the same way over and over again given similar cases they don’t have a policy. They have some cool aspirations and then chaos, right?”

Finally, our concerns about what is allowed or prohibited should be paired with concerns about how platforms organize, distribute, and oversee new forms of labor—from soliciting the labor of users as a voluntary corps of fl aggers, to employing thousands of crowdworkers under inequitable employment arrangements, to leaving the task of setting and articulating policy to a tiny community of San Francisco tech entrepreneurs who have been socialized and professionalized inside a very specific worldview.

The challenges of coordinating the distributed and decentralized work of moderation are probably most acute at the points of contact. Between flaggers and crowdworkers, and between crowdworkers and policy teams, are membranes across which must flow both expressions of principle in one direction and expressions of concern in the other. At these points of contact, these expressions get rewritten, translated into a new form that is meant to both retain the meaning and fit it to the work that will respond to it. These translations can introduce distortions, ambiguities, and even new meanings.

Many of the problems with these systems of platform moderation lie in the uncertainties of this distributed system of work, and they breed in the shadow of an apparatus that remains distinctly opaque to public scrutiny.

Platforms must stand accountable for constructing this chain of labor, for making cogent decisions about how this work should be done and by whom, for articulating why moderation should be parceled out in this particular way, and for articulating clearly and publicly how they plan to make their moderation effective and responsive while also protecting the labor rights and the emotional and social investments of everyone involved. And while it may be too much to ask, these statements of accountability would be stronger and more persuasive if they came not just from each platform individually, but from all the platforms together as industry-wide commitment.