International Journal of Communication 13(2019), 1526–1543 1932–8036/20190005

# What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation

NICOLAS P. SUZOR[1]
Queensland University of Technology, Australia

SARAH MYERS WEST
AI Now Institute, USA

ANDREW QUODLING
Queensland University of Technology, Australia

JILLIAN YORK
Electronic Frontier Foundation, USA

This article seeks to provide greater specificity to demands for transparency in the commercial content moderation practices of digital platforms. We identify gaps in knowledge through a thematic analysis of 380 survey responses from individuals who have been the subject of content moderation decisions. We argue that meaningful transparency should be understood as a component of a communicative process of accountability (rendering account) to independent stakeholders. We make specific recommendations for platforms to provide people with clear information about decisions that affect them, including what content is moderated, which rule was breached, and a description of the people and automated processes responsible for identifying content and making the decision. Beyond providing more information to individuals about particular decisions, however, we note the major challenge of improving understanding of content moderation at a systems level. General demands for greater transparency should be reframed to focus on enhanced access to large-scale disaggregated data that can enable new methods and collaborations among academia, civil society, and journalists to make these systems more understandable and accountable.

*Keywords: content moderation, platforms, transparency, due process*

Nicolas P. Suzor: n.suzor@qut.edu.au
Sarah Myers West: smw10@nyu.edu
Andrew Quodling: andrew.quodling@gmail.com
Jillian York: jillian@eff.org
Date submitted: 2018–05–24

[1] Nicholas P. Suzor is the recipient of an Australian Research Council DECRA Fellowship for Project DE160101542. This project was funded by the Internet Policy Observatory and the Australian Research Council. Many thanks to Rosalie Gillett and Georgia Robertson for excellent research assistance on this project.

Internet platforms play a major role in governing the material that users can share and view online (UNESCO, 2014). These providers are subject to pressure from law enforcement agencies and private actors around the world to moderate content in different—and sometimes conflicting—ways. Particularly for social media platforms, the decisions that providers make are increasingly the subject of public interest, and a number of major controversies have erupted over the last decade as different stakeholders try to understand and influence how content is moderated online (Gillespie, 2018).

Unfortunately, there are few good data available about how platforms make content moderation decisions (Suzor, Van Geelen, & West, 2018). This leads to confusion among users (West, 2018) and makes it more difficult to have an informed public debate about how to regulate Internet content in a way that protects freedom of expression and other legitimate interests. In this context, there have been frequent calls for platforms to disclose more information about how content moderation systems operate (Bankston & Woolery, 2018; Leetaru, 2018). The concept of transparency, however, is often used in a general sense; Gillespie (2018) goes so far as to say that "calls for greater transparency in the critique of social media are so common as to be nearly vacant" (p. 212).

In this article, we directly address the conceptual problem of vagueness in calls for transparency in the moderation of social media content. We argue that there is a pressing need for more specificity in identifying what information should be provided and to whom. We begin by considering the expressed needs of users, as the first and most immediate audience of information about the decisions that platforms make. Our analysis draws on 380 survey responses submitted by users who had been adversely affected by the removal of content they posted on social media platforms or by the suspension of their account. Through this empirical work, we propose a more targeted concept of transparency that focuses on the information that is required to better understand systems of content moderation in a way that renders them more accountable.

We found that users frequently express confusion about what action has triggered a moderation action or account suspension. In our data, more than a quarter of respondents (108 reports) were uncertain about what content triggered a moderation decision. Even when the offending content was clearly identifiable, users too frequently had insufficient information to understand why a moderation decision was made. Only half of our participants (190 users) expressed confidence in their understanding of the platform's moderation action. Through close analysis of the complaints and apparent confusion evident in these reports, we identify several specific deficiencies in the way that platforms communicate with users about moderation decisions. We use this analysis to develop a set of more detailed recommendations for platforms to increase the transparency of their content moderation systems.

We argue in favor of a more nuanced conception of platform transparency not only as information provision, but also as a necessary component within a system of accountability. Transparency, in this sense, is part of the communicative process of rendering account to stakeholders and independent institutions. Focusing on accountability, we suggest, helps to understand why simply providing more detailed information about moderation will not be sufficient to help users understand moderation systems or to hold platforms to account for their decisions. There is a danger, we argue, in treating the disclosure of information as a goal in itself. We provide a critique of transparency practices that can be used by platforms to obscure the inner workings of content moderation systems and defend against calls for greater accountability. In particular, the provision of aggregate statistics of content moderation in regular "transparency reports" is

insufficient to help identify how well moderation systems are working or how they can be improved. We consider what sorts of information may help a diverse range of scholars and civil society organizations make meaningful analyses of the operation of content moderation systems in ways that might help to build real accountability. This, we argue, requires the development of new collaborations and institutions to monitor systematic trends and inform both policy debates and public understanding.

## The Limits of Aggregate Transparency: Understanding Complex Systems and Holding Platforms to Account

Few platforms provide aggregate transparency reporting information about how they enforce their terms of service or community guidelines, particularly where third parties are not involved (Suzor et al., 2018). The Santa Clara Principles (2018) are the latest in a series of demands for platforms to make more information available about how they moderate content. The principles are a joint declaration from civil society groups and academics (including ourselves) that outlines the types of aggregate statistics that present a useful starting point for analysis of content moderation at a systems level, based in part on the preliminary findings of our research below. They urge platforms to improve transparency at two levels of abstraction: individual notice about specific decisions and regularly aggregated information.

