Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices

PRERNA JUNEJA, Virginia Tech, USA
DEEPIKA RAMA SUBRAMANIAN, Virginia Tech, USA
TANUSHREE MITRA, Virginia Tech, USA

Transparency in moderation practices is crucial to the success of an online community. To meet the growing demands of transparency and accountability, several academics came together and proposed the Santa Clara Principles on Transparency and Accountability in Content Moderation (SCP). In 2018, Reddit, home to uniquely moderated communities called subreddits, announced in its transparency report that the company is aligning its content moderation practices with the SCP. But do the moderators of subreddit communities follow these guidelines too? In this paper, we answer this question by employing a mixed-methods approach on public moderation logs collected from 204 subreddits over a period of five months, containing more than 0.5M instances of removals by both human moderators and AutoModerator. Our results reveal a lack of transparency in moderation practices. We find that while subreddits often rely on AutoModerator to sanction newcomer posts based on karma requirements and moderate uncivil content based on automated keyword lists, users are neither notified of these sanctions, nor are these practices formally stated in any of the subreddits' rules. We interviewed 13 Reddit moderators to hear their views on different facets of transparency and to determine why a lack of transparency is a widespread phenomenon. The interviews reveal that moderators' stance on transparency is divided, there is a lack of standardized process to appeal against content removal and Reddit's app and platform design often impede moderators' ability to be transparent in their moderation practices.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**;

Keywords: online communities, content moderation, rules, norms, transparency, mixed methods

ACM Reference Format:
Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices . In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4, GROUP, Article 17 (January 2020). ACM, New York, NY. 35 pages. https://doi.org/10.1145/3375197

1 INTRODUCTION

In 2019, moderators of subreddit r/Pics removed an image depicting the aftermath of the Tiananmen Square massacre that was well within the rules of the subreddit [22]. The post was later restored in response to the uproar. However, the event calls attention to two important shortcomings in the moderation process: the moderators did not indicate *what* part of the content triggered the moderation and *why*. To date, the content moderation process has mostly been non transparent and little information is available on how social media platforms make moderation decisions; even transparency reports only provide number of removals without capturing the context in which the

Authors' addresses: Prerna Juneja, Virginia Tech, Blacksburg, USA, prerna79@vt.edu; Deepika Rama Subramanian, Virginia Tech, Blacksburg, USA, dramasubramanian@vt.edu; Tanushree Mitra, Virginia Tech, Blacksburg, USA, tmitra@vt.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2020 Association for Computing Machinery.

2573-0142/2020/1-ART17 $15.00
https://doi.org/10.1145/3375197

removals occurred [60]. Content removal without justification and change in content moderation rules without notification have been common occurrences in all major social media platforms, like Facebook and Reddit [21, 57]. When asked to publicly comment on these contentious decisions, platforms respond with short, formal statements that rarely indicate the overall ideology of the organization's moderation systems. [21]. This lack of information can puzzle users [65], making it difficult for them to understand rules governing the platform's moderation policies or even learn from their experience after their content is sanctioned on the platform [64]. Lack of proper feedback also leads users to form folk theories and develop a belief of bias in the moderation process [64]. Thus, in recent times, there has been a huge demand for transparency and accountability in content moderation of all social media platforms [36, 60].

In an attempt to address what transparency and accountability entails in social media platforms, three partners—Queensland University of Technology (QUT), University of Southern California (USC) and Electronic Frontier Foundation (EFF) [1] jointly created the Santa Clara Principles (henceforth referred to as SCP). SCP outlines a set of minimum guidelines for transparency and an appeal process that internet platforms should follow. For example, it requires companies to provide detailed guidelines about what content is prohibited, explain how automated tools are used to detect problematic content and give users a reason for content removal. In response to these principles, Reddit, for the first time in 2018, included some statistics regarding the content that was removed by subreddit moderators and Reddit admins for violating their Content Policy [26]. The report stresses the fact that Reddit is aligning its content moderation practices with the SCP.

> *"It (transparency report) helps bring Reddit into line with The Santa Clara Principles on Transparency and Accountability in Content Moderation, the goals and spirit of which we support as a starting point for further conversation"*

While Reddit as a company claims to abide by the SCP, are its communities following these principles too? It is important to note that calls for transparency are not limited to Reddit's Content Policy, the company has also issued moderator guidelines (MG) [50] that reiterate how transparency is important to the platform. They ask moderators to have *"clear, concise and consistent"* guidelines and state that *"secret guidelines aren't fair to the users"*. In this study, we examine if content moderation practices in Reddit communities abide by the transparency guidelines outlined in the SCP and the moderator guidelines issued by the platform (MG).

Reddit is one of the largest and most popular discussion platforms. It consists of millions of communities called subreddits where people discuss a myriad of topics. These subreddit communities are governed by human moderators with assistance from an automated bot called AutoModerator. Thus, the platform provides us with a unique opportunity to study the transparency of human-machine decision making, an aspect rarely studied in previous literature (with an exception of a few studies like [29]). Reddit is also unique in its two-tier governance structure. While the company has a site wide Content Policy [51], every subreddit has a set of rules and norms. While rules are explicitly stated regulations, norms are defined as unspoken but understood moderation practices followed by moderators of the online communities. Looking through the dual lens of SCP and MG, we sketch how transparency works in Reddit's two tier governance enforced via platform's policies, community's rules and norms. We ask:

RQ1: How do content moderation practices in Reddit sub-communities align with principles of transparency outlined in the SCP guidelines?

RQ1a: Do all sanctioned posts correspond to a rule that was violated, where the rule is either explicitly stated in the community's rule-book or on Reddit's Content Policy?

[1] http://eff.org

RQ1b: Do moderators provide reasons for removal after taking a content moderation action on a user's post or comment?
RQ1c: What are the norms prevalent in various subreddit communities?
RQ1d: Are rules in subreddits clearly defined and elucidated? Are they enforced and operationalized in a consistent and transparent manner by human and auto moderators?

We employed a mixed methods approach to answer these research questions. First, we quantitatively identified "meta communities"— cohorts of subreddits that sanction similar kinds of transgressions. Next, we qualitatively mapped high ranking post/comment removals from every meta community to their corresponding subreddit rule violation as well as Reddit's Content Policy violations that led to the removals. Unmapped instances revealed unsaid norms that moderators followed to sanction such posts and comments. Among the several norms that surfaced through our analysis, the following are the most unsettling: (1) moderators sanctioned comments containing criticism of their moderation actions and (2) moderators themselves posted and removed rule-violating expletives, sometimes even while providing feedback to the community that they were moderating. Through our qualitative analysis, we also uncovered several moderation practices that violate SCP transparency guidelines. We observed that most of the subreddits present in our dataset do not notify users about the reason for content removal. We also found that enforcement and operationalization of rules is not transparent. AutoModerator configurations such as inclusion of blacklisted words and karma requirements have not been publicly revealed in subreddits' rules.

To understand the rationale behind the widespread transparency violations, we decided to interview moderators to comprehend their side of the story. SCP, along with RQ1 findings inspired our RQ2 questions.

RQ2: How do moderators view the following facets of transparency?
(a) Communities silently removing the content without notifying the user
(b) Revealing reasons for removals or citing rules while sanctioning content
(c) Vaguely worded subreddit's rules
(d) Transparent/non-transparent enforcement of rules
(e) User appeals against content removal

We interviewed 13 Reddit moderators. They had a divided stance on transparency. While one half believed being transparent during the moderation process is essential for healthy communities, the other half provided several insights on the pitfalls of being transparent. They believed notifying users about content removal and providing reasons for these removals act as negative reinforcement, thereby making users more uncivil. Being transparent about rule enforcement is also problematic. Miscreants and trolls could use this information to game the system. We found how the design of the Reddit platform acts as a hindrance and prevents moderators from being more transparent. Although Reddit does provide a way for users to appeal suspension or restriction of their accounts[2], moderators revealed that there is a lack of standardized appeal process for content removal across subreddits. In Reddit's world, where moderation is voluntary with scarce resources, it is important to appeal against a moderator's decision in an appropriate way. Based on the responses from moderators, we have compiled a complaint etiquette —guidelines that users should follow to appeal against content removal in order to get their case heard. Taken together, our findings suggest that while existing moderation practices such as not providing proper feedback are problematic, practicing complete transparency in rule enforcement is not pragmatic for social media platforms. There is a need to find the 'juste-milieu' in content moderation where healthy communities are cultivated by providing appropriate feedback while simultaneously avoiding abuse of this information by miscreants.

[2] https://www.reddit.com/appeals

2 STUDY CONTEXT: SANTA CLARA PRINCIPLES ON TRANSPARENCY AND ACCOUNTABILITY

In this section, we briefly discuss the Santa Clara Principles of Transparency and Accountability in content moderation. SCP includes a set of three recommendations that serve as a starting point, providing minimum levels of transparency and accountability. The recommendations provided are [49]:

(1) *Numbers*: Companies should publish the number of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines.
(2) *Notice*: Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension. They should also provide: (a) detailed information as to what content is disallowed, including examples of permissible and non-permissible content, (b) information about the guidelines used by the moderators and (c) explanation of how the automated tools are used for detection of non-permissible content during moderation.
(3) *Appeals*: Companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension.

In order to formulate these recommendations, the three partners (QUT, USC and EFF) undertook a thematic analysis of 380 survey responses that were submitted to the EFF's—a leading non-profit advocating free speech and digital privacy—website `onlinecensorship.org` by users who were seriously affected by the censorship of their content or temporary suspension of their accounts on social media platforms [44]. In addition to this, several deliberative sessions that took place at the "All Things in Moderation"[3] conference—a symposium organized by UCLA[4] on online content review, also contributed to the formulation of the guidelines [44]. As of 2019, twelve companies including Reddit and Facebook publicly support the Santa Clara Principles [19, 48, 62]. In this study, we focus mostly on *Notice*, the second SCP, while briefly discussing the accepted methods for appeals during our interviews with moderators. Examining the aspects of the second and third SCP gives the HCI community an excellent opportunity to 'design for transparency'. We leave the investigation of subreddits' adherence to the first SCP to future work.

The rest of this paper is organized as follows. We start by reviewing related research. After briefly describing our dataset, we describe RQ1 methods followed by results detailing SCP violations. Next, we describe RQ2 method followed by describing in detail the insights gained from our interviews with moderators. Finally, we discuss our results, design implications of our work and future directions before concluding with the limitations of our study.

3 BACKGROUND

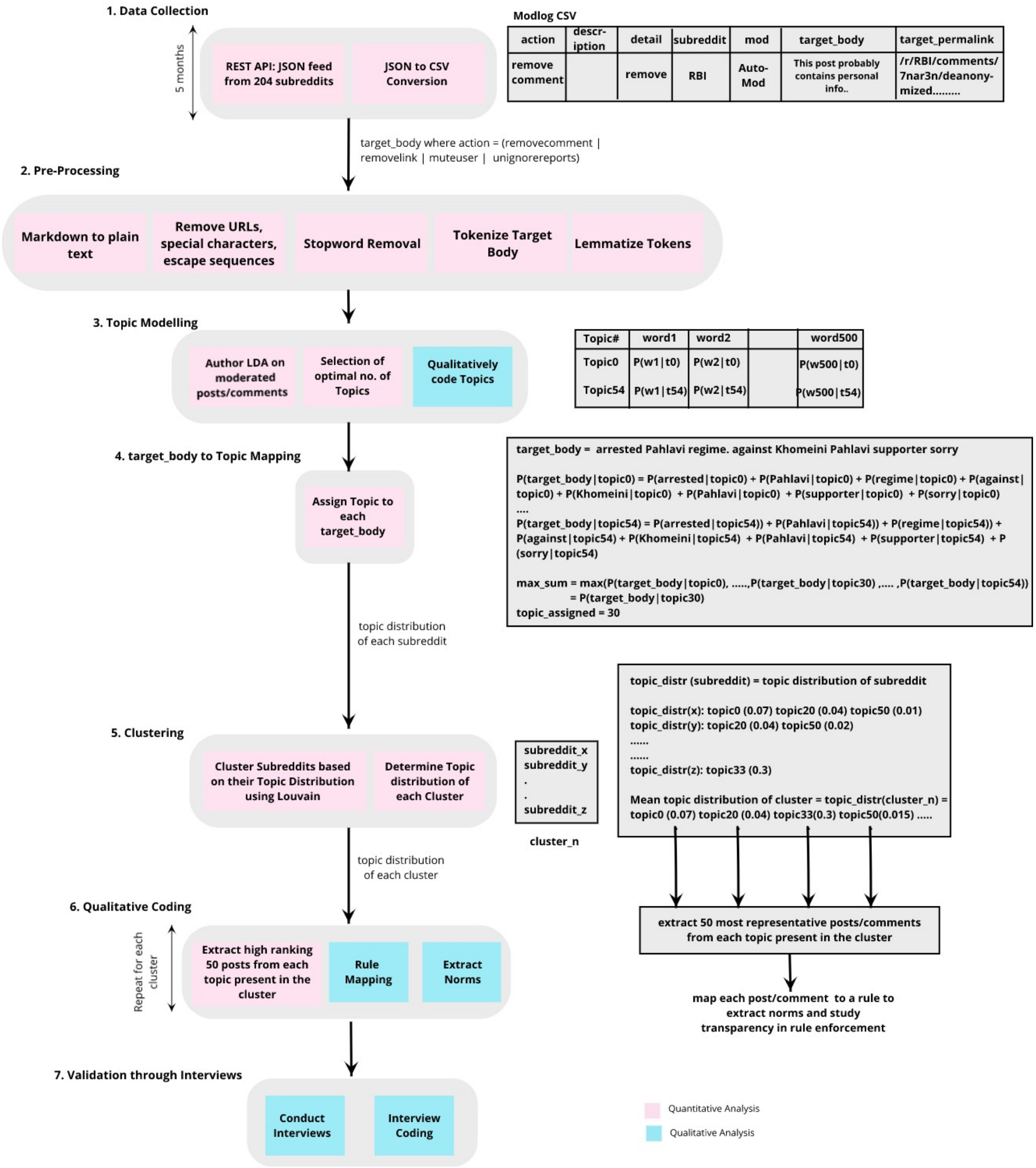
Rules, Norms and Policies: Deviant behavior is a serious and pervasive problem in online communities [11]. One way to tackle this kind of behavior in online platforms is by enforcing rules [31], norms [6] and content policies [5]. Policies and rules are ubiquitous through terms and conditions which are set either by the platform [17] or by the community [30] or collaboratively by both [4]. On the other hand, norms are emergent and develop from the interactions among users of the community [45]. For example, in a platform like Reddit that has a two tier regulatory structure, the company has a site wide Content Policy [51] that is enforced throughout the platform. In addition, each subreddit has its own formal rules (formulated by the moderators of those subreddits) and cultural norms that arise from the interactions among users and moderators. Past studies have focused on rules, norms and policies together [3, 6] and individually [16, 17, 37]. Lessig et

[3] https://ampersand.gseis.ucla.edu/ucla-is-presents-conference-on-all-things-in-moderation/
[4] University of California, Los Angeles —http://www.ucla.edu/

al. studied how policies shape the behavior of online communities and play a role in regulating them [37]. Fiesler et al. characterized subreddits' rules, studied the frequency of rule types across the subreddits and examined the relationship between rule types and subreddit characteristics [17]. Chandrasekharan et al. analyzed content that has been removed from various subreddits and extracted micro (violated in a single subreddit), meso (violated in a small group of subreddits) and macro (violated all over Reddit) norm and rule violations [6]. While the aforementioned work [6] studied rule and norms as loosely coupled entities, we consider the dichotomy between them. We use this distinction to propose a unique way of discovering community norms by mapping each sanctioned post/comment with site wide as well as subreddit level rules.

Content Moderation: It is not sufficient for social media platforms to craft rules and policies, they have to be enforced through moderation processes. There are a variety of ways in which content is moderated online. Some platforms (Facebook, Reddit, Twitter, YouTube) rely on the community to "flag" inappropriate and offensive content that violates the platform's rules and policies [9]. Others (Facebook and Twitter) employ commercial content moderators [53]. They hire large number of paid contractors and freelancers who police the platforms' content [8, 32]. Some platforms (Wikipedia, subreddits) rely on volunteer moderators instead of commercial moderators to moderate the content posted in their communities [12, 15, 66]. These moderators and their actions build and shape online communities [41]. Seering et al. examined the process to become a moderator and studied how moderators handle misbehavior and create rules for their subreddits. However, their study does not handle aspects of moderator feedback for community development which this paper will address in future sections. Moderator feedback is given not only by humans, but sometimes by automated tools as well. These tools assist the moderators in their day-to-day moderation tasks but their configuration is a black box to the users [28]. We explore this phenomenon of hidden AutoModerator configuration along with other practices followed by moderators to examine whether they adhere to SCP and Reddit's moderator guidelines.

Transparency in Moderation: The way content moderation is done by moderators in many online platforms is non transparent and murky with content disappearing without explanation [21]. Users find the lack of communication from the moderators very frustrating [27]. In some platforms, even the presence of moderators themselves is sometimes unknown [54]. This lack of transparency can confuse users and lead to high drop out rates [65]. Thus, in recent times, researchers [21, 23, 61] and the media [22, 24] have called for greater transparency in the moderation process followed by the platforms. But when we talk about transparency in online content moderation, the first problem is the lack of a formal definition of what transparency actually means [61]. Several researchers have tried to define it [18, 42]. For example, Ann Florini [18] defines transparency as *"the degree to which information is available to outsiders that enables them to have informed voice in decisions and/or to assess the decisions made by insiders"*. Researchers have also proposed several models and guidelines using which platforms can transparently govern themselves. One such guideline that has recently gained a lot of traction and is widely embraced by many social media platforms is SCP [49]. Another framework that has become popular in recent times is the 'fairness, accountability and transparency' (FAT) model. This model has especially been used to understand the algorithmic decision making process [34, 35, 52]. Many content moderation systems such as YouTube, Facebook use artificial intelligence techniques to carry out their moderation. However, these techniques are not interpretable and often questioned for being a black box. Transparency can have its pitfalls too. Providing too much information to the user can lead to inadvertent opacity —a situation where important piece of information gets concealed in the bulk of data made visible to the user [59]. It can also endanger privacy, suppress authentic conversations [56] and allow miscreants to game the system [13], a sentiment also shared by moderators during the interviews.


1. Data Collection
5 months
REST API: JSON feed from 204 subreddits
JSON to CSV Conversion
target_body where action = (removecomment | removelink | muteuser | unignorereports)
2. Pre-Processing
Markdown to plain text
Remove URLs, special characters, escape sequences
Stopword Removal
Tokenize Target Body
Lemmatize Tokens
3. Topic Modelling
Author LDA on moderated posts/comments
Selection of optimal no. of Topics
Qualitatively code Topics
4. target_body to Topic Mapping
Assign Topic to each target_body
topic distribution of each subreddit
5. Clustering
Cluster Subreddits based on their Topic Distribution using Louvain
Determine Topic distribution of each Cluster
topic distribution of each cluster
6. Qualitative Coding
Repeat for each cluster
Extract high ranking 50 posts from each topic present in the cluster
Rule Mapping
Extract Norms
7. Validation through Interviews
Conduct Interviews
Interview Coding
cluster_n
extract 50 most representative posts/comments from each topic present in the cluster
map each post/comment to a rule to extract norms and study transparency in rule enforcement
Quantitative Analysis
Qualitative Analysis

Modlog CSV

action	descr-iption	detail	subreddit	mod	target_body	target_permalink
remove comment		remove	RBI	Auto-Mod	This post probably contains personal info..	/r/RBI/comments/7nar3n/deanony-mized.........

Topic#	word1	word2		word500
Topic0	P(w1|t0)	P(w2|t0)		P(w500|t0)
Topic54	P(w1|t54)	P(w2|t54)		P(w500|t54)

target_body = arrested Pahlavi regime. against Khomeini Pahlavi supporter sorry

P(target_body|topic0) = P(arrested|topic0) + P(Pahlavi|topic0) + P(regime|topic0) + P(against|topic0) + P(Khomeini|topic0) + P(Pahlavi|topic0) + P(supporter|topic0) + P(sorry|topic0)
....
P(target_body|topic54) = P(arrested|topic54)) + P(Pahlavi|topic54)) + P(regime|topic54)) + P(against|topic54) + P(Khomeini|topic54) + P(Pahlavi|topic54) + P(supporter|topic54) + P(sorry|topic54)

max_sum = max(P(target_body|topic0), ...,P(target_body|topic30) ,.... ,P(target_body|topic54))
= P(target_body|topic30)
topic_assigned = 30

subreddit_x
subreddit_y
.
.
subreddit_z

topic_distr (subreddit) = topic distribution of subreddit

topic_distr(x): topic0 (0.07) topic20 (0.04) topic50 (0.01)
topic_distr(y): topic20 (0.04) topic50 (0.02)
......
......
topic_distr(z): topic33 (0.3)

Mean topic distribution of cluster = topic_distr(cluster_n) = topic0 (0.07) topic20 (0.04) topic33(0.3) topic50(0.015) .....

Fig. 1. Flowchart depicting methodology and data processing pipeline. Additional method details with respect to the Quantitative Analysis are outlined in Appendix A

4 DATASET

We use /u/publicmodlogs[5], a Reddit bot, to gain access to the public moderation logs from subreddits. Each subreddit that wants to provide public access to their moderator action logs (referred to as modlog in the rest of the paper) invites this bot to be a moderator with read-only permission. Once the bot accepts the invitation, it gains access to the subreddit's moderation logs and starts

[5] https://www.Reddit.com/user/publicmodlogs

Modaction	Description
removelink	Removal of a link
removecomment	Removal of a post/comment
muteuser	Mute a user from the moderator mail (ModMail)
ignorereports	Ignore all reports made for a particular post/comment
spamcomment	Content of post/comment is marked spam
spamlink	Link(s) present in a post/comment is marked spam

Table 1. Description of a sample of modactions present in our modlog dataset

generating a JSON feed containing those logs. This bot is one of the many third party initiatives that provide access to the moderation logs and Reddit has no say on why/how/when a subreddit can opt for it.

At any given time, the modlogs contain complete moderation action data from the previous 3 months. Using this bot, we have been continuously scraping data of 204 subreddits from March 2018 to September 2018. The collected logs have several fields capturing the moderation action (henceforth referred to as modaction) that was performed (`action`), who performed it (`mod`, `mod_id`), time of moderation (`created_tc`), short description for the action (`description`, such as *removecomment, removelink*, etc.), detailed explanation stating the reason for the modaction (`details`), community on which it was performed (`subreddit_name_prefixed`, e.g. r/conspiracy, sr_id, subreddit,), user whose contribution was moderated (`target_author`), content of the post or comment that was moderated (`target_body`), permanent static link of the moderated content (`permalink`) and title of the post (`title`). Note that the moderation action can be performed either on the post published by the user or comment made by other users on existing subreddit posts. Hence, we will refer to the moderated content (captured by the `target_body` field) as post/comment throughout the paper.

Our data contains 44 unique modactions. Table 1 presents a sample along with a brief explanation. Since our goal is to study norms and transparency in rule enforcement, we focus on modactions that explicitly represent removal of content. For this purpose, we shortlist four modactions—*removelink*, *removecomment*, *muteuser* and *unignorereports*—all of which represent removing either a post or comment by the moderator because they deemed it was unsuitable for that particular community. After filtering for these modactions, we had 479,618 rows in our modlog data-set. Our data also revealed that the 'description' field corresponding to these modactions was mostly blank. While some moderators provide detailed explanations behind their modactions, others do not. The 'details' field includes explanations as vague as *remove* to as detailed as *Section 1B-2 - Blacklisted Domains. Domain detected: Coincodex.com*. There is a lack of common standards when it comes to explaining and justifying modactions—a phenomenon which this study investigates.

Figure 1 shows the methodology flow chart and data processing pipeline used in our project to answer the research questions. It consists of five stages that involve both machine computation and human qualitative coding. We describe them in detail in the following sections.

5 RQ1: HOW DO CONTENT MODERATION PRACTICES ON SUBREDDITS ALIGN WITH SCP'S PRINCIPLES OF TRANSPARENCY

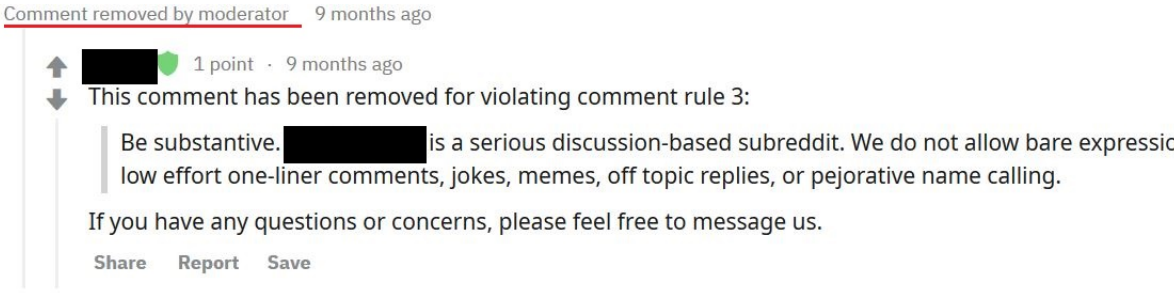
In order to study how rules and norms are operationalized and enforced, it is essential to study *why* content gets removed from the platform. The reasons could range anywhere between violation of rules and norms to deviation from site wide Content Policies. We make use of both quantitative and qualitative methods to determine these reasons. The quantitative part aims to find the "meta communities"—groups of subreddits—that share the same types of transgressions. We use qualitative methods to analyze the transgressions in each of the meta communities in order to determine the reason for removal. The detailed methodology as well as the results are discussed in sections below.

Meta Community	Sample subreddits	Mean subscribers	Freq. of Violations
Gaming, erotic & political meta community (MC1)	ElderScrolls, AVGN, POTUSWatch, SugarBaby, CuckoldPregnancy, moderatepolitics	62248.48	🟥 17.69% 🟦 49.45%
Pro free speech & anti-censorship meta community (MC2)	dark_humor, BadRedditNoDonut, AllModsAreBastards	6720.7	🟥 64.90% 🟦 77.07%
Cryptocurrency, news & special interest meta community (MC3)	CryptoCurrency, ethereum, conspiracy, MakingaMurderer	63305.9	🟥 23.78% 🟦 41.09%
Conspiratorial (MC4)	ConspiracyII	15000	🟥 28.25% 🟦 69.30%
Academic (MC5)	Indian_Academia	2100	🟥 62.50% 🟦 68.75%

Table 2. Meta-communities along with a few constituting subreddits. Mean subscribers stands for mean subscriber count per meta-community. Note that the last two meta communities consist of only one subreddit. 🟥 denotes the percentage of sanctioned posts that coders could not map to a rule. 🟦 denotes the percentage of sanctioned posts where moderators do not provide an explanation for content removal.

5.1 Quantitative Method: Find Meta-Communities Sanctioning Similar Content

The aim of the quantitative method is to determine "meta communities" that sanction similar content. We posit that communities that remove similar content will have similar rules and norms. To find the types of sanctioned content that were removed from the subreddits, we empirically find topics in the modlog dataset using the Author Topic Modelling (ATM) [55] algorithm—an extension of LDA (Latent Dirichlet Allocation) [1], a widely used topic modeling technique. Since topic modelling algorithms are highly susceptible to noisy data, robust pre-processing of the dataset is necessary in order to obtain interpretable topics. We applied standard text pre-processing techniques (see Appendix A) on the sanctioned content and fed it to the ATM algorithm. ATM extracts common themes and topics from documents and groups them by their authors. For this study, we considered sanctioned posts/comments as the documents and subreddits in which they were posted as the authors. By using ATM, we obtained 55 topics that represent the types of transgressions sanctioned across various subreddits. Appendix A.0.2 provides our empirical process behind choosing 55 topics. To interpret the common themes spanning these topics, two authors independently coded each of the topics by examining the top 25 posts and comments. Many of these topics related to spam posts. For example, we found 10 topics that were dominated by spam. Top posts representing these spam topics promoted and highlighted certain websites, cryptocurrencies and social issues. The remaining topics discussed conspiracy theories, controversial political and geopolitical themes, erotic lifestyles, and miscellaneous content. Appendix B lists the 55 extracted topics and details of our qualitative investigation and interpretation of these topics. Finally, to find groups of communities that sanction similar content, we applied Louvain Community Detection [2] algorithm to cluster the subreddits spanning the 55 topics discovered by ATM. Louvain's algorithm is a popular method used to detect communities from large networks. This algorithm has also been used on various Reddit datasets that deal with inter-subreddit relationships [10, 46]. After applying Louvain, we obtained 5 clusters. Each cluster represents a "meta community" that sanctions similar kinds of transgressions. Table 2 enlists all the discovered meta communities along with a few selected subreddits. We allocated each meta community a shorthand name to improve readability. We present descriptions of each meta community along with all of its constituting subreddits in Appendix A.0.4



Comment removed by moderator 9 months ago

[redacted] 1 point · 9 months ago

This comment has been removed for violating comment rule 3:

> Be substantive. [redacted] is a serious discussion-based subreddit. We do not allow bare expressio
> low effort one-liner comments, jokes, memes, off topic replies, or pejorative name calling.

If you have any questions or concerns, please feel free to message us.

Fig. 2. Example of an instance where human moderator specified the reason for removing a comment

5.2 Qualitative Methods: Investigating adherence to SCP through Rule Mapping and Norm Discovery

In order to study moderation practices, rules and their enforcement in each of the 5 "meta communities", we conducted a deep content analysis of 50 high ranking posts and comments from each of the topics present in the 5 cohorts. High ranking posts/comments for a topic in a meta community are the ones that have high probability of belonging to that topic; hence are the most representative of that topic. Note that in some meta communities, the probability of occurrence of few topics was very low. Thus, we did not get 50 posts from those topics to annotate. Finally, we analyzed a total of 6260 posts/comments. Two authors qualitatively mapped these posts/comments to subreddits' rules and Reddit's Content Policy. The absence of a rule results in the discovery of a norm that has not been formalized as a rule. In the process, we also study how transparently rules are enforced by moderators in the communities. Contrary to popular belief, content is not the only reason why posts/comments are sanctioned. They are sanctioned due to a variety of other reasons. Minimum karma and account age, title and format of the post and history of the user's rule violations can all result in sanctions. Posts/comments can also be removed if their theme clashes with the ideology of the subreddit. For example, Pro Trump comments are removed from subreddit Socialism. Socialism is a subreddit that is dedicated to discussing current events from a socialist/anti-capitalist perspective. Since there is a clash in ideology, Socialism disallows any posts supporting capitalism. The views of Donald Trump and the Republican party are the exact opposite of those of socialism. Therefore, the subreddit disallows any posts supporting Trump. This indicates that understanding the context surrounding the post/comment is essential to discover norms and study transparency in operationalizing rules.

To understand the context in which a comment was removed, we relied on the 'target_permalink' field present in the modlog. This field provides us with either the link to the sanctioned post or the link to the post to which the comment was posted by a user. At this link, even though we are unable to see the actual sanctioned post/comment, we have access to all the other surrounding content and the overall discussion. Removed posts are represented by the placeholder text *[removed]* and removed comments are represented by the placeholder text *"comment deleted by moderator"*. In some cases an explanation provided by the moderator (human moderator or AutoModerator) is present in the form of a comment after this placeholder as shown in Figure 2. We relied on this comment, the 'details' field present in the modlog and the context in which the sanctioned comment was posted to determine why the content was removed.

The mapping task was performed by two authors independently in an iterative manner. Both authors are passive users of Reddit and have been on the platform for seven and fifteen months respectively. They mapped each post/comment, keeping in mind the context and subreddit in which it was posted, with the subreddit's rules and Reddit's content policy. They then separated

Transparency guidelines	Transparency violations
Provide detailed guidance to the community about what content is prohibited	Secret guidelines in the form of community norms (RQ1a and RQ1c) 23% of the annotated posts were either ambiguous removals or norm violations. For example, moderators of meta community 1 remove derisive posts/comments.
Provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension	Lack of explanations and reasons for content removal (RQ1b) ~50% of annotated posts were not accompanied by reason for removal.
Provide an explanation of how automated detection is used across each category of content.	Hidden AutoModerator configuration (RQ1d) Subreddits do not reveal banned words, projects and karma requirement. For example, subreddit MurderedByWords removes posts/comments containing the word feminazi.

Table 3. Table summarizing SCP transparency guidelines and corresponding moderation practices that violate these guidelines. 🟧 represents the remarks and examples corresponding to the violations.

posts/comments that were removed because of a norm or a reason that was not clearly reflected in rules. Finally, the authors came together and discussed the codes. Disagreements were resolved by discussions and by re-iterating over codes. Through the qualitative analysis, we discovered hidden norms and transparency violations in several moderation practices. We present our findings in the next section.

5.3 RQ1 Results

Our qualitative analysis revealed several moderation practices across meta communities that violate the SCP guidelines. We discuss these practices in detail in the subsequent sections. Table 3 summarizes the SCP guidelines and corresponding violations briefly.

5.4 RQ1a: Do all sanctioned posts correspond to a rule that they have violated that is either explicitly stated in the community's rule book or Reddit's Content Policy?

Content on Reddit can be removed due to a violation of the site-wide Content Policy, subreddit's rules or norms. During the qualitative analