Automated Platform Governance Through Visibility and Scale: On the Transformational Power of AutoModerator

Lucas Wright

Abstract
When platforms use algorithms to moderate content, how should researchers understand the impact on moderators and users? Much of the existing literature on this question views moderation as a series of decision-making tasks and evaluates moderation algorithms based on their accuracy. Drawing on literature from the field of platform governance, I argue that content moderation is more than a series of discrete decisions but rather a complex system of rules, mechanism, and procedures. Research must therefore articulate how automated moderation alters the broader regime of governance on a platform. To demonstrate this, I report on the findings of a qualitative study on the Reddit bot AutoModerator, using interviews and trace ethnography. I find that the scale of the bot allows moderators to carefully manage the visibility of content and content moderation on Reddit, fundamentally transforming the basic rules of governance on the platform.

Keywords
automated content moderation, platform governance, algorithmic visibility
the literature as novel software systems—tools introduced to make community governance more efficient rather than a transformative force that can establish a new regime of governance (Jhaver et al., 2019b; Kiene & Hill, 2020; Seering et al., 2019). As I show in this article, previous work on automated content moderation has largely focused on the decision-making capabilities of algorithms. But the expansion of the Reddit shadowban enabled by this third-party bot raises a different set of questions. Regardless of whether AutoModerator content removals are accurate or biased, research must ask how it has transformed the rules of governance on Reddit and by what mechanisms it is able to do so. How does a platform’s regime of governance—the collection of software, design, people, policies, and processes that orient a system of content moderation toward certain values and assumptions about how governance is done—get reconfigured by automation and scale?

This article examines that transformation, asking how AutoModerator establishes new rules of governance on Reddit. I find that the scale of the bot allows it to carefully manage visibilities—of content and of the act of moderation itself. In doing so, AutoModerator alters basic rules about account creation and the meaning of content removals. In this case, visibility is not merely an attribute of governance (i.e., the algorithm’s decisions are more or less transparent) but a mechanism by which new forms of governance are achieved. The question of researching automated moderation systems is therefore not merely “Does the algorithm make the correct decision?” or even “How is the bot used by moderators?” but also “What does the bot make more or less visible within a system of governance?” I argue that our understanding of automated moderation systems cannot stop at audits of their accuracy but must also theorize how introducing automation transforms governance through scale and visibility, even when an algorithm’s decisions are considered accurate and free of bias.

**Background**

Why should scholars of automated governance care about content moderation on Reddit? It is a good place to study the form of platform governance enabled by algorithmic tools due to its open design, reliance on volunteers, and its federated system of governance (Massanari, 2015). Reddit is divided into a number of subreddits based upon interests, ranging from the ordinary to the absurd (e.g., r/hockey, r/breadstapledtotrees), which are created and moderated by volunteer users. This distributed system of governance allows individual subreddits to develop their own moderation tools, and researchers can examine these systems more easily than proprietary algorithms developed by companies like Facebook and Google.

On 19 January 2012, Chad Birch, a volunteer moderator for the largest video gaming subreddit, r/gaming, announced that he had created a bot account capable of automating moderation tasks via the platform’s Application Programming Interface (API). Initially, moderators had to give the AutoModerator account moderator privileges in their subreddit and send their configurations to Birch to program into the bot. In 2013, Birch was hired by Reddit and developed a system where subreddits could edit a wiki page dedicated to their subreddit to set their subreddit’s configuration, eliminating the need for Birch to configure the bot manually. In 2015, most of the bot’s functions were integrated into the internal processes of the platform, rather than as an external process operating through an account. This integration allows AutoModerator to act on posts as they are uploaded to the platform, meaning that its comments and removals are almost instantaneous.

A configuration for the bot consists of one or more checks and one or more actions. A check looks for the presence or the lack of some condition, be it a word or list of words in the title, username of the author, link in the body, and so on. Actions can be set to remove a post, deleting it permanently; report a post, leaving it visible in the subreddit and adding it to the modqueue (a list of posts for moderators to review) for human review; filter a post, removing it from the subreddit and adding it to the modqueue for review; comment on a post with a pre-set message; or message the moderators about the post through “modmail.” Unlike other moderation algorithms, AutoModerator does not utilize machine learning algorithms, meaning that it does not become better at recognizing certain patterns as it receives more data. Instead, it is a pattern-matching system that only acts in the well-defined cases described in the conditions set by moderators.

**Literature Review**

In recent years, scholarship on how platform companies manage user data, design their products, make and enforce rules for content, algorithmically sort and promote posts, and manage groups has coalesced with legal and policy research on how governments do and should regulate the technology industry. Gorwa (2019) offers a useful overview of this new field of platform governance, showing the overlapping agenda of research on algorithmic culture, the political economy of platform companies, and studies of user experience design and social computing. The goal of knitting together these adjacent areas of study is to understand how a complex assemblage of regulatory regimes, technologies, and political and economic incentives enact particular forms of governance.

Much of the recent work on content moderation has focused on platforms that regulate their users’ speech on the largest scales with teams of paid moderators. Work in this area has focused on companies like Facebook and Google, covering the quasi-legal nature of their regimes of governance (Klonick, 2017; Suzor, 2019), the exploitative, hidden labor that enables that governance (Roberts, 2019), and the urgent need for greater accountability from platforms that
mediate global speech (Gillespie, 2018; Kaye, 2019). Another line of work has concentrated on platforms like Reddit and Wikipedia, where moderation is largely handled by user volunteers (Grimmelmann, 2015). Research on these platforms has shed light on the community-developed rules and volunteer labor that sustains these platforms (Fiesler et al., 2018; Matias, 2019).

Importantly, the literature on both community and commercial moderation has established that platforms do more than host content—they use the mechanisms of content moderation to structure, guide, and influence people’s behavior, or, in other words, they engage in governance (Grimmelmann, 2015). And the collection of rules, processes, design, and decision-making systems they develop to moderate forms a particular regime of governance. The regime is not merely a means of enacting governance but also a representation of the purpose and values of moderation on that platform. For example, questions like whether the barrier to entry on a platform is low or high, whether a removal comes with an explanation or not, whether users can appeal takedowns, and so on are all asking what the basic premise of the purpose of moderation is on a platform—its regime of governance. In this article, I consider how introducing automation does not merely extend a regime of governance by making it bigger and more efficient but transforms it and alters the basic rules of participating on Reddit.

It is therefore worth considering how automation has been treated in the existing literature and public debate on platform governance. First, many scholars take a decision-making approach to researching automated content moderation systems. This usually means that they focus on evaluating the accuracy of a moderation algorithm. For example, studies that measure bias in algorithms inherited from training data (Binns et al., 2017), test an algorithm against difficult edge cases to identify weaknesses (Hosseini et al., 2017), and assess whether automated content removals were correctly identified (Urban et al., 2017) all take a decision-making approach to evaluating moderation algorithms.

Much of the public debate over automated content moderation takes place within the narrow bounds of accuracy. Critics point to errors in algorithmic decisions as evidence that automated content moderation poses a risk of over-censoring speech (Citron, 2017; Duarte et al., 2018; Keller, 2020; Llansó et al., 2020). Platform company executives tout the benefits of implementing and using it but must also include a deontological analysis of the regime of governance it imposes.

While questions about decision-making are important, the public’s evaluation of automated content moderation cannot stop there, as an automated intervention can be perfectly accurate and transparent and still lead to undesirable outcomes (Chancellor et al., 2016). This is often because the impact of a moderation algorithm is determined by the assumptions people make about automated decision-making and the behavioral adjustments they make in response (Esami et al., 2016; Jhaver et al., 2019a). This is the focus of a second common approach to studying automated content moderation—researching how moderation algorithms are experienced by both users and moderators.

For example, research has found that automation changes how users perceive moderation of their posts. Users who lack a clear understanding of how moderation works are left to develop their own “folk theories” about how the system works, and these theories guide how the users interpret and adapt to moderation, even when they are wrong (Myers West, 2018). Other studies have observed how users strategically adapt to automated systems to circumvent bans on certain behavior, such as Instagram’s ban on content promoting eating disorders (Chancellor et al., 2016). Finally, even when moderation algorithms work correctly, they may have unintended chilling effects on other forms of protected speech if recipients of takedown notices choose to err on the side of speaking less to avoid future sanctions (Matias et al., 2020).

Research in computer science has focused on describing the software systems and user experience of automated moderation tools. This pragmatic line of research describes how they are developed, how moderators incorporate bots into their labor, and proposes recommendations for how the software can be improved (Halfaker et al., 2013; Kiene & Hill, 2020; Kiene et al., 2019; Seering et al., 2019). For example, in their study of AutoModerator, Jhaver et al. (2019b) ask how moderators use the bot to enact content regulation and what benefits and challenges are involved in its use. They describe how the software behind the bot works, how it was adopted and eventually integrated into the platform, and how moderators navigate its limitations. As computer scientists focused on understanding the design and use of systems, they show that automated moderation is a collaborative task between humans and machines.

Focusing on how individual moderators use AutoModerator in their own subreddits can miss how the bot scales across the entire platform, taking the form of an invisible infrastructure with its own social and political influence on the platform itself, regardless of how specific subreddits deploy it. A different but complementary research agenda would not only ask how automation extends and enacts an existing regime of governance but would also ask how it transforms it. Specifically, I ask how automation transforms the premise of governance on Reddit through scale and visibility. While these concepts come up in Jhaver et al. (2019b), they are viewed as qualities of the system that moderators must learn to engage with, adapt to, and evaluate in a utilitarian sense. However, as Burk and Gillespie (2008) argue in their analysis of digital rights management software, the evaluation of a new technology that regulates human behavior cannot be limited to the social costs and benefits of implementing and using it but must also include a deontological analysis of the regime of governance it imposes.
A third approach, which I take in this article, theorizes how we think and talk about moderation by clarifying how an algorithm transforms the broader system of governance in which it operates. For example, Emma Llansó (2020) argues that, regardless of how accurate an algorithm is, proactive automated filtering of content cannot escape the pitfalls of prior restraint on speech and the threat it poses to freedom of expression. This is because automated filtering is not merely a faster version of ex post human review of content, but it alters the basic premise of how speech is governed by removing it before anyone sees it. Stuart Geiger’s (2014) research on third-party Wikipedia editing bots shows that the platform is largely governed by a third-party code hosted on third-party servers, undermining the perspective in which platforms are cast as spatially bounded territories, code is cast as having the force of law over such territories, […] and users are cast as subjects who enter territories and are subsequently governed by code. (p. 9)

These bots not only substitute less efficient human decision-making for more efficient automated decision-making, but they also provide for new sources of architectural governance on the platform, external from the company that builds and runs the platform.

An often overlooked question within this literature is the specific transformation enabled by the scale of automation. Platforms often refer to automation as the only means of moderating at “scale,” but as Tarleton Gillespie (2020) points out, scale is not the same as size. This is because scaling an action is not the same as doing it many, many times—it introduces a new force that makes something small into something large, or vice versa. But this shift in a regime of governance is often overlooked when scale is conceived as merely a matter of size. As Gillespie (2020) says, The claim that moderation at scale requires AI is a discursive justification for putting certain specific articulations into place—like hiring more human moderators, so as to produce training data, so as to later replace those moderators with AI. In the same breath, other approaches are dispensed with, as are any deeper interrogations of the capitalist, “growth at all costs” imperative that fuels these massive platforms in the first place. (p. 2)

This raises the question, does scaling moderation via automation merely extend a regime of governance to a greater number of cases or does it transform it? Legal scholar Jonathan Zittrain’s (2007) critique of automated law enforcement can serve as a guide for understanding how the scale of automation can transform a regime of governance. According to Zittrain, the norms of individual freedom of liberal societies depend on the discretion of law enforcement officials and the friction that makes enforcement difficult and expensive, yet automation offers complete, simultaneous enforcement of the law and constant surveillance of behavior. Rather than simply extending a pre-existing law enforcement regime to more cases and doing it many times, the perfect enforcement of automation at scale transforms the basic premises of the relationship between the public and the law.

In this article, I focus specifically on how the scale afforded by AutoModerator enables a form of governance that carefully manages the visibility of content in new and powerful ways. Echoing Zittrain’s (2007) analysis of automated law enforcement, I show that scale in automation does transform moderation, not merely by allowing moderators to remove more posts but by allowing them to manage visibility in new ways, making some acts of moderation visible to some users and not others. This simple incongruence fundamentally changes the regime of moderation on Reddit and undermines common conceptions about how platform power is exercised.

Existing literature on algorithmic visibility and platform governance has shown how algorithms can manipulate the visibility of information and people in powerful ways that reinforce historical injustices (Introna & Nissenbaum, 2000) and exert control over users by creating a constant threat of being made invisible (Bucher, 2012). As a result, people are incentivized to change their behavior to appeal to the algorithm and reach a larger audience, reorienting their practices according to its logic (Gillespie, 2014). This literature establishes that visibility is a key mechanism through which algorithmic systems govern users and their behavior.

In addition to algorithms making people more or less visible, an algorithmic system can itself be more or less visible, and researchers have expressed concern over systems that govern invisibly. For example, Gorwa et al. (2020) argue that the opacity of moderation algorithms further shields the decision-making processes of platforms from public scrutiny. And when platforms do introduce transparency, they do so strategically—often to direct the public’s attention toward certain practices and away from others (Flyverbom, 2016).

Visibility is therefore essential to platform governance, both in terms of how automated systems make people and information more or less visible and in terms of the visibility of the algorithm itself. My project builds on this understanding by asking how the scale afforded by automated moderation systems opens up new possibilities for managing visibilities, not only by making some information visible or invisible to all, but by allowing moderators to carefully manipulate the visibility of governance itself, making an act of moderation visible to some but not others. I ask: how does managing visibility by moderating at scale with AutoModerator create new forms of power and transform the basic premise of governance on Reddit?

**Methods**

How can researchers theorize how automation changes a platform’s regime of governance when a bot and the governance it imposes is dispersed across thousands of subreddits and its
actions are often designed to be invisible? To investigate the regime of governance imposed by AutoModerator, you need to understand the challenges it was developed to solve, how its simple design allowed it to move beyond its original purpose, and the mechanisms through which it transforms the regime of governance on Reddit. To capture these transformations, I conducted interviews with the bot’s creator and moderators who use it and conducted textual analysis of the bot’s documentation and posts from a subreddit where moderators can receive help using the bot. My approach is inspired by Geiger and Ribes’ (2011) description of trace ethnography, a method of participant observation that uses computational tools to collect and analyze trace data from online interactions. Combining trace ethnography with interviews allows me to produce a thick description of how AutoModerator scales governance and articulate how it allows moderators to manage visibilities to govern users in new ways.

I conducted interviews with Reddit moderators in June and July 2018 and analyzed public documents and Reddit posts about AutoModerator to better understand how Reddit moderators make sense of the bot and how they incorporate it into their work.¹ The first interview I conducted was with Chad Birch, the creator of AutoModerator and a former Reddit moderator and employee. I also interviewed three moderators who use AutoModerator in their work. My interviews with Birch and two of the moderators were conducted over the phone, with the audio recorded, and the third was conducted via Reddit’s private messaging system. The moderators were recruited through posts on r/AutoModerator, r/AskModerators, and r/ModerationTheory. First is a long-time moderator of a large news subreddit, the second moderates a small technology subreddit devoted to discussing a particular laptop model, and the third moderates several smaller, niche subreddits.

In addition to interviews, I also collected and analyzed a series of documents and Reddit posts that either introduce AutoModerator mechanics or announce changes to the bot. This collection includes six Reddit posts from Birch (u/Deimoroz on Reddit), one from Steve Huffman (co-founder and CEO, u/spez on Reddit), and one from u/sodypop, a Reddit administrator. These posts were selected because they represent snapshots in time over the course of the development of the bot. In addition, I analyzed comments left in response to these posts that give a sense of the reaction to each announcement. The posts range from Birch first announcing the creation of the bot in 2012 to Huffman announcing Birch’s departure from Reddit in 2016 and u/sodypop announcing a new AutoModerator feature in 2019.

My sample also includes four public documents: the AutoModerator documentation, a library of 45 common AutoModerator rules, an introduction to writing new rules, and an AutoModerator FAQ. These documents were useful for my own understanding of how the bot works, and they also demonstrate how Reddit communicates about AutoModerator and the uses of the bot encouraged by these.

The “Stupid” Efficiency of AutoModerator

Tracing AutoModerator’s origins and the logic behind its creation is essential for understanding the original purpose it was designed for, how it has grown beyond that purpose, and the changes it has imposed on Reddit’s system of governance. Birch’s account of the inspiration for AutoModerator focuses on a need for greater efficiency in handling simple and mundane enforcement of obvious rule violations. For example, when I asked why he created the bot, he told me:

Almost every text post [...] would get filtered because the spam filter was stupid. [...] One of the major reasons I ended up writing AutoModerator is because I spent much time just fixing really stupid decisions the spam filter made. [...] I thought, “I can probably just write a stupid bot that will go in there and automatically approve all of the text posts.” [...] So I figured I could write a bot that could do simple things like that, like look at the spam filtered posts and be like, “OK, all of these are text posts, they’re probably fine.”

Birch describes both Reddit’s spam filter and AutoModerator as stupid, but he uses the term in different ways. The spam filter is stupid because it tries to learn from
Social Media + Society

moderator behavior, it but does so poorly and over filters. AutoModerator is stupid because it is simple and flexible, and this allows it to only act in well-defined cases. This is evident in Birch’s examples of imagined uses for the bot:

[Reddit’s] tools are so minimal that you can’t even ban a domain from the subreddit. So, for example, if you’re a gaming subreddit or something, there might be, whatever, maybe people are spamming by posting some porn site or something. Obviously that should never be submitted to the subreddit, so you could, ideally you’d have a tool that would say, “OK, if anyone ever submits this site, just remove it because it never belongs here.” So other subreddits will need that for, “OK, our subreddit is just for pictures, why is someone submitting a news site? That’s wrong and needs to be removed.”

This perspective is also clear in his original post announcing AutoModerator, in which he describes the ideal use of the bot: “Check the domain, maybe check the user, see if some obvious good/bad keywords are in the title, etc.” Throughout our interview, Birch described aspects of moderation that are complex and intensive, but these were not the tasks that he had in mind for the bot when he created it.

After announcing the bot on r/TheoryofReddit on 19 January 2012, he submitted a post in the subreddit with the title “Data on automating in a default subreddit (/r/gaming)” on 23 February. At this point, AutoModerator had already been adopted by several other subreddits. After listing the conditions that he uses for removal and approval in r/gaming, Birch provides the number of actions taken by all human moderators versus the number taken by AutoModerator in the previous week on the subreddit (Image 1).

After citing these statistics, Birch highlights the time saved by AutoModerator approving 1,013 submissions that were wrongly removed by Reddit’s anti-spam filter. His statistics also show, however, that most of the bot’s actions were limited to confirming or reversing the decisions of the default anti-spam filter, not removing posts on its own. Birch recognizes this in the post, saying, “Humans are still handling most of the removals of visible posts, which is probably the part of the job that requires the most judgment, and impossible to automate for most cases.” Once again, Birch shows that his motivation for creating AutoModerator was fairly narrow: correcting an anti-spam filter that made a lot of obvious mistakes.

This explanation of the role of automation in Reddit’s governance is one of extension—it takes a process for regulating users’ speech (moderators checking posts against a series of simple rules), replicates the decision-making process, and

| Action type          | AutoModerator | %   | Humans | %   |
|----------------------|---------------|-----|--------|-----|
| Un-spam link         | 1013          | 94% | 63     | 6%  |
| Confirm spam link    | 539           | 73% | 196    | 27% |
| Remove link          | 120           | 30% | 278    | 70% |
| Re-approve reported link | 386    | 46% | 457    | 54% |
| Re-approve reported comment | 0     | 0%  | 223    | 100%|
| Un-spam comment      | 12            | 48% | 13     | 52% |
| Remove comment       | 4             | 9%  | 41     | 91% |
| Confirm spam comment | 0             | 0%  | 2      | 100%|
| **Total actions**    | **2075**      | **62%** | **1284** | **38%** |

Image 1. Birch’s table of the number of actions taken by AutoModerator and human moderators in 1 week on the r/gaming subreddit. The table is meant to show a large reduction in the number of tasks required by humans through the use of the bot.
extends that same system of governance to more posts than a human could check, increasing efficiency. The moderators I interviewed echoed this explanation of AutoModerator, with an emphasis on the need for automation to moderate large subreddits. For example, one moderator told me:

I couldn’t imagine moderating without AutoMod. It does so many important things. Image [sic] having a subreddit with 500k subscribers and getting hundreds of posts a day and going through them all by hand, reading thousands of comments to find bad language or phone numbers etc. That just wouldn’t work. And . . . 500k isn’t even that big, I think reddit’s [sic] biggest subs have over 15 million subscribers [sic].

This idea was frequently echoed in posts on r/AutoModerator. Many moderators I observed asking for help described the overwhelming volume of content in their communities as the challenge they needed the bot to solve.

This is an attractive explanation of the role of automation in platform governance, especially when the purpose of moderation is viewed as a series of decision-making tasks. In this view, the goal of automation is to create an algorithm that can match the decision-making of a human and scale it, preserving the original system of governance while making it more efficient. After all, AutoModerator does not have access to privileges beyond those of any human moderator. When it deletes a post (or moves it to the modqueue), it is taking the same action a human moderator would, only doing so programmatically rather than through the click of a button.

Yet doing something at scale is not equivalent to doing something many, many times. The extension explanation of how AutoModerator works holds the broader system of governance on Reddit in place, assuming that AutoModerator extends it rather than transforms it. As I will show, the scale and completeness of the simple, rote decisions AutoModerator makes powerfully transforms the form of governance it can enact, and visibility is the mechanism through which this transformation takes place. An evaluation of AutoModerator must therefore analyze not only its decision-making capabilities, but also the transformation in governance that it enacts.

Managing Visibilities

All moderators manage visibility of content when they delete posts or ban users. Even users have some input into content visibility on Reddit through the platform’s voting system. Like human moderators, AutoModerator can manage content visibility through removing posts, but the completeness and scale of automation allows the bot to systematically manipulate it. As a result, AutoModerator is capable of not only managing the visibility of content but of the act of moderation itself. It does this by making some aspects of moderation more visible to some people, and some aspects less visible to others.

Moderators have the ability to ban users from their subreddit as a default feature of the platform. The banned users are automatically notified of the ban, allowing them to create a new account and continue posting in the subreddit. Shadowbanning is a practice that is enabled by AutoModerator’s scale, completeness, and the administrative limbo of the spam queue. When a post is moved to this queue, it is removed from the subreddit for everyone except the author until a moderator approves or denies it. When AutoModerator is configured to shadowban a user, all of their posts are automatically moved to the spam queue where they can sit forever, leaving their contributions to the subreddit visible to the user but making them invisible to everyone else. When a single post from a user is moved to the spam queue, the effect is limited, but when every post a user makes is automatically moved there immediately, they are effectively erased from the subreddit.

Removals of content on social media are often assumed to be an all or nothing intervention—before the removal a post is visible to everyone with permission to view it and after the removal it is visible to no one (Grimmelmann, 2015). The AutoModerator shadowban plays on this assumption of total removal by carefully managing who can see what, thereby creating a new form of platform governance, one in which governance is uncertain and automatic, where new users must earn the right to contribute. By carefully manipulating the visibility of a user’s posts, AutoModerator fundamentally changes the platform’s rules about bans and moderator power.

This makeshift shadowban is only possible through automation and the scaling of individual removals. Because AutoModerator acts on every submission and comment as it is posted to the platform, a series of removals takes on the effect of the shadowban. A human moderator could theoretically search for new posts from a user and manually move them to the spam queue, but the effect would not be the same. The moderator might miss posts, the posts would likely be visible for at least a brief time before being removed, and they could never feasibly do this for a large number of users. So while the rule programmed into the bot is quite simple (“move any posts from x account to the spam queue”), the effect of the individual removals scaled is far beyond the sum of its parts.

AutoModerator shadowbans can be implemented manually by simply adding a username to a list of banned accounts in an AutoModerator configuration. But sometimes shadowbans are enforced automatically based on simple conditions like an account’s age or karma score (upvotes minus downvotes on all of an account’s posts). For example, moderators I spoke with described setting AutoModerator to automatically shadowban accounts that were created within the past week or had fewer than zero karma points. This practice of automatically blocking new accounts is often referred to as “age gating” and typically used to weed out sock puppet accounts and spam bots, but two moderators I interviewed expressed concern that this practice unfairly discriminates against new users and may even discourage some Reddit users from participating. One moderator expressed concerns
about the effect this might have on new Reddit users who get swept up in an age gating configuration:

Especially if a lot of those removals happen silently so that you don’t know that anything that you’ve written in the first month of being on the site hasn’t been available to any other person to see [. . . ] We get a bunch of outbursts from really distraught users who have been trying to participate and who wonder, “Why is nobody talking to me?” [. . . ] When we explain to them what has been going on, they get really disillusioned with the site and a lot of them end up being toxic users.

Another moderator echoed these concerns:

I think one topic that is really, really important to how Reddit functions is how different communities use age gates—so filtering out content from brand new accounts. [. . . ] A lot of other communities have a lot harsher age gating conditions for being able to participate. Which means that if you’re exposed to Reddit for the first time and 20 of the 30 largest communities don’t allow you to write a comment for the first month you’re on the site, is that a site you ever want to spend time on?

Because the default assumption is that new users can make posts right away and many subreddits don’t advertise their age gating conditions, AutoModerator quietly alters the rules of the account creation process, creating new requirements for participating on the platform.

This transformative force enabled by the scale of automation is concealed by language that presents moderation algorithms as narrow solutions to specific problems. Media scholar Finn Brunton argues that the concept of spam is “expansive and vague” and has a “productive blurriness” that opens up space for groups to deliberate about its meaning and the acceptable boundaries for speech in their communities (Brunton, 2013, p. 8). What is commonly viewed as a fixed category of speech is instead a site for discussion and negotiation. But this productive blurriness can also conceal the role of automated solutions in closing off deliberation, as was pointed out by one moderator I interviewed:

As you speak with different mods of different communities that talk about spam, spam is a word that has many different meanings on Reddit [. . . ]. When people say that they have antispam conditions that use AutoModerator, that can be a lot of different things, and it’s very difficult to sort of pin down what is spam and therefore how is AutoModerator used to combat an extreme amount of content that obviously isn’t written by humans. [. . . ] It makes talking about AutoModerator and how AutoModerator deals with spam and sort of inorganic contributions very difficult.

When AutoModerator is implemented in a subreddit, it adopts the productive blurriness of spam and other contested categories of speech like hate speech and misinformation. The specific configurations of the bot in a subreddit are usually kept private, and the framing of automation as a narrow solution to a clear problem rather than a transformational force limits our ability to discuss that transformation.

Take, for example, how age gating is often justified in the name of fighting spam, as one moderator explained:

There’s this real issue regarding how to keep spammers out and how new users to a site that’s growing really, really quickly are caught by these automatic conditions. It’s a massive concern, and I think it’s something that the moderator community as a whole takes way, way too lightly because it’s really satisfying to get rid of all these spammers automatically and it’s easy to forget all the people, all the new users who effectively don’t have a voice.

This is reflected in the library of common AutoModerator rules, in which age gating is presented as a simple means of weeding out “throwaway accounts,” a problem frequently associated with spam (“Library of Common Rules,” n.d.).

As discussed above, the automation that enables age gating is a transformative force that can alter some of the fundamental rules of governance on Reddit, but when wrapped in the vague language of fighting spam, the platform-altering power of that transformation is lost. Using AutoModerator to automatically block these accounts is reduced to a narrow solution to spam, obscuring the change to the fundamental premise of account creation and newcomer participation on Reddit. This language limits our understanding of AutoModerator to a tool for decision-making rather than a means of establishing new systems of governance.

And with this blurry lens, the transformation in governance enabled by AutoModerator is made to look like a simple, narrow solution to a narrow problem. By invoking the blurriness of spam and deploying automation as an easy, “satisfying” solution, AutoModerator resolves the problem of spam by collapsing the space for deliberation described by Brunton. And it does so frictionlessly, as the role of automation is not immediately apparent to onlookers—Reddit’s account creation process does not mention the barriers to participation created by common age gating configurations, and the shadowban allows moderators to silently enforce it.

In practice, the productive blurriness of automated governance can obscure the sources and motivations for certain uses of a bot like AutoModerator, as new forms of governance are routed through its configurations. Consider for example, a controversy in a technology subreddit over moderating discount codes. One moderator I spoke with described how the scale of AutoModerator enabled new regulations on user behavior while also obscuring the role of external interests in shaping community policies. The moderator had known for some time that the members of his subreddit were sharing corporate employee discount codes, but he chose to let the content stay up due to his “laissez-faire” approach to moderation. The manufacturer of the computers the codes were for eventually contacted him and asked him to remove them. He agreed to manually delete the codes, but he found
that this was time consuming and inefficient, so he gave up until he learned about AutoModerator. At the time of our interview, he was considering creating an AutoModerator rule to automatically remove any post that contained text matching the unique pattern used in the discount codes.

As this moderator made clear, enforcing a policy against discount codes would have been impossible without AutoModerator, but this enforcement is not perfect. For example, AutoModerator cannot read text in images, so users of the subreddit could bypass the bot by uploading an image of the code rather than sharing it as a text post. There are several concerns that could be raised in response to this example about accuracy, censorship, and a lack of transparency in rule creation and enforcement. But I want to highlight how, despite the clear and obvious cases Birch had in mind when creating the bot, the productive blurriness of spam and AutoModerator conceived as a narrow solution allows a behavior broadly accepted by the community to become easily problematized and made governable according to external, corporate interests.

In this example, a form of speech that was previously accepted by both the community and the moderator becomes governable through a system of automation designed for spam. And even though the enforcement would be easily circumvented, AutoModerator takes a prior system of governance in which enforcing a ban on discount codes would have been time consuming, ineffective, and likely visible to the community and enables an easy, mostly invisible solution. This is not to say that this use of AutoModerator is illegitimate. It is well within the Reddit terms of service and broader societal norms to limit the spread of the discount codes. The key point in this case is how automation enables a new regime of governance via scale and visibility, not merely making the existing work of removing codes more efficient or accurate, but to alter the boundaries of the platform by altering what is governable, how it is governed, and then covering over this change with the blurriness of spam.

**Conclusion**

Regulators and critics of social media often call for platforms to remove more content, pushing them toward automation as a silver bullet that will scale their moderation regimes to meet these demands (Bloch-Webha, 2021). To ensure these moderation algorithms are making the correct decisions when removing content, many of these proposals also require that platforms publish transparency reports disclosing how their moderation algorithms work, and platforms defend their automated systems by touting accuracy statistics. But these proposals are premised on the idea that automation preserves and extends a regime of governance, as long as the decision-making is replicated accurately. And the challenge of scale is described as repeating this millions of times a day. But what if automation, and the scale it affords, transforms the regime, instituting new conditions about who makes and enforces rules, transforming governance as a result?

In this article, I’ve shown how the Reddit moderation bot AutoModerator both transforms the broader system of governance on Reddit and adopts the blurriness of spam, concealing that transformation. I argue that researchers and regulators need to move past the widespread focus on accuracy and bias and also ask what sort of regime an automated moderation system imposes. Such an analysis requires an understanding of the history of an algorithm to understand the conditions from which it emerged, as well as a critical analysis of new forms of governance it enables. It may be the case that an algorithm which is perfectly accurate is found to be undesirable because of the regime of governance it creates. At the very least, regulators need to be aware of the unintended consequences that pushing platforms toward automation may create.

These unintended consequences are often difficult to foresee because of the blurriness of the categories of speech automation is introduced to eliminate, such as “spam” and “online harms.” Because of this blurriness, it becomes harder to see how introducing automation to platform governance can transform it. The extent of the transformation enabled by AutoModerator wasn’t clear to Birch when he created the bot and it’s not always clear to moderators when they adopt it. And while other moderation algorithms will function differently than AutoModerator, the possibility for poorly understood transformations remains. As regulators and platforms push for greater automation, researchers must study these transformations in addition to auditing decision-making.

Governance cannot be understood simply as a series of discrete decisions, regardless of the conditions under which those decisions are reached. We must also theorize moderation tools as forms of governance that transform the broader system and impose new forms of power—establishing new rules about who influences moderation decisions and how those decisions are enforced. Errors and bias in decision-making should be a clear reason for critiquing and regulating an algorithmic system, but audits of automated moderation systems must also scrutinize the mechanisms through which an algorithm governs and consider how this might impose a new regime within a platform.

Other automated moderation systems are likely to produce different regimes of governance than what I have described here. AutoModerator emerged from the very particular conditions of Reddit: an open API through which third-party bots can easily act on the platform and a federated system of volunteers who could choose to adopt new tools that were not developed internally by the platform company. On more computationally closed platforms like Facebook or YouTube, this transformation is more likely to come from automated moderation tools developed and deployed internally, and they might utilize other mechanisms for altering the existing governance regime. This may be especially true in the case of algorithms that use machine learning rather
than AutoModerator’s style of pattern matching. But I propose that a key insight of this study will remain true across these cases: that scaling governance through automation does more than add efficiency to an existing regime of governance, it transforms it. And, as a result, we must carefully analyze this change in governance, even when the algorithm’s decision-making is accurate.

Finally, regulators and auditors should recognize that power to transform governance through automation does not lie only with platform companies in every instance. Of course, this varies from platform to platform according to factors like degree of community-based moderation and level of public API access, but the example of AutoModerator shows that automated governance is not the sole domain of platform—indeed, moderator-developed bots often serve the explicit purpose of circumventing platform power and introducing new regimes. As the AutoModerator model is adopted on other platforms like Twitch and Discord, it is more important than ever for researchers to investigate the array of sources of transformative power within platform governance.

Acknowledgements

The author would like to express his gratitude to J. Nathan Matias, Andrew Przybylski, and Amelia Hassoun for their guidance and generous support at various stages of this project. He is also thankful for the thoughtful feedback he received on drafts of this piece from Tarleton Gillespie, Brooke Duffy, attendees of the 2019 Rise of Platforms ICA post-conference, the Cornell New Media & Society working group, and two anonymous reviewers.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Lucas Wright https://orcid.org/0000-0002-3392-1928

Note

1. This research was approved by the University of Oxford Central University Research Ethics Committee. CUREC approval number: SSH OII C1A 18 058.

References

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), Social informatics (pp. 405–415). Springer. https://doi.org/10.1007/978-3-319-67256-4_32

Bloch-Wehba, H. (2021). Automation in moderation. Cornell International Law Journal, 53, 41–96.

Boellstorff, T., Nardi, B., Pearce, C., & Taylor, T. L. (2012). Ethnography and virtual worlds: A handbook of method. Princeton University Press. https://doi.org/10.1515/9781400845286

Brunton, F. (2013). Spam: A shadow history of the Internet. MIT Press.

Bucher, T. (2012). Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. New Media & Society, 14(7), 1164–1180. https://doi.org/10.1177/1461444812400159

Berk, D. L., & Gillespie, T. (2008). Autonomy and morality in DRM and anti-circumvention law (SSRN scholarly paper no. ID 1146448). Social Science Research Network. https://papers.ssrn.com/abstract=1146448

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016). #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing (pp. 1201–1213). http://www.munmund.net/pubs/cscw16_thyghgapp.pdf

Citron, D. K. (2017). Extremist speech, compelled conformity, and censorship creep. Notre Dame Law Review, 93(3), 1035–1072.

Duarte, N., Llansó, E., & Loup, A. (2018). Mixed messages? The limits of automated social media content analysis [Conference session]. Conference on Fairness, Accountability, and Transparency. https://cdt.org/wp-content/uploads/2017/12/FAT-conference-draft-2018.pdf

Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., & Kirlik, A. (2016). First I “like” it, then I hide it: Folk theories of social feeds. In CHI conference on human factors in computing systems (pp. 2371–2382). Association for Computing Machinery. https://doi.org/10.1145/2858036.2858494

Feiner, L. (2021, May 28). New bill would “open up Big Tech’s hood,” make companies explain how they decide which content to show. CNBC. https://www.cnbc.com/2021/05/28/markey-matsui-bill-would-force-social-media-content-algorithm-transparency.html

Fiesler, C., Jiang, J., “Aaron” McCann, J., Frye, K., & Brubaker, J. R. (2018). Reddit rules! Characterizing an ecosystem of governance [Conference session]. Twelfth International AAAI Conference on Web and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17898

Flyverbom, M. (2016, 28 May). Disclosing and concealing: Internet governance, information control and the management of visibility. Internet Policy Review. https://doi.org/10.14763/2016.3.428

Geiger, R. S. (2014). Bots, bespoke, code and the materiality of software platforms. Information, Communication & Society, 17(3), 342–356. https://doi.org/10.1080/1369118X.2013.873069

Geiger, R. S., & Ribes, D. (2011, January 4–7). Trace ethnography: Following coordination through documentary practices [Conference session]. 2011 44th Hawaii International Conference on System Sciences, Kauai, HI, Unite States. https://doi.org/10.1109/HICSS.2011.455

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), Media technologies (pp. 167–194). MIT Press. https://doi.org/10.7551/mitpress/9780262525374.003.0009

Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.
Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2), Article 94323. https://doi.org/10.1177/2053951720943234

Gorwa, R. (2019). What is platform governance? Information, Communication & Society, 22(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), Article 989794. https://doi.org/10.1177/2053951719897945

Grimmelmann, J. (2015). The virtues of moderation. Yale Journal of Law & Technology, 17, Article 69.

Halfaker, A., Geiger, R. S., Morgan, J. T., & Riedl, J. (2013). The rise and decline of an open collaboration system: How Wikipedia’s reaction to popularity is causing its decline. American Behavioral Scientist, 57(5), 664–688. https://doi.org/10.1177/0002764212469365

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google’s perspective API built for detecting toxic comments. Arxiv:1702. http://arxiv.org/abs/1702.08138

Introna, L., & Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matters. The Information Society, 16(3), 169–185. https://doi.org/10.1080/01972240050133634

Jhaver, S., Appling, D. S., Gilbert, E., & Bruckman, A. (2019a). Did you suspect the post would be removed?: Understanding user reactions to content removals on Reddit. Proceedings of the ACM on Human-Computer Interaction, 3, 1–33. https://doi.org/10.1145/3359294

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019b). Human-machine collaboration for content regulation: The case of Reddit automoderator. ACM Transactions on Computer-Human Interaction, 26(5), 1–35. https://doi.org/10.1145/3338243

Kaye, D. (2019). Speech police: The global struggle to govern the Internet. Columbia Global Reports.

Keller, D. (2020). Facebook Filters, Fundamental Rights, and the CJEU’s Glawischni-Piesczek Ruling. GRUR International, 69(6), 616–623. https://doi.org/10.1093/grurint/ikaa047

Kiene, C., & Hill, B. M. (2020). Who uses bots? A statistical analysis of bot usage in moderation teams. In Extended abstracts of the 2020 CHI conference on human factors in computing systems (pp. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3334480.3382960

Kiene, C., Jiang, J. A., & Hill, B. M. (2019). Technological frames and user innovation: Exploring technological change in community moderation teams. Proceedings of the ACM on Human-Computer Interaction, 3, Article 44. https://doi.org/10.1145/3359146

Klontick, K. (2017). The new governors: The people, rules, and processes governing online speech. Harvard Law Review, 131, Article 1598.

Library of Common Rules. (n.d.). Reddit. https://www.reddit.com/r/AutoModerator/wiki/library

Llansó, E., van Hoboken, J., Leeressen, P., & Harambam, J. (2020). Artificial intelligence, content moderation, and freedom of expression. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

Llansó, E. J. (2020). No amount of “AI” in content moderation will solve filtering’s prior restraint problem. BigData & Society, 7(1), Article 92068. https://doi.org/10.1177/20539517209200686

Massanari, A. L. (2015). Participatory culture, community, and play: Learning from Reddit (1st New ed.). Peter Lang: International Academic Publishers.

Matias, J. N. (2019). The civic labor of volunteer moderators online. Social Media + Society, 3(2), Article 98377. https://doi.org/10.1177/2056305119837778

Matias, J. N., Mou, M. E., Penney, J., Klein, M., & Wright, L. (2020). Do automated legal threats reduce freedom of expression online? Preliminary results from a natural experiment. https://doi.org/10.17605/OSF.IO/NC7E2

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media & Society, 20(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Roberts, S. T. (2019). Behind the screen: Content moderation in the shadows of social media (Illustrated ed.). Yale University Press.

Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. New Media & Society, 21(7), 1417–1443. https://doi.org/10.1177/1461444818821316

Suzor, N. P. (2019). Lawless: The secret rules that govern our digital lives. Cambridge University Press.

Urban, J. M., Karaginis, J., & Schofield, B. (2017). Notice and takedown in everyday practice (SSRN scholarly paper no. ID 2755628). Social Science Research Network. https://doi.org/10.1177/1461444818783659

Zittrain, J. (2007). Perfect enforcement on tomorrow’s internet. In K. Yeung & R. Brownsword (Eds.), Regulating technologies: Legal futures, regulatory frames and technological fixes (pp. 125–156). Bloomsbury.

Author Biography

Lucas Wright (MSc, University of Oxford) is a PhD student in the field of communication at Cornell University. His research interests include digital governance of online communities, social media content policy, and the relationship between algorithms and human behavior.