Community Moderation

John Crawford and Aaron Mak

SMGI Lab Spring 2023

THE JUSTICE COLLABORATORY
Yale Law School

SOCIAL MEDIA GOVERNANCE INITIATIVE

Project Overview

The Community Moderation Project (CMP) is a qualitative study of how community moderation facilitates online discourse. This project was conceived of and worked on by students, John Crawford and Aaron Mak, as part of the Justice Collaboratory's Spring 2023 Social Media Governance lab offered to Yale Law School students through the generous support of a grant from the Stavros Niarchos Foundation. The goal of this project is to better understand the work, insights, and needs of community moderators, as well as to determine if community moderation should be utilized more widely in the future. This semester, we examined the role of volunteer moderators as intermediaries between Reddit administrators (admins) and subreddit community members. We paid special attention to controversial subreddits in order to understand how moderators enforce compliance with platform-wide rules with which they do not necessarily agree.

Reddit is a social media platform that markets itself as "a large community made up of thousands of smaller communities." These smaller communities, known as 'subreddits,' "are created and moderated by" users. Reddit has a content policy that all subreddits must obey. This policy consists of eight general rules that are common among social media platforms, such as prohibitions against harassment and illegal content. Enforcement is typically left to community moderators, with admins only intervening in cases of repeated or egregious violations. Community moderators may also establish and enforce their own rules specific to the subreddit(s) they oversee. These subreddit rules, which exist on top of platform-wide rules, vary greatly between subreddits in terms of both scale and focus.

Process

Our initial idea was to interview moderators dealing with a wide range of subjects which might present unique moderation challenges. We identified 78 subreddits and loosely separated them into five categories:

1. **topic-focused** (i.e. r/Pets)
2. **nonpolitical** (i.e. r/ShowerThoughts)
3. **identity** (i.e. r/Filipino)
4. **politics** (i.e. r/politics)
5. **controversial** (i.e. r/TheRedPill)

17 of the 29 'controversial' subreddits we found were active and accessible without moderator permission.

We then decided to narrow the focus of our project so that any meaningful results we produced would not be lost in a sea of superfluous information. We replaced our five categories with two subjects which were contentious but not overtly political: the men's rights movement and vaccine/lockdown skepticism. We chose these subjects for two reasons. First, content related to the men's rights movement and vaccine/lockdown skepticism seem to frequently break platform-wide rules. Moderators of subreddits which allow posts on these subjects therefore likely have to grapple with enforcing Reddit's content policy or risk action by admins. Second, we imagined that it would be easier to contact and interview moderators of these subreddits than moderators of subreddits dealing with many other controversial subjects. Some insular, identity-based communities like r/BlackPeopleTwitter, r/FemaleDatingStrategy, or r/aznidentity have been accused of letting problematic content go unmoderated. These subreddits, however, severely limit communication for those who are not members of the in-group. Some far-right political communities have made national headlines for the toxicity of their content. These subreddits, however, have been banned en-mass by Reddit in recent years.

We eventually identified 25 active and accessible subreddits dealing with the men's rights movement or vaccine/lockdown skepticism. 12 of these subreddits were related to the men's rights movement, and 13 were related to vaccine/lockdown skepticism. Some of these subreddits promoted posts related to our chosen subjects, while others sought to correct, shame, or reject users who shared such content. For reasons which will be

discussed later, we also subsequently added another seven subreddits dealing with content unrelated to the men's rights movement or vaccine/lockdown skepticism, bringing the total number of subreddits we investigated to 32.

The second step in our process was to gather basic information on the 25 subreddits whose moderators we had originally decided to contact. The purpose of this effort was to be able to later evaluate if there were any differences in the perspectives shared by our interviewees based on the scale or focus of the subreddit(s) they moderated. We created a Reddit account, YaleSMGI, and joined each subreddit that we were investigating. We then recorded each subreddits' member and moderator counts, loosely estimated how active each subreddit was, and determined which subreddits were quarantined or otherwise had restricted access.

We next developed a spreadsheet of all of the moderators for our selected subreddits. This list eventually totalled 373 moderators. We then screened these moderators in the hopes of only sending outreach messages to those who might respond. We tried to deduce whether each account was run by a person or a bot, software applications used on many Reddit subreddits to perform basic moderation tasks, based on usernames and past posts. We also determined whether each account was active by checking if they had engaged with any content on Reddit in the last few weeks. 226 of the moderators we investigated appeared to be current, human Reddit users.

The next step in our process was to reach out to the moderators on our list. We developed [1] an outreach message formatted as a short letter which could be personally addressed and sent to each moderator via Reddit's chat function; [2] a survey for moderators interested in our project; and [3] a brief description of our project posted on the Justice Collabotary's website to lend us legitimacy. [1] then incorporated both [2] and [3] as links. We then created a document containing all of the rules for each of our selected subreddits, as listed on the subreddits themselves, for future reference. We also later created a Calendly for moderators who had filled out our survey to sign up for an interview slot.

Ironically, our initial attempt at outreach quickly failed because of an imprecise, presumably automatic moderation action against our account. On March 10th, 2023, we contacted 13 moderators. Unfortunately, Aaron discovered on March 11th, 2023 that Reddit had shadowbanned our account, Yale-SMGI, presumably because it flagged our messages to moderators as spam. While we were logged in, there was no obvious sign that the account had been banned. The chat function simply kept failing to load. When we logged out and checked our account page, however, an error message appeared stating the

following: “Sorry, nobody on Reddit goes by that name. The person may have been banned or the username is incorrect.” Newly-created Reddit accounts are more likely to be shadowbanned for spam as they have no established posting and chat history.

Shadowbanning is a well-known and highly problematic action conducted by many social media companies. Users receive no notice as to why this action was taken, or even notice that the action was taken at all, hence the term ‘shadowbanning.’ While we could still navigate Reddit as normal, any posts we made would not be seen by anybody but us. We attempted to reinstate our account using Reddit’s appeal suspension form on March 11th, 2023. This form allows users to submit a brief statement explaining that they did not violate the platform’s rules and were suspended in error. As of April 24th, 2023, however, over a month after we were shadowbanned, we have not received any response.

Our work was delayed by about a week and a half as we tried to figure out a new game plan. We eventually decided to use Aaron’s personal Reddit account, aarontmak, as it had an established, albeit limited, posting and chat history and was therefore presumably less likely to be shadowbanned for spam. The risk was of course that using aarontmak to send the same outreach messages could result in a second shadowban. Furthermore, we had no idea how many messages we could send in a given time period before triggering Reddit’s spam filters since the platform had never acknowledged that we had surpassed such a threshold in the first place. We therefore restarted slowly, sending five to eight messages a day for the first week or so, eventually working our way up to eight to ten or more messages a day in subsequent weeks. Thankfully, Reddit ultimately did not shadowban aarontmak.

Our response rate was consistently low. Approximately one in forty moderators would complete our survey. Only about half of survey respondents signed up for interviews. This eventually forced us to reach out to moderators for communities beyond our initial selection. We expanded the scope of our research to include seven subreddits related to the far right and conspiracy theories. In the end, we received 13 survey responses and conducted six interviews from about 200 outreach messages. We could not reach out to all 226 active moderators on our list because some had disabled their chat function. Additionally, we ultimately decided not to message the moderators of r/WitchesVsPatriarchy, as further research indicated that this community rarely dealt with posts related to the men’s rights movement, the subject which initially caused us to add it to our list.

When moderators responded to our outreach message, we chatted and encouraged them

to complete our survey and sign up for an interview. Several moderators explicitly declined to complete our survey or sign up for an interview. One moderator informed us that their subreddit "[doesn't] allow these types of surveys." Another moderator declined to formally participate but explained how they moderated in a brief message via chat. Still others told us that they were not interested. We respected the wishes of these moderators and did not message them further.

We conducted our interviews via Zoom or Reddit chat based on the preference of the moderator we were interviewing. Five of the six interviews were conducted via Zoom. We informed our interviewees that we would maintain their privacy by omitting Reddit usernames and any identifying personal information from our public findings. In this report, we have therefore replaced all Reddit usernames with stand-ins. We also offered to omit the names of the subreddits that our interviewees moderated, instead using generalized descriptions of each community. No one accepted this offer. As such, this report includes subreddit names, which allows us to discuss the insights shared by our interviewees with more specificity.

Our interviews began with a formulaic introduction based on a script. We reminded our interviewees who we were and what kind of work we were doing, as some of our interviews were conducted several weeks after our initial conversations with the moderators in question. We also reiterated that their Reddit usernames would not be included in our final report. We then asked a series of questions selected from a prewritten list based on the direction of the conversation. We loosely sorted our questions into six categories for internal purposes. These categories were:

1. **Introductory Questions**
2. **Structure of Subreddit Moderation**
3. **Subreddit Members' Experience**
4. **Views on Reddit's Platform-Wide Rules**
5. **Responding to Controversy**
6. **Thoughts on the Future.**

The table below contains some sample questions from our list as well as shortened versions of some interviewees' answers.

Question Category	Question	Abbreviated Sample Answer
Introductory Questions	What drew you to Reddit?	"I joined because I was [] losing my mind from these lockdowns. And my views on this subject were apparently diametrically opposite to everybody else's. So I was just kind of looking to see if there was any other like-minded souls out there."
Structure of Subreddit Moderation	How does your subreddit select new moderators?	"Every now and then they would post on the subreddit... If you're interested, send us a message."
Subreddit Members' Experience	How do you think members of your subreddit view Reddit administrators?	"I think it's not really transparent to them when certain content gets removed, or decisions get made... sometimes it's hard to tell if it came from the admins at Reddit or if it came from the moderation team."
Views on Reddit's Platform-Wide Rules	Has Reddit changed its platform-wide rules since you became a moderator?	"I don't really follow the [Terms of Service]. I just kind of go on what's been told to me. In general, they just change it to make it more strict or less strict based on basic parameters."
Responding to Controversy	Has your subreddit faced any controversy?	"Some of the biggest controversies... is that there's a couple users that people want banned and they do a really good job of walking the line."
Thoughts on the Future	What, if any, changes do you think Reddit should make to better support its community moderators?	"If they could, it's already been done. Because frankly it's just a matter of manpower."