To help individuals understand decisions that affect them, the Santa Clara Principles require notice to users about what specific content was at issue, which specific rule the content was alleged or found to violate, how the content was detected, and by whom. The provision of better information to users by platforms is necessary to understand moderation systems and hold platforms to account for the enforcement of their rules. In the history of common law legal systems, the provision of public reasons to explain decisions has long been seen as fundamental to avoiding arbitrariness and promoting good decision making (Bosland & Gill, 2014) and the rule of law (Waldron, 2008). The analogy from legal decision making is "that justice should not only be done, but should manifestly and undoubtedly be seen to be done" (*R v. Sussex Justices, Ex parte McCarthy*, 1924). When full reasons are not provided to explain decisions, meaningful due process is impossible.

But understanding individual decisions is insufficient to understand the massive systems of content moderation. Given the importance of commercial content moderation in determining how and what billions of users can communicate online, there have been increasing demands for Internet platforms to release details of their moderation processes at a systems level (Bankston, Schulman, & Woolery, n.d.; Newland, Nolan, Wong, & York, 2011). At a higher level of abstraction, the Santa Clara Principles urge platforms to release regular information about the total numbers of posts and accounts flagged or reported and the proportion of content removed or accounts suspended. Demands of this type are positioned around the value of understanding content moderation not only to the individual user, but also to the broader community. This value might be understood as a public's right to hear that democratic self-governance requires that individuals not only have the right to speak, but that publics have the right to encounter a rich diversity of viewpoints and perspectives (Ananny, 2018).

Platforms are slowly improving their transparency practices in response to mounting pressure from a wide range of stakeholders. In April 2018, YouTube released its first transparency report that included detailed numbers on its terms of service takedowns. Days later, Facebook provided a more detailed explanation of its content moderation processes including a slightly redacted version of the internal

guidelines it provides to moderators. Other companies have made promises that additional information will be forthcoming. The continuous improvements in transparency reporting over the last decade (Ranking Digital Rights, 2018) are welcome changes, but there is a danger in the language of transparency.

In calls for greater transparency, there is often an explicit or implicit assumption that transparency—greater information disclosure—leads to greater accountability and trust (Albu & Flyverbom, 2016). Good governance is often thought to be contingent on transparency (Braithwaite & Drahos, 2000). But the extent to which transparency actually leads to greater accountability and better outcomes, however, is often unclear at best. Transparency, deployed strategically by an organization as theater to ward off claims for greater accountability, can ultimately work to obscure understanding (Losey, 2015; Parsons, 2019). From this perspective, the production of transparency reports could provide platforms a "market friendly" response to demands for greater accountability while avoiding regulation that imposes real accountability (and is backed by authoritative sanctions; Fox, 2010).

For transparency to be meaningful, it has to be targeted—not just increasing information, but communicating in a way that can be used to help hold decision makers to account (Fung, Graham, & Weil, 2007). This approach to transparency makes clearer the need to pay attention to the capacity of different audiences to interpret the information disclosed and to the politics of disclosure that can influence the quality and scope of information (Albu & Flyverbom, 2016).

Aggregate statistics can provide an overview of content moderation processes and identify particular broad areas of concern, but they are not sufficient to enable the detailed analysis that is required to hold platforms accountable. Notices about individual decisions, by contrast, are not often provided in a form that can be used by researchers and other independent stakeholders to understand moderation at a systems level. Understanding the complex moderation systems of digital platforms in detail and at scale is a critical challenge. As platforms begin to disclose aggregated data on terms of service enforcement, it is therefore important to consider what information is required for transparency to be meaningful and how it can be used. We approach this task from the provocation of Ananny and Crawford (2018), who urge those of us thinking about transparency to ask, "What is being looked at, what good comes from seeing it, and what are we *not* able to see?" (p. 985). Transparency, Ananny and Crawford point out, is not sufficient to generate accountability, but it can be used as a starting point "for reconstructing accountability for systems that cannot be seen into, held still, or fully traced" (p. 985).

We argue that a partial answer to this question may be derived by learning from the experiences of users who have already been affected by content moderation, who constitute the most immediate audience of information about the decisions that platforms make. These experiences allow us to provide a much more specific grounding for the types of information that are required to improve understanding of moderation decisions. We then can use these insights to consider the form in which information is disclosed and what types of information may be useful to help scholars and civil society organizations to understand—and hold accountable—the operation of content moderation systems.

## Method

We thus undertook a thematic analysis of 380 survey responses submitted by users who have been adversely affected by the removal of content they shared on social media platforms or by the suspension of

their account. These survey responses were submitted by users between November 2015 and August 2017 to the website Onlinecensorship.org, an advocacy project launched in 2015 by members of the Electronic Frontier Foundation and Visualizing Impact that seeks to shed greater light on content moderation practices by collecting reports from users when their content or accounts are removed from social media sites.[2]

Many of these reports contained detailed textual personal summaries of the moderation events that users had been affected by on Internet platforms. Most of the reports dealt with moderation on Facebook, Instagram, and Twitter, although some other social media platforms were also represented. These reports were self-selected: The users who took the time to submit a report were only a tiny proportion of all users who have been affected by content moderation decisions, and they were perhaps likely to feel more seriously aggrieved than the average user. Nevertheless, these reports are a useful source to understand the types of issues that are of greatest significance to users in trying to understand or contest content moderation decisions.

We read through each of these reports to understand the themes that emerged as the most significant concern to users. Our focus in this study was on the information that was available to users; we looked specifically for concerns that emerged from lack of good information to explain the moderation process in general or the particular decision at hand. Importantly, the survey reports primarily represented users who felt that they had been unfairly censored by platforms. There are other important stakeholders in disputes over content moderation. The users who complain about content they object to, either through flagging or other channels, also express concerns about the lack of transparency in how their complaints are evaluated and determined. At a higher level of abstraction, civil society groups and government regulators also have complained about the lack of information that major platforms make available about how they deal with different types problematic content. Neither of these sets of concerns was visible in the survey data, but they are documented more thoroughly in the literature. We consider these issues in more detail in the final section of this article, when we consider the needs that different audiences have for information about content moderation decisions.

## What Is Meaningful Transparency?

In the sections that follow, we organize the concerns users raise into four thematic categories, and offer suggestions for improvements in transparency. First, we note major concerns about the prevalence of confusion from users about the exact content or behavior that triggered a sanction from the platform. Second, there is a systemic failure on the part of platforms to provide good reasons to explain the decisions they reach. Third, because the content moderators who are ultimately responsible for making decisions are hidden from view, users may more readily infer bias or attribute a lack of contextual knowledge and cultural sensitivity to human moderation teams of major platforms. Fourth, we note that users are not often told how their content was flagged for review, and are sometimes left to guess whether they have been the subject of a complaint by an acquaintance, a government agency, or an opaque algorithm designed to

[2] Two members of the team that developed Onlinecensorship.org are coauthors of this article. Data were collected and shared with the research team in anonymized form under the privacy policy terms outlined at https://onlinecensorship.org/privacy-policy.

identify potentially problematic content. Taken together, the lack of clear information available to users breeds folk theories and concerns about conspiracies and systemic bias in content moderation processes.

### *Confusion About What Content Triggered Moderation*

In the Onlinecensorship.org data, we saw users frequently express confusion about what post or act triggered a moderation action or account suspension. For example, one Australian owner of a Facebook page that includes contributions from many people said the suspension notification they received was not specific enough to identify which post might have triggered the complete suspension of the page. The owner of the page was only told that one of the posts contained "malicious or misleading content," which led them to speculate about why their page was banned: "I can only guess that it was in relation to either a post about the rise of authoritarianism in U.S. politics, or a post about China's plans to build a global power grid. In any case it seems to be politically motivated censorship" (#F83).[3] The page's owner explains that when they managed to contact someone from Facebook support and offered to remove the offending post, the response "didn't provide any detail at all, it simply said, 'We have a no-tolerance policy concerning this infraction and your page is ineligible to be republished.' I still don't know which post triggered the censorship" (#F83).

The lack of information about what content triggered a breach appears to be a serious issue that could be avoided relatively easily. There appears to be no major technical impediment to providing users with a notice containing a link or extract of the content, given that presumably it is always a particular post—or collection of posts—that reviewers are evaluating for compliance with the rules of the platform. Some platforms may have to develop new workflows to ensure that this information flows through the moderation process to the notification message provided to users, but this is presumably not difficult to implement.

Sometimes, users are not provided with any notice at all when their content is removed. In some cases, this is required by law (Losey, 2015; Woolery, Budish, & Bankston, 2016). For example, a judicial order for content removal may include a suppression order that prevents platforms from notifying the user. In other cases, platforms introduce so-called "shadowbans" voluntarily, when they may hide or deprioritize content without informing the user. The reasons behind shadowbans can sometimes be relatively benign; when dealing with spammers who create bots to post content, for example, some platforms decide to use shadowbans to make it more difficult for the spammers to know when they have been detected and create a new account (Massanari, 2017). By their very nature, however, shadowbans create uncertainty, and it is not surprising to see users suspect that they may have been surreptitiously censored.

Seventeen of the participants in the Onlinecensorship.org survey reported that they suspected that they had been shadowbanned or that the platform's actions in moderating their content exceeded what had been communicated to them. A German Instagram user who posts risqué fanart paintings of computer game characters that are routinely removed noted that the images do not appear to show up in the search results of hashtags: "I'm also something like Shadowbanned on Instagram, I'm not the first one with that. Even when I tag a image with a clearly not blacklisted hashtag, I'm not visible in the public search" (#I7).

[3] The quotes drawn from the OnlineCensorship.org survey are identified in this article by a # symbol followed by the unique identifier used in the original data set.

Some users reported going to some lengths to identify potential shadowbanning. One Twitter user, who advertises their organic farm shop, believed they had been shadowbanned for posting politically sensitive content calling out brands that use GMO ingredients and negative content about Hillary Clinton in the lead-up to the 2016 U.S. presidential election. They used the Tor browser to view their tweet stream from a different location, and noted in side-by-side results that certain tweets in their history were not visible in their normal browser. We had no way to verify these claims, but the fact that users are investigating potential hidden moderation suggests a level of anxiety about content moderation processes. If these claims are correct, the implication is most likely that Twitter has withheld the tweets in question in certain jurisdictions, and the Tor browser was able to bypass the geoblocking by appearing to access the tweet from outside the user's country (United States). The fact that this appeared to be happening without notice created a great deal of distrust from the user, who was left to assume that Twitter censors tweets either for political reasons or in support of large commercial interests.

The lack of information makes it very difficult for users to understand the rules and learn from the experience, or to understand whether their content is moderated by the platform or removed by another user. For example, a Facebook user reported,

> Every post I've ever made on to any Facebook group has been removed with no explanation, yet my account is still active. This is worse than just having one post blocked because at least then you know what they didn't like. I have asked them many times why they have blocked me but they won't answer. I have been completely silenced, and by not revealing why, they avoid showing their exact bias.

The user explains that they want to abide by the rules, but "if Facebook doesn't ever explain why they censored me, then how am I ever able to stop it from happening again?" (#F148).

As a first step toward meaningful transparency in content moderation, we suggest that platforms ensure that users are informed when their content is removed or made invisible to particular audiences. Good moderation practices should ensure that the URL of the prohibited content or a sufficiently detailed extract is available in the notification sent to the user. These notifications should be permanently available to the user in some form, and not just sent as potentially ephemeral in-app notifications. If shadowbanning is used to target certain users, platforms should exercise great care. Our results show how users are likely to distrust the system if they suspect that they are being censored without notification, and the mere existence of shadowbanning as a practice may well lead users to draw inferences of direct, targeted intervention even to explain technical problems or changes to the platform's architecture.

### *Lack of Clear Reasons for Moderation Decisions*

Users too frequently have insufficient information to understand why their content was moderated or account suspended. In our data set, only half of the users (190 users) expressed confidence in their understanding of the platform's moderation action. In response to the limited information available and their distrust of the explanations given by platforms, we saw users develop and use vernacular explanations about why their content was removed. The lack of reliable information leads users to develop their own rationalizations to explain moderation decisions they are subject to, frequently blaming biased moderators and undue external influence (West, 2018).

Some users explained that they had no real understanding of why their content was removed or their account suspended. A Facebook user who was suspended for 30 days on two separate instances explained that she was unable to understand the reason for the suspension. Her most recent post was, in her eyes, a reasonable complaint about the way she was treated in a misogynistic way by a car dealership salesman who had refused to talk to her and only spoke to her sons: "It had no threats, no harassment, and no vulgarity." The cursory level of information provided by Facebook in both instances made it difficult for her to understand why her posts were prohibited: "It honestly doesn't make sense to me why I was banned in either instance. Facebook provided no clarification in the matter. Only that I should read the community standards, which I did, and found nothing which indicated I had violated the rules" (#528). Another user believed that Twitter was deliberately censoring conservative political opinions: "Twitter appears to just suspend conservative writers in the hopes that they (Twitter) can limit free speech and conservative voices on Twitter" (#484).

More than half of our participants (214 users) explained the moderation actions of platforms using what appeared to be self-rationalized statements about the motives of platform operators. They were offered in terms of what appeared to be folkloric vernacular discussion of how the users believe that the platform has operated. This is a common and understandable response that users have in the face of limited available information (Burgess & Green, 2009); we suspect that users offer their own understandings and rationales for their bans in response to the lack of convincing explanations from platform operators.

This conspiratorial knowledge-creation behavior seems to manifest most often when users attempt to explain why an Internet platform seems to be overtly policing content in relation to domestic politics or broader sociopolitical, geopolitical, and commercial issues. For example, we saw many responses that made no reference to the rules of the platforms, but instead alleged that their content was removed or account suspended because it "wasn't politically correct" (#537) or that a particular business interest did not like their views. Approximately a quarter of participants (96 users) expressed suspicion that platforms were operating in a manner that was politically motivated. In our data, users expressed arguments that platform operators were biased against a wide range of participants, including conservatives, Donald Trump supporters, "alt-right" figures, Gamergaters, Bernie Sanders supporters, antivaccination campaigners, vegans, and more.

The vernacular rationalizations that users develop breed further distrust and can encourage the development of conspiracy theories. One user in our data, for example, reported that although they had been censored by a platform, they felt that the act of moderation had validated their position. Following a suspension as a result of a discussion about the "Pizzagate" conspiracy theory (a widely debunked hoax alleging that senior officials in the Democratic Party were involved in a child sex-trafficking ring operated in the basement of a pizza restaurant; Robb, 2017), the user noted, "I'm actually really glad that they suspended me because it validates the work I'm doing. The cover-up proves conspiracy. So I wear my Twitter suspensions like badges of honor" (#619).

The lack of clear justifications makes it difficult for users to learn the rules. In some cases, it may be sufficient for platforms to provide a clear indication of which rule was deemed to have been broken by the user. In other cases, however, it seems that users could benefit from fuller explanations of the rules. So, for example, one user whose Facebook posts were blocked explained,

> Not quite sure [why I was banned] to be honest, all I said was "bill nye is a fag" and it got removed and I am banned for 3 days. Last time I got banned it was for saying "rape is a natural instinct." . . . I do not feel I have done anything wrong. All I have done is expressed my opinions. I see pages that make fun of autism that don't ever get banned. (#575)

These distasteful comments establish a moral equivalency among offending content: How does the platform justify the removal of comments advocating rape, but not those that express hate toward people with autism? After all, the user explains, these are all just "opinions." As this comment suggests, platforms have a substantive challenge in educating users about the guidelines that are used to explain what kinds of content are allowed and disallowed on the platforms. Many social media users may not know these guidelines exist, until they violate them. Approaching content moderation from a position of user education, as opposed to conceiving of it as a system to punish violating users, may go farther in addressing these challenges (West, 2018).

### *Perceived Bias in Policies and Moderation Teams*

Major commercial platforms deal with content moderation at a massive scale, and major social platforms employ or outsource large teams of human moderators to make decisions. High-quality, consistent moderation is a skilled task: Trained moderators need to quickly judge whether content is prohibited or not according to a large set of rules and exceptions. In recent years, the lack of transparency in the process of commercial content moderation has created concerns about the process. Moderation is often outsourced to workers in developing countries and freelancers on task-working platforms, and moderators sometimes report not having sufficient time, training, or support to make considered decisions (Chen, 2012; Roberts, 2016).

In our data, some participants expressed a great deal of confusion about the people and processes responsible for making content moderation decisions. This leads directly to allegations of explicit bias and an implicit lack of understanding of complex cultural issues. So, for example, the perception that social media companies are run by young people led a conservative poster to complain about those who "run afoul" of the "social justice warrior" orthodoxy being unfairly targeted: "Facebook has hired many millennials who've been educated in systems that emphasize identity politics and thus, personal offense is indistinguishable to them from actual offense" (#F39). At the same time, people on the other side of the political spectrum reported having the exact opposite perception and experience:

> I suspect that Facebook has one or more male mods who are hostile to feminists and minorities in general. I have had posted [*sic*] removed and been suspended five times for relatively innocuous posts such as talking about how women attempt suicide three times as much as men, and how men are expected to limit their emotional range. Meanwhile hundreds of vile posts I have reported have stayed up, from individuals referring to me directly as a cunt, to men advocating for women to be rounded up into "rape camps," raped and repeatedly impregnated until they die. (#F23)

Another user thought that Facebook had an "issue with anti trump folks and male sexuality" (#518). He took issue with a perceived double standard when images of women whose genitalia can be seen in outline through their clothing seem common, whereas his account was suspended for posting images of clothed males where the shape of their penis is visible to a closed group "of like-minded men." This is a

recurrent theme; the 2012 leak of Facebook's training documents for oDesk moderators showed that the company prohibited "blatant (obvious) depiction of camel toes and moose knuckles," but Facebook has repeatedly been criticized for applying its standards in ways that reflect homophobic biases (Chen, 2012). YouTube too was heavily criticized in 2017 for hiding LGBTIQ educational videos for users who enabled its "restricted mode" settings and for excluding videos from receiving advertising revenues (Lang, 2017). The vague language that is common in terms of service and community guidelines, combined with the lack of good information about how platforms moderate content in practice, can work to hide actual systemic bias and create serious apprehension about its existence (Noble, 2018; Roberts, 2016).

Moderation practices vary greatly across platforms and over time as platforms grow and respond to criticism. Major platforms apparently do a lot of work internally to ensure consistency across moderation decisions at a large scale, but this information is not readily visible to users who are subject to particular decisions. It would be useful for platforms to provide more general demographic information about the makeup of their moderation teams, with particular regard to age, nationality, race, and gender. This type of information is more useful at an aggregate level; it is not necessary to provide specific details about the actual moderator of any given decision, but providing general information might go some way to alleviate the concerns we see users express about potential bias.

In order to enable users to better trust the outcomes of moderation decisions, we think it is necessary for platforms to disclose detailed information about the training and guidelines associated with the moderation process, including what processes exist to support moderators to make consistent and well-informed decisions in the context of potential ambiguity. One of the core challenges of content moderation at scale is the trade-off between precision and nuance. The rules that platforms present to users are generally expressed in quite vague terms. Because rules are encoded in language, they are "open textured"; they may have a clear core meaning, but there will always be areas of uncertainty—what Hart (1961/1994) called the "penumbra of doubt" (p. 123).

Instagram's Terms of Use, for example, generally prohibit 'nudity,' but the description of this category of content remains unclear (Witt, Suzor & Huggins, forthcoming). This lack of clarity leads to controversies over how the rules are applied in practice, and allegations of bias are common in these debates. Instagram's nudity policy is a particularly vague example, but there will always be some degree of uncertainty in the rules of a platform. To reduce this uncertainty and increase consistency across a large team of moderators, platforms typically create detailed training materials with examples about content that is prohibited and permitted for each rule. Effectively, moderators are enforcing a separate set of rules to the ones that are displayed to users. When these materials leak, they provide a rare glimpse into how content moderation decisions are actually made in practice: The popularity of exposés from such publications as *Gawker*, *ProPublica*, and *The Guardian* (Angwin & Grassegger, 2017; Chen, 2012; Hopkins, 2017) reveals a real hunger from users for these insights into an otherwise opaque system.

One of the particularly difficult areas of content moderation is the difficulty that transnational platforms have in dealing with locally specific contexts. Standards about the acceptability of speech differ among regions; usually, these standards have evolved in the context of mass media regulation over time, and reflect significant variation in regulatory instruments, legal opinions, and community understandings of acceptable content (Gillespie, 2018). Commentators have for several years noted the lack of information about the policy teams responsible for setting the rules that moderators apply, which can lead to allegations of implicit

bias entrenched in the rules themselves (Rosen, 2013). In addition, the interpretation of particular content may differ significantly based on local and context-specific knowledge; for example, the use of racist imagery may seem more innocuous to moderators who do not have experience with the cultural dynamics of racism in specific national contexts (Matamoros-Fernández, 2017). The lack of local knowledge can lead to major controversies, as when the UN Independent International Fact-Finding Mission on Myanmar found that Facebook had played a "determining role" in spreading racial vilification in Myanmar (Miles, 2018). Representatives from Facebook noted that the lack of local knowledge in its U.S.-based moderation and policy teams led to a decision to ban the accounts of an insurgent organization resisting state-led ethnic cleansing of the Rohingya Muslim minority in the country, as well to prohibit posts that demonstrated support for the organization (Glaser & Oremus, 2018).

Building trust in moderation processes in circumstances such as the Myanmar case will require more detailed accounting by platforms about the entire process of content moderation systems. A good understanding of the system requires more than just a statement of reasons for particular decisions; if users are to have more faith in the moderation system, they need to be able to trust the process. This suggests that platforms need to invest in finding ways to help users understand how rules are created and by whom; how potential breaches are identified; how moderators learn and enforce the rules; what the composition, training, and working conditions of moderation teams are like; how the platform ensures consistency; how consistent decisions are in practice; and how mistakes and novel issues are dealt with.

### *Biased Algorithms and Undue Influence*

Modern content moderation systems are highly complex and involve many different actors and automated agents. The potential for bias across the system exists not just in the rules and decisions of the platform, but also in how content is identified for review. The identification of potential breaches of the rules relies heavily on user labor in flagging objectionable material for review, but flagging systems hide complex contestations of values among users (Crawford & Gillespie, 2014). Major platforms are now investing heavily in the development of automated tools to detect and flag content for review, but there is little information about how automated detection systems work, what data they are trained on, how effective they are, and how they integrate with the other components within a moderation system. Users in our data set expressed distrust of other users, worrying that they were being targeted in coordinated flagging campaigns, resulting in effective bias in moderation decisions. This distrust is worsened by a perception that flagging systems can be gamed, particularly when decisions are made with the assistance of algorithms that automate part of the process. In our data set, 50 users reported that they believed that their content was moderated automatically or with the use of an algorithm. These users often believed that these algorithms were being exploited by coordinated groups of malicious users. For example, a Syrian activist believed that their Facebook "posts have been reported repeatedly by pro-Assad/Putin users," and that "these mass reporting campaigns trigger automatic algorithms that result in post deletion and account suspension." After repeated suspensions, a Twitter user who posts conservative content was left to wonder whether "a group of Twitter abusers [were] 'reporting' me or something, for exercising my free speech like everyone else on Twitter?"

Other users in our data expressed concern that the moderation system was being influenced by powerful external interests, including law enforcement officials or other government actors. Some platforms have granted additional powers to particular trusted groups to allow them to more easily flag a larger volume of posts or access expedited moderation processes (Crawford & Gillespie, 2014). Many platforms are

responsive to requests from law enforcement agencies, but do not explain that content was identified by law enforcement when moderating content under their terms of service. This leads to confusion about decisions and fears about potential overreach from public agencies in ways that would raise due process or constitutional issues if the action were taken directly by the public agency (Birnhack & Elkin-Koren, 2003). For example, a Turkish Facebook user who posts news and political content expressed fear that their Facebook account was suspended because of an organized campaign by the Turkish government. When asked why their account was suspended, the user responded that President Erdogan "has got an army of paid cyber agents who are busily making complaints about posts that criticize the ruling party" (#F76). The user continued, explicitly drawing Facebook as complicit in political censorship,

> Facebook is now reflecting the authoritarian regime in Turkey by just deleting or taking sides with tyrants who dictate agendas upon people that people are not ok with. Facebook is not a democracy there is no internal [dispute] resolution. They just ban it with general explanations while the same content is all over facebook being freely [distributed].

The user believes that their high-profile pages and posts, which can reach millions of views, make them a target for organized complaints, which culminated in three successive bans for 30 days each at the time of the report.

Good moderation practices should try to provide more context to users about how their content is identified for review. The lack of transparency across the entire content moderation process creates confusion about how breaches of the rules are detected and enforced. This confusion in turn leads to distrust about the role of algorithms, other users, law enforcement agencies, other third parties, and internal decision makers in flagging, identifying, or evaluating prohibited content. Users may in some cases have legitimate concerns that they are being targeted by organized attempts to stifle their speech by other users, governments, or corporate interests. There are, however, important privacy concerns that platforms must take into consideration here. The comparable legal principle of due process requires that people subject to sanctions are entitled to know their accuser, but this could expose people to serious risk of harm (Feerst, 2018). Recently, for example, CloudFlare revealed the identity of people who were reporting hate speech hosted by neo-Nazi website The Daily Stormer in the notifications they sent to their clients, which were then used to fuel campaigns of harassment and abuse against those who criticized the content (Schwencke, 2017). Some level of increased disclosure is likely necessary to inform users about the source of complaints, but it is not yet clear what best practices in this regard might look like. The risk to the complainant is much lower in the case of government complaints (whether legal or extrajudicial) or material detected by the platform itself (algorithmically or manually); moderation decisions that result from these sources should presumably clearly disclose the particular reporting process.

## Analysis: From Individual Understanding to Systemic Accountability

For individuals who are directly subject to decisions about the enforcement of rules of social spaces, transparency is a critical prerequisite to understanding why a decision has been reached. But it is not enough: The accounts we analyzed point to a greater need not only for transparency, but also for due process. By focusing on providing aggregate statistics at scale, platforms present a false dichotomy between efficiency and due process that can work to obscure more fruitful opportunities to improve content moderation systems. Fostering meaningful transparency means, at least in part, providing more detailed and individualized explanations of the content moderation process. This might be addressed with relatively

little impact on efficiency, by providing users with notice that includes the specific URL of the content that triggers a decision, for example. These types of improvements are relatively simple changes that could go a long way to improving the experience for users.

The much more difficult sets of concerns are about bias and undue influence. Epistemologically, the provision of highly aggregated statistics about the numbers of complaints and moderation decisions made cannot be used to draw conclusions about the quality of decisions, and the results of any given decision are hard to generalize to understand complex, interrelated systems at scale (Ananny & Crawford, 2018). If users (and regulators) are to develop greater confidence in the operations of content moderation systems, much more granular information will be required. Understanding bias in moderation decisions and the algorithms that support them requires careful attention to the inputs and outputs of these systems and their differential social impact (Noble, 2018; Sandvig, Hamilton, Karahalios, & Langbort, 2014). Analysis of this type will require large-scale access to data on individual moderation decisions as well as deep qualitative analyses of the automated and human processes that platforms deploy internally.

This is the type of transparency that platforms have been reluctant to provide. We might speculate that platforms resist this type of meaningful transparency precisely because it might eventually lead to greater accountability, and therefore greater restrictions on their actions. Whatever the reason, platforms have historically not provided sufficiently granular data that could enable this type of analysis. Scholars have complained for years that the application program interfaces and terms of service of major platforms are designed primarily for commercial exploitation and are not suitable for research, effectively making a great deal of important research impractical or impossible for scholars (Bechmann & Vahlstrup, 2015; Burgess & Bruns, 2015). Recently, platforms have even started to clamp down further on application program interface access for researchers, even as questions of undue influence and bias in digital media environments become more pressing. Facebook's revamped security settings in the wake of the Cambridge Analytica scandal restrict application program interface access in a way that further limits access to researchers, and continues a trend to provide limited access to small groups of scholars to undertake research on selected defined questions (Bruns, 2018).

As calls for greater transparency continue to intensify, we suggest that they should be expanded to focus on how platforms can provide the data that will enable researchers to understand a wide variety of concerns about moderation systems. This is the difference between transparency that merely provides aggregate information and transparency that can help to foster accountability. Improving accountability in these systems will require a degree of openness that can enable a diverse range of stakeholders to investigate particular issues as they arise in a way that cannot be easily specified in advance. Specifically, this means at least that platforms should work to find ways to provide access to fine-grained data on moderation decisions and the operation of different components of moderation systems in a way that enables independent public interest research that "can diagnose emergent problems and suggest possible remedies" (Bruns, 2018, para. 11). It will require new methods to investigate bias and efficacy of enforcement in a way that allows comparison among platforms and over time as rules, architectures, processes, and social norms change (Sandvig et al., 2014; Suzor et al., 2018). This research will require the cooperation of research institutions and granting agencies that can provide resources to support them, as well as platforms to provide access to more granular data on moderation processes and outcomes. It will also need new and ongoing collaborations with journalists (Diakopoulos, 2015) and civil society organizations (MacKinnon, Maréchal, & Kumar, 2016) that are able to make content moderation systems

understandable to wide audiences in a way that can be used to hold platforms to account against a set of shared public values (Parsons, 2019). Given the complexity of moderation systems and the contested values at stake, this work is likely to be difficult, and will require a large and diverse set of collaborations to help monitor and communicate concerns to platforms, users, and regulators in a way that can improve understanding and progress the political debates about accountability.

## Conclusion

Our analysis provides additional context to ground calls for greater transparency in the specific information needs expressed by a group of users directly affected by moderation decisions. In particular, we suggest that individual users whose posts have been removed or accounts suspended need specific explanation about which content breached the rules, how that content was identified, who was responsible for making the decision, and why the conclusion was reached that a rule had been breached. We argue that some of these concerns can be alleviated, at least in part, by better disclosure and education about the rules of digital platforms and the enforcement decisions they reach. We suggest that platforms should pay particular attention to these expressed areas of uncertainty, because they are breeding mistrust among users, who in turn are developing their own vernacular explanations of content moderation processes in the absence of clear and reliable information.

We ultimately support calls for greater transparency around content moderation policies and decisions, but note that the ongoing provision of explanations for individual decisions and abstract aggregated statistics will not be sufficient to help understand moderation systemically. Meaningful transparency also requires a better understanding of the deeper complexities of content moderation processes, and should be conceptualized as a communicative process of rendering account to the many stakeholders implicated in them.

We suggest future work should focus on methods for understanding content moderation at a systems level. This may require building new kinds of institutions that can support diverse, global, and distributed independent research that is grounded in localized contexts and can provide the kind of nuance necessary to make sense of how content moderation impacts users around the world. None of this will be easy, but given the important role that digital platforms play in mediating public communication, it is vital that ongoing calls for transparency be framed in a way that can actually help to improve understanding. Without this extra work, the major threat is that transparency will be deployed in a way that obscures the responsibilities of platforms and helps resist demands for systemic change.

**References**

Albu, O. B., & Flyverbom, M. (2016). Organizational transparency: Conceptualizations, conditions, and consequences. *Business & Society*, *58*(2), 268–297. doi:10.1177/0007650316659851

Ananny, M. (2018). *Networked press freedom: Creating infrastructures for a public's right to hear.* Cambridge, MA: MIT Press.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. doi:10.1177/1461444816676645

Angwin, J., & Grassegger, H. (2017, June 28). Facebook's secret censorship rules protect White men from hate speech but not Black children. *ProPublica*. Retrieved from https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms

Bankston, K., Schulman, R., & Woolery, L. (n.d.). Case study #3: Transparency reporting. *New America.* Retrieved from https://www.newamerica.org/in-depth/getting-Internet-companies-do-right-thing/case-study-3-transparency-reporting/

Bankston, K., & Woolery, L. (2018). We need to shine a light on private censorship. *Techdirt.* Retrieved from https://www.techdirt.com/articles/20180130/22212639127/we-need-to-shine-light-private-online-censorship.shtml

Bechmann, A., & Vahlstrup, P. B. (2015). Studying Facebook and Instagram data: The digital footprints software. *First Monday*, *20*(12). doi:10.5210/fm.v20i12.5968

Birnhack, M. D., & Elkin-Koren, N. (2003). The invisible handshake: The reemergence of the state in the digital environment. *Virginia Journal of Law and Technology*, *8*, 6–13. doi:10.2139/ssrn.381020

Bosland, J., & Gill, J. (2014). The principle of open justice and the judicial duty to give public reasons. *Melbourne University Law Review*, *38*(2), 482–524. Retrieved from https://search.informit.com.au/fullText;dn=973862058950027;res=IELHSS

Braithwaite, J., & Drahos, P. (2000). *Global business regulation*. Cambridge, UK: Cambridge University Press.

Bruns, A. (2018, April 25). Facebook shuts the gate after the horse has bolted, and hurts real research in the process. *Internet Policy Review.* Retrieved from https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786

Burgess, J., & Bruns, A. (2015). Easy data, hard data: The politics and pragmatics of Twitter research after the computational turn. In G. Langlois, J. Redden, & G. Elmer (Eds.), *Compromised fata: From social media to big data* (pp. 93–111). London, UK: Bloomsbury Publishing. Retrieved from http://www.bloomsbury.com/au/compromised-data-9781501306525/

Burgess, J., & Green, J. (2009). *YouTube: Online video and participatory culture* (1st ed.). Cambridge, UK: Polity Press.

Chen, A. (2012, February 16). Inside Facebook's outsourced anti-porn and gore brigade, where "camel toes" are more offensive than "crushed heads." *Gawker*. Retrieved from http://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads

Crawford, K., & Gillespie, T. (2014). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. doi:10.1177/1461444814543163

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, *3*(3), 398–415. doi:10.1080/21670811.2014.976411

Feerst, A. (2018, February 1). Implementing transparency about content moderation. *Techdirt.* Retrieved from https://www.techdirt.com/articles/20180131/22182339132/implementing-transparency-about-content-moderation.shtml

Fox, J. (2010). The uncertain relationship between transparency and accountability. In A. Cornwall & D. Eade (Eds.), *Deconstructing development discourse: Buzzwords and fuzzwords* (pp. 245–256). Warwickshire, UK: Practical Action.

Fung, A., Graham, M., & Weil, D. (2007). *Full disclosure: The perils and promise of transparency*. Cambridge, UK: Cambridge University Press.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media* (1st ed.). New Haven, CT: Yale University Press.

Glaser, A., & Oremus, W. (2018, March 15). *Facebook's alleged role in Myanmar's violence is "deeply concerning," says Facebook's News Feed Chief*. Retrieved from https://slate.com/technology/2018/03/facebooks-alleged-role-in-myanmars-violence-is-deeply-concerning-says-facebooks-news-feed-chief.html

Hart, H. L. A. (1994). *The concept of law* (2nd ed.). New York, NY: Oxford University Press. (Original work published 1961)

Hopkins, N. (2017, May 21). Revealed: Facebook's internal rulebook on sex, terrorism and violence. *The Guardian*. Retrieved from http://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence

Lang, N. (2017, May 16). YouTube has supposedly stopped restricting LGBT content: So why are queer YouTubers still leaving? *New Now Next.* Retrieved from http://www.newnownext.com/youtube-censorship-lgbt-monetize/05/2017/

Leetaru, K. (2018). Without transparency, democracy dies in the darkness of social media. *Forbes.* Retrieved from https://www.forbes.com/sites/kalevleetaru/2018/01/25/without-transparency-democracy-dies-in-the-darkness-of-social-media/#462f60257221

Losey, J. (2015). Surveillance of communications: A legitimization crisis and the need for transparency. *International Journal of Communication*, *9*, 3450–3459. Retrieved from https://ijoc.org/index.php/ijoc/article/view/3329/1495

MacKinnon, R., Maréchal, N., & Kumar, P. (2016). *Corporate accountability for a free and open Internet* (Global Commission on Internet Governance Paper Series No. 45). Retrieved from https://www.cigionline.org/publications/corporate-accountability-free-and-open-internet

Massanari, A. L. (2017). Contested play. In R. W. Gehl & M. Bakardjieva (Eds.), *Socialbots and their friends: Digital media and the automation of sociality* (pp. 110–127). New York, NY: Routledge.

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, *20*(6), 930–946. doi:10.1080/1369118X.2017.1293130

Miles, T. (2018, March 12). U.N. investigators cite Facebook role in Myanmar crisis. *Reuters*. Retrieved from https://www.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigators-cite-facebook-role-in-myanmar-crisis-idUSKCN1GO2PN

Newland, E., Nolan, C., Wong, C., & York, J. (2011). *Account deactivation and content removal: Guiding principles and practices for companies and users.* Cambridge, MA: The Berkman Center for Internet & Society and The Center for Democracy & Technology.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York, NY: New York University Press.

Parsons, C. (2019). The (in)effectiveness of voluntarily produced transparency reports. *Business & Society*, *58*(1), 103–131. doi:10.1177/0007650317717957

Ranking Digital Rights. (2018). *2018 corporate accountability index.* Retrieved from https://rankingdigitalrights.org/index2018/assets/static/download/RDRindex2018report.pdf

*R v. Sussex Justices*, *Ex parte McCarthy*, 1 KB 256 (High Court of Justice 1924).

Robb, A. (2017, November 16). Anatomy of a fake news scandal. *Rolling Stone.* Retrieved from https://www.rollingstone.com/politics/news/pizzagate-anatomy-of-a-fake-news-scandal-w511904

Roberts, S. (2016). Commercial content moderation: Digital laborers' dirty work. In S. U. Noble & B. Tynes (Eds.), *The intersectional Internet: Race, sex, class and culture online*. New York, NY: Peter Lang. Retrieved from http://ir.lib.uwo.ca/commpub/12

Rosen, J. (2013, April 29). The delete squad. *New Republic*. Retrieved from http://www.newrepublic.com/article/113045/free-speech-internet-silicon-valley-making-rules

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). An algorithm audit. In S. P. Gangadharan, V. Eubanks, & S. Barocas (Eds.), *Data and discrimination: Selected essays* (pp. 6–10). Washington, DC: Open Technology Institute and New America. Retrieved from https://www.newamerica.org/oti/policy-papers/data-and-discrimination/

Santa Clara Principles on Transparency and Accountability in Content Moderation. (2018, May 7). Retrieved from https://newamericadotorg.s3.amazonaws.com/documents/Santa_Clara_Principles.pdf

Schwencke, K. (2017, May 4). How one major Internet company helps serve up hate on the Web. *ProPublica*. https://www.propublica.org/article/how-cloudflare-helps-serve-up-hate-on-the-web

Suzor, N. P., Van Geelen, T., & West, S. M. (2018). Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette*, *80*(4), 385–400. doi:10.1177/1748048518757142

UNESCO. (2014). Fostering freedom online: The role of Internet intermediaries. Retrieved from http://unesdoc.unesco.org/images/0023/002311/231162e.pdf

Waldron, J. (2008). The concept and the rule of law. *Georgia Law Review*, *43*, 1–62. Retrieved from https://ssrn.com/abstract=1273005

West, S. M. (2018, May). *Policing the digital semicommons: Researching content moderation practices by social media companies*. Paper presented at the International Communication Association Conference, San Diego, CA.

Witt, A., Suzor, N.P., & Huggins, A. (forthcoming). The rule of law on Instagram: An evaluation of the moderation of images depicting women's bodies. *UNSW Law Journal, 42*(2).

Woolery, L., Budish, R. H., & Bankston, K. (2016). *The transparency reporting Toolkit: Best practices for reporting on U.S. government requests for user information*. Cambridge, MA: Berkman Center for Internet & Society. Retrieved from https://dash.harvard.edu/handle/1/28552578