We did not limit ourselves to the questions on our prewritten list. We asked interviewees follow-up questions as needed. We also departed from our prepared questions altogether as necessary. In one interview, the individual that we were chatting with indicated that they performed only limited moderation duties on one of the subreddits on our original list of 25. We instead shifted the conversation to talk about their more robust responsibilities on two other subreddits.

Findings

The six interviews we conducted provided a variety of insights into moderators' experiences and interactions with Reddit admins, as well as more general information as to how social media platforms might envision moderation more effectively. Our interviewees were:

1. **Alpha**, a moderator for **r/MensLib**
2. **Beta**; a moderator for **r/HermanCainAward**
3. **Gamma**, a moderator for **r/HermanCainAward**
4. **Delta**, a moderator for **r/rape**
5. **Epsilon**, a moderator for **r/LockdownSkepticism**
6. **Zeta**, a moderator for **r/Conservative**

The table below contains the stated purpose of each subreddit moderated by one of our interviewees.

Subreddit	Interviewed Moderator(s)	Purpose	Number of members	Number of moderators
r/MensLib	Alpha	"[B]uilding a new dialogue on the real issues facing men through positivity, inclusiveness, and solutions-building."	227 thousand	18
r/HermanCainAward	Beta + Gamma	"Nominees have made public declaration of their anti-mask, anti-vax, or Covid-hoax views, followed by admission to hospital for Covid. The Award is granted upon the nominee's release from their Earthly shackles."	498 thousand	14
r/rape	Delta	"All survivors/victims of sexual violence, their families, and friends are welcome here."	52.5 thousand	9

Subreddit	Interviewed Moderator(s)	Purpose	Number of members	Number of moderators
r/LockdownSkepticism	Epsilon	"Interdisciplinary examination of lockdowns & other pandemic policies. We acknowledge the threat of COVID-19. We are also concerned about the policies' impact on our physical & mental health, human rights, and economy."	55.7 thousand	34
r/Conservative	Zeta	"We provide a place on Reddit for conservatives, both fiscal and social, to read and discuss political and cultural issues from a distinctly conservative point of view."	1.0 million	36

The following analysis examines the interviewees' responses in light of five themes which emerged over the course of our conversations.

1. **Reddit's Decentralized Moderation System**
2. **Drawbacks of Customization**
3. **Moderator Selection and Member-Moderator Interactions**
4. **Reconciling Reddit Rules with Subreddit Values**
5. **The Role of Outside Systems**

Reddit's Decentralized Moderation System

> "r/MensLib is a space for constructive discussion of men's issues. Moderators reserve complete discretion to maintain a positive atmosphere, including removing comments and submissions, and banning offenders."
> *—Rule Zero, r/MensLib*

Unlike most social media platforms, Reddit polices content using a decentralized moderation system. All subreddits have one or more volunteer moderators who implement both Reddit's content policy and subreddit-specific rules. In practice, they can also decide how strictly to interpret certain rules, limited only by the threat that Reddit admins will intervene in extreme cases for failure to enforce platform-wide rules. Subreddits can therefore try out different approaches to moderation on a smaller scale in a manner akin to Justice Brandeis's idea of the fifty states as "[l]aboratories of democracy." Many view this system as preferable to Reddit constantly rolling out experimental rules which must be enforced across the entire platform.

The moderators we interviewed often noted that the penalties they imposed for violations of Reddit's platform-wide rules were more severe than those for subreddit-specific rules. This is understandable, as violations of Reddit's rules can result in action against an entire subreddit, up to and including a ban of the community itself. According to Gamma, r/HermanCainAward issues three-day bans for violations of subreddit rules, and five-day bans for violations of Reddit-wide rules. However, some interviewees claimed that subreddit-specific rules were more effective for keeping discourse in check. For example, Epsilon noted that r/Lockdownskepticism's civility rule is useful for tamping down on public shaming and ad hominem attacks. Reddit's platform-wide rules only prohibit harassment. Another important rule on r/LockdownSkepticism is the prohibition against "low quality vaccine content," which refers to ungrounded speculation about coronavirus vaccines. The subreddit instituted this rule during the initial rollout of the vaccines, which Epsilon says has helped to restrict conspiratorial content. Gamma also pointed to r/HermanCainAward's prohibition against celebrations of death as being particularly useful. Decentralization allows subreddits to tailor rules to the kind of content that they typically host, and test out rules to determine which are the most effective.

Another customizable tool that moderators have at their disposal is an automoderator, which is a bot available to every subreddit that automatically takes action against certain kinds of posts and comments. It was presumably a platform-wide automoderator that

shadowbanned our first account after we sent dozens of nearly-identical interview inquiries via Reddit's chat function. Subreddits can configure the settings of their automoderator to target specific types of content, such as posts containing spam content. Automoderators can also help subreddits ensure that posts stay on topic. "We've really tailored our automod configuration to cover things specific to our subreddit that we see a lot of," said Gamma. "We also use it to filter out posts. If the post isn't an award, that's not a story of somebody kicking the can, it gets screened."

Epsilon shared that r/LockdownSkepticism's automoderator regulates content that is taboo because it is beyond the subreddit's Overton window. While r/LockdownSkepticism often pushes the bounds of what is allowable under Reddit's coronavirus misinformation policies, it takes a hard line against bigotry and conspiracy theories that the moderators consider outlandish. "If there's a comment containing racist terms or whatever, that'll get flagged, of course, and auto-removed," said Epsilon. "If there's something containing a word that tends to cause trouble, like for example 'plandemic' – that's a word associated with the more conspiracy-oriented types – it won't be auto-removed, but it will be auto-flagged." ('Plandemic' is a viral video from 2020 that falsely claims that Microsoft CEO Bill Gates was partly responsible for the spread of the coronavirus). Overall, this decentralization allows Reddit to offload much of their content-moderation responsibilities onto individual subreddits. At the same time, the subreddits have more leeway to shape their communities' cultures within the bounds of Reddit's platform-wide rules.

Drawbacks of Customization

> "[I]deally, if you have good rules, its easy to make a decision on whether something should be approved or removed"
> *—Beta, moderator for r/HermanCainAward*

There are two main drawbacks to Reddit's decentralized moderation system that were clearly apparent in our interviews. The first is that moderators are volunteers and may therefore be absent when they are needed most. Subreddits often go through cycles of activity and inactivity. Periods of high activity can overwhelm subreddit moderation teams as they cannot pay to bring in extra assistance. Gamma identified the period between August 2021 and January 2022, when the Omicron variant was sweeping across the country, as the busiest for r/HermanCainAward. According to Gamma's estimates, there were 10 to 15 active moderators putting in 20 hours of work per week at the time to keep up with a daily stream of 300 posts, 20,000 comments, and 50 million page views per post. This workload led some to leave. "We had one moderator who was doing the work of 20. For two weeks, he just put in so many hours," Gamma said. "I think he burnt himself out."

Zeta similarly noted that certain news events can set off a wave of posts that the moderators of r/Conservative have trouble handling. Zeta specifically pointed to the aftermath of mass shootings and the run up to elections as being particularly hectic, with a high volume of posts and rule-breaking. There are certain strategies that the moderation team has developed for such circumstances, such as relying on the 'flair' system that allows subreddits to identify established users. During periods of controversy, the subreddit will often only allow users with 'flair' to write posts. Despite these restrictions, however, Zeta admitted that r/Conservative is often unable to effectively handle an influx of traffic. "When there's major stuff happening, especially in the first few days, we don't have the manpower to deal with all of that," Zeta said. "Sometimes I think the Reddit admins have spared us from a really bad fate, because they understand that it's just a giant amount of work that we have to do."

In addition to the issue of periods of high activity, the moderation problems faced by r/Conservative may also be related to the second drawback to Reddit's decentralized moderation system: the need for technical expertise on subreddits' moderation teams. Beta explained that r/HermanCainAward currently uses a combination of Reddit's moderation tools and a third-party plugin to automate most of their moderation activities. When removing a post, Beta can automatically send messages to the user explaining why

the post was removed. Beta estimated using this method “98 times out of 100.” In contrast, Beta described r/conservative’s automoderator as “junk” after being given a copy from a “defector.” If a subreddit has one or more moderators with technical expertise, then content can be effectively policed with minimal efforts. In the absence of this knowledge, however, moderation becomes much more time-consuming.

Our interview with Delta demonstrates the challenges faced by subreddits without technical expertise. r/rape uses its automoderator to posts and comments containing certain keywords or phrases, such as slurs or rape threats. The rest of the moderation, however, is done manually. Delta thought that the current moderation system works “fairly effective[ly],” but also shared that the automoderators often failed to catch spam, forcing moderators to remove such content themselves.

Moderator Selection and Member-Moderator Interactions

> "People get upset about moderating decisions all the time. A lot of people don't view mods in the best light and think they are power-hungry."
> *—Delta, a moderator for r/rape*

Another potential issue inherent to Reddit's decentralized moderation system is that a small group of users within any given subreddit, the moderators, have immense power over the community as a whole. As has been mentioned, moderators can establish rules for their subreddits above and beyond those mandated by Reddit admins, for better or for worse. They can also take actions against users in the absence of or in contravention of existing subreddit rules, subject only to oversight by their fellow moderators and potentially Reddit admins. Furthermore, a user can only become a moderator of an established, active subreddit through an invitation by its current moderators. In theory, this system ensures that new moderators will enforce rules in a manner consistent with the community's values. The status quo, however, has also led to community members accusing the moderators of various subreddits of being 'oligarchs' or 'tyrants.' Despite these accusations, our interviewees all seemed to think that their subreddits, at least, had fair practices for selecting new moderators and moderating content.

Several of the moderators we interviewed indicated that the process for selecting new moderators on their subreddits was ad-hoc but not arbitrary. Delta became a moderator for r/rape after "a current mod reached out to me privately to ask me to consider joining" based on Delta's posting history on another subreddit. Zeta indicated that while r/Conservative sometimes encourages community members to apply to open moderator positions, "it's really a matter of . . . what the top guys think of you." Similar to how Delta became a moderator, active users on r/Conservative might receive a direct message from current moderators offering them the position. These findings are consistent with those of Joseph Seering et al. in the New Media & Society study "Moderator Engagement and Community Development in the Age of Algorithms." As the researchers found, "[m]oderators are most commonly selected for the position because they were standout members of the community; head moderators tend to look for members who understand the community's values, have the maturity to set an example, and can enforce the rules appropriately."

These informal selection processes clearly indicate that many subreddits select new

moderators non democratically, although neither Delta nor Zeta framed their experiences negatively. This might be because, as current moderators, they are now members of the in-group, and therefore see nothing wrong with the current system. It also could be, however, that informal selection processes benefit communities as a whole by ensuring that no one becomes a moderator without adhering to shared subreddit values. After all, current moderators can likely vet a handful of users that they have frequently seen post on the subreddit far more effectively than dozens of users applying for a position of power. In fact, Gamma stated that r/HermanCainAward purposefully does not consider candidates who reached out to indicate an interest in becoming a moderator. “There are a lot of people out there who are just trying to become a moderator on all of the popular subreddits,” Gamma said. “They’re collecting subreddits, and they’re not really helpful.” For this reason, r/HermanCainAward does not use public moderator recruiting posts, but instead also messages candidates individually.

Subreddits can have relatively formal processes for selecting new moderators. Some subreddits strive to ensure that new moderators are in tune with the existing moderators, the subreddit, and the real-world community that the subreddit is meant to represent. For one such subreddit, interested subreddit members first need to complete a survey asking them about their experience and beliefs, as well as answer several open-ended questions about what they would do as a moderator in certain situations. The existing moderators then invite prospective moderators into a group chat to get to know them better. The moderation team then presents their final candidate(s) to the community in a discussion post to allow the community to have some say in the matter and conduct their own vetting. New moderators are then provided with a wiki document on the subreddit’s moderation practices to guide their work. Finally, they must complete a probationary period in which they have more limited powers and robust oversight than established moderators. This system seems more akin to some companies’ hiring and onboarding processes for new employees than the selection process for a niche, voluntary, part-time position.

Regardless of how they were selected for their position, every moderator must deal with how community members perceive their actions. Some interviewees indicated that there is little friction between their moderation teams and the subreddit members. They attributed the lack of discord to members understanding the delicate conditions under which moderators often have to work. For instance, Epsilon maintained that most members of r/LockdownSkepticism understand that pressure from Reddit admins often coerces moderators into implementing rules with which they do not agree. “We did make it very clear during these periods where we had to be a little more careful that the reason we’re

doing this is to protect the sub from any action that we might not want, like quarantining or banning," Epsilon said. "I think most of [the members] understood that." In certain cases, there is a symbiotic relationship between moderators and Reddit admins when it comes to unpopular rules. Subreddit moderators can blame Reddit admins to shield themselves from criticism. In turn, individual Reddit admins do not face much blowback, as they rarely implement policy decisions themselves.

A common sentiment shared amongst other interviewees was that users frequently disagree with moderators' decisions and are often not shy about sharing their perspective. Delta said that "[p]eople get upset about moderating decisions all the time," and has personally "received a lot of harassment [] on [R]eddit as a result of being a mod." Alpha suggested that the majority of users become hostile to moderators after their content is removed for violating subreddit rules, with many complaining about censorship or echo chambers. Beta offered an intuitive explanation as to why there is often a divide between how moderators view their role and how their decisions are perceived by community members. "A lot of it's apathy [on the part of moderators]," Beta said. A community member "will write out a really long post or a really long comment, put a lot of effort into it, but... it gets removed... and that sucks." From the perspective of the moderator, however, "that's just one of... twenty decisions they've had to make that morning and they haven't had their coffee yet."

The opportunity to appeal seems vital given that moderators and users seem to often disagree as to whether a content removal or other adverse action was justified. Several interviewees claimed that they at least occasionally reversed previous moderation decisions made by themselves or another moderator. Their descriptions, however, suggested that some moderators were more willing to change their mind than others, and some subreddits had better appeals processes than others. Furthermore, some subreddits' rigidity or fluidity regarding these decisions seems to be baked into their purpose and design.

Our interviews with Delta and Alpha clearly demonstrate that some moderators and subreddits are more open to appeals than others. Delta said that Reddit users often reach out to moderators from r/rape in the aftermath of an adverse action. These decisions might be reversed if the moderators and user "come to an agreement." This agreement typically involved an "acknowledge[ment] [of] their actions," an "apology," and a "commit[ment] to not doing it again." In considering whether to reverse a decision, Delta looks for "someone [who] just didn't realize the gravity of their actions" and "seem[s]

genuinely remorseful." This system of moderator discretion regarding reversals is likely valuable given the sensitive nature of the topics discussed on r/rape, but falls short of an actual merits-based appeals process which is likely to satisfy all users.

In contrast, Alpha indicated that the moderators of r/MensLib have created a robust appeals system which is "largely underutilized by any user that might actually have a complaint." Users can appeal an adverse action taken by one moderator to the moderation team as a whole, and another moderator will then act as a second set of eyes. Should an adverse action create significant disagreement within the moderation team, then the best course of action will be decided based upon a seniority system. According to Alpha, however, users who participated in the appeals process tend to be "fairly loaded with assumptions" that their content was removed because a moderator disagreed with their politics rather than because they violated any rules. As part of the appeals, the moderators "will explain [their] perspective." Most appeals do not get overturned. Alpha believes that this is because the "moderators have the wiki [which is accessible to all community members] right in front of them and [the] rules are laid out in pretty explicit terms."

Our interviewees' thoughts on moderator selection and member-moderator interactions raise an important question which might be explored in a future research project: Do subreddit members agree with their moderators' belief that the current system is effective? Our interviewees all described different systems for choosing new moderators and hearing members' complaints about content removals and other adverse actions. Each system seems relatively functional. Our only source of information, however, are the individuals tasked with enforcement. In order for community moderation to be viable, it will need buy-in from members as well as moderators.

Reconciling Reddit Rules with Subreddit Values

> "Reddit admins only contact us when they really, really need to. It's very, very sparse communication. And thats a good thing, they're leaving us to do our own thing."
> *—Beta, a moderator for r/HermanCainAward*

Several moderators discussed the dilemma of having to enforce Reddit's platform-wide rules when they run counter to the purposes of their subreddits. Enforcing such rules can result in community members protesting and leaving the subreddit en masse. Yet, failing to enforce these rules can result in Reddit quarantining or even banning the entire subreddit. (Reddit CEO Steve Huffman told the New York Times in 2020 that the platform banned r/The_Donald, a large pro-Donald Trump subreddit, after executives were unable to convince the community's moderators to more strictly enforce the rules.) All of our interviewees who had faced this conflict said that they chose to enforce Reddit's platform-wide rules rather than risk the survival of their respective subreddits. As we only interviewed moderators from subreddits that have not been banned, however, it is unclear if this view is universal to all Reddit moderators or only those who choose to stay on the platform.

According to r/Conservative moderator Zeta, the subreddit is keen on avoiding a ban. Senior leadership stresses the importance of survival to moderators. "We have the mission statement pretty clear, and it's nailed into all of our heads: we want to keep the sub alive," said Zeta. As a result, Reddit makes r/Conservative's moderators enforce rules with which Zeta does not agree, such as the platform's prohibitions against misgendering and harassing people of a protected status. Zeta explained that enacting rules that go against one's political beliefs is "kind of a muscle memory." As a cautionary tale, Zeta cited Reddit's ban of r/LowderWithCrowder, a subreddit dedicated to the far-right political commentator Steven Crowder, in early April 2023. In Zeta's view, this was a result of the moderators' failure to distance themselves from their duties. "Their personal thoughts on the subject got in the way of them moderating stuff," Zeta said.

Even though there is a disconnect between Zeta's beliefs and moderation practices, Zeta asserted that the survival of r/Conservative is worth the sacrifice given its significance. Zeta argued that "[r/Conservative] is the only subreddit that is right-leaning of that size, and it's also the oldest," so a ban would "se[t] a precedent" which could lead to the removal of other, smaller right-wing political subreddits. As an example, Zeta pointed to Reddit removing r/The_Donald in 2020 as part of an effort to more explicitly combat hate speech. When Reddit banned r/The_Donald, it also banned about 2,000 other subreddits

to conform with its new policy initiative. Zeta claimed to not be too concerned about members leaving the r/Conservative because of a rule change. The subreddit has one million members, so Zeta thinks that it's unlikely that enough people could leave at once to significantly impact the community.

Gamma, a moderator for r/HermanCaidAward, described a similar challenge when Reddit admins contacted the subreddit's moderators. r/HermanCainAward highlights people who had made social media posts expressing skepticism about the COVID-19 pandemic and the efficacy of vaccines, and then subsequently died of the disease. During the height of the pandemic, the subreddit attracted a significant amount of attention not only from the wider Reddit community, but also the mainstream press. According to Gamma, Reddit admins expressed concern about posts on the subreddit featuring screenshots of Facebook posts from deceased coronavirus skeptics. The subreddit's original policy was to redact surnames but allow first names and profile pictures to remain public. Reddit admins informed the subreddit's moderators that full names and faces must be omitted from all posts. They also wanted r/HermanCainAward to more strictly moderate comments that violated the platform's rules regarding harassment and bullying, as some celebrated or called for the deaths of coronavirus skeptics.

The Reddit admin's directives led r/HermanCainAward to recruit more moderators and implement an automoderator. The moderators also wrote posts addressed to the members of the subreddit explaining the changes. These changes were controversial. Gamma said, "There were pitchforks. People were upset. I was upset." Gamma notes that the subreddit aimed to demonstrate that the coronavirus was not a hoax, but rather a dangerous disease that was killing people. Some members asserted that this message was not as compelling without displaying the profile pictures of the deceased. Gamma said, "If you couldn't see their face, it didn't really resonate. You couldn't feel that this was happening at the scale that it was." Subreddit members also argued that the people in the posts had a weaker right to privacy, as they were already deceased. "A lot of the members thought that we shouldn't listen to Reddit, and that we should just keep doing what we were doing," Gamma explained.

Despite community uproar, the moderators were more concerned about the continued survival of the subreddit. "When [the Reddit admins] reached out to us, we thought for sure that we were going to get quarantined and banned, and we were definitely moderating so that did not happen," said Gamma. "We didn't want to give [the Reddit admins] any room to quarantine or ban the subreddit." Gamma notes that r/HermanCainAward was appearing

almost every day on the front page of Reddit, where the most popular communities get the most exposure. The subreddit's mission was to widely broadcast the message that refusing to get vaccinated could result in death. Despite members' and moderators' disagreement with Reddit's platform-wide rules, the subreddit's survival and mission were more important.

Not all r/HermanCainAward's moderators were on the same page regarding complying with Reddit's platform-wide rules. Beta, for example, thought that the Reddit admins' message to the subreddit's moderators informing them that the community was in violation of the platform's terms of service was "totally reasonable." The platform simply wanted the subreddit "to enforce the rules that [are] enforce[d] for everyone else." Another, newer moderator, however, was "very, very angry at the world, and very angry at conservatives, anti-maskers, and anti-vaxxers, and really wanted to light fire to their world," and was "completely against" changing the subreddit in any way to comply with Reddit's rules. "He started doing the social media equivalent of burning bridges," Beta said, "he ragequit, and then he came back, and started airing our dirty laundry on the subreddit itself." Senior leadership therefore "had to let him go."

The subreddit r/LockdownSkepticism promotes a point of view contrary to that of r/HermanCainAward. Members of this subreddit generally criticize coronavirus mitigation rules from a libertarian standpoint. According to one moderator, Epsilon, the subreddit has also had to manage the tension between Reddit's rules and community values. While Reddit's admins have never contacted r/LockdownSkepticism directly, the subreddit nevertheless changed its moderation rules in order to avoid a ban.

Epsilon pointed to a moment in 2021 when Reddit quarantined and banned a host of subreddits promoting coronavirus skepticism, particularly those opposing masks and vaccines. In response, r/LockdownSkepticism began cracking down on anti-mask and anti-vax posts. For instance, the moderators took a stricter approach to posts asserting that the vaccines constituted a form of gene therapy. "We limited the discourse more than we would have liked to," said Epsilon. Epsilon claimed that, while some members left over the changes, most were understanding of the position that the subreddit was in.

Again, for Epsilon, the survival of the subreddit took precedence. Epsilon viewed r/LockdownSketpicism as a community of like-minded people that the interviewee had been unable to find elsewhere. The value of having that community was paramount. "This is really the only place that people had," Epsilon said. "Personally, before I found this subreddit, I felt like I was just the only sane [person] in the world, and everyone else was

losing their minds. Then I found some other sane people, and lots of other people felt the same. So we all value this place quite a lot and want to keep it around." The moderators did have discussions about what they would do if Reddit ever banned r/LockdownSkepticism. They considered the possibility of moving to a different platform, or making an archive of the subreddit's post so that it could be reinstated elsewhere.

The moderators who prioritized the survival of their subreddits over ideological disagreements with Reddit's sitewide rules justified their decisions by emphasizing the practical benefits of staying on the platform. Zeta predicted that a hypothetical ban of r/Conservative would set a precedent for Reddit to ban other right-leaning subreddits. Gamma contended that, because r/HermanCainAward was appearing on the front page of Reddit, it was instrumental in publicizing the dangers of the coronavirus. Epsilon stressed that r/LockdownSkepticism provided a community that members were having trouble finding offline.

These decisions indicate that community moderators can exist within a wider content-control ecosystem given the right incentives. This has important implications for the potential scalability of community moderation, as current and future social media platforms will likely continue to establish terms of service which maximize profitability and limit legal exposure. If community moderators understand that their decisions exist within an umbrella of acceptable discourse established by the platform, then more platforms might be willing to rely upon them. This finding, however, is limited by the nature of our project's process, as we only reached out to moderators of active subreddits rather than all subreddits that have ever existed on Reddit.

The Role of Outside Systems

> "[O]verall, of social media sites, Reddit's doing a pretty good job."
> *—Beta, a moderator for r/HermanCainAward*

Some of our interviewees relied on outside systems to fill in perceived limitations of Reddit's moderation tools. For example, Beta suggested that Reddit's 'modmail' tool was not ideally suited to the needs of r/HermanCainAward. The platform describes this feature as "[a] shared messaging system that moderators use to communicate with members of their communities and other redditors." Beta said that modmail was meant to allow moderators to communicate with their communities, Reddit admins, and amongst themselves. Beta felt, however, that modmail "was really a clunky interface" for discussions between moderators, and was confident that Reddit administrators could read moderators' messages if they wanted to. r/HermanCainAward moderators therefore primarily communicate with one another outside of Reddit. Alpha seconded the idea that moderator-to-moderator communication on Reddit "could definitely use more improvement."

Many Reddit moderators heavily rely on Discord, a social media platform focused on instant messaging, to communicate with one another regarding their work. r/HermanCainAward operates a Discord channel which is "split in two" sections. The first is just for moderators. The second is for both moderators and "[their] rockstars," the members of the subreddit who actually create content. Beta thought that "what [r/HermanCainAward] do[es] is fairly similar to other subreddits." Zeta said that r/Conservative primarily selects new moderators based on activity on its Discord server, indicating that key aspects of the subreddit's operations are offshored to another platform. Zeta also characterized Discord as r/Conservative's "second line of defense" in case Reddit banned the subreddit.

Reddit moderators' use of Discord does not seem to reflect negative feelings towards the platform's moderation tools as a whole. Beta said that "Reddit does give some very good tools" even if "a lot of people complain" and Beta personally uses a third-party plugin to more effectively moderate content. The current importance of outside systems to the operation of some subreddits instead seems to suggest that Reddit should simply continue to work on improving its features. Our interviewees did indicate that they noticed when Reddit made improvements to its moderation tools. In our discussion, Alpha specifically noted that Reddit recently added an anti-harassment feature to modmail which could filter

out people insulting moderators for removing their posts.

The lesson that can be learned from some subreddits' reliance on Discord is that moderators need an effective method to directly communicate with one another. If a subreddit is a community, then its moderators are a vital subgroup within that community keeping the whole group together. Moderators cannot do their job if they cannot talk to one another. On Reddit, moderators have worked around the platform's limitations. In order for community moderation to be adopted on a wider scale, however, social media platforms would need to ensure that moderators are able to remain in constant communication with one another without having to work around existing systems.

Recommendations For Reddit

Some interviewees recommended a number of small, concrete fixes to make moderation easier and more effective. As has been mentioned, several interviewees wanted Reddit to improve its modmail to allow moderators to more easily communicate with one another. Additionally, Gamma recommended that the platform develop a tool that allows subreddits to monitor where traffic is coming from when there is an influx of posts. Gamma said that one of the moderators on the team programmed a makeshift bot that monitors other hostile subreddits for mentions of r/HermanCainAward. This is typically a warning sign that outside users looking to cause trouble are about to invade r/HermanCainAward and post content that violates the subreddit's rules. Gamma notes while the makeshift bot is helpful, it would be better if Reddit made an official version of the tool.

Zeta further suggested that the platform import more tools from "old Reddit," which is an older version of the interface that was phased out with a redesign in 2018. One feature in particular that Zeta misses from old Reddit is the 'toolbox,' which was a suite of tools including a monitor that kept track of the number of posts that a certain user has made on other individual subreddits. Moderators can then check whether a user is active in subreddits that are hostile to r/Conservative and thus more likely to cause trouble if they enter the community. At the end of the day, though, the main problem that Zeta sees is a lack of manpower. Zeta does not think Reddit can solve this problem by simply creating more tools. Ironically, however, Beta indicated that r/HermanCainAward, which is likely one of r/Conservative's 'rival' subreddits, uses a third-party plugin that fulfills the role formerly occupied by toolbox.

Recommendations
For Justice Collaboratory's Future Research

Based on the results of our study, we have developed a list of potential steps that Justice Collaboratory's SMGI can take to further pursue this avenue of research:

Locate and interview individuals who formerly moderated subreddits which the platform has banned. Did these individuals attempt to save their subreddit by enforcing compliance with Reddit's platform-wide rules, or reject these requirements in favor of supporting community values? Did their community move elsewhere?

Conduct a survey of subreddits to empirically determine how new moderators are chosen. Compare subreddits with formal and informal moderator selection processes. Does how community moderators are chosen impact their behavior? Does it matter if moderators are chosen by existing moderators or the community as a whole?

Contact Reddit to collaborate on research, especially to get permission to send a higher volume of interview inquiries to moderators.

Interview Reddit admins about their views of, and interactions with, subreddit moderators. Subreddit moderators seemed to view Reddit admins as distant, looming figures. Do Reddit admins intend to give this impression?

Interview Reddit admins about when and why they decide to implement platform-wide rule changes or quarantine/ban large numbers of subreddits for noncompliance.

Extend this study to controversial Facebook groups and Discord servers.

Contact other social media platforms to determine why they have chosen not to adopt a community moderation model. While basic research seems to suggest that it is simply a problem of scale, this answer seems unconvincing given that subreddits with millions of members can be as effectively moderated as those with hundreds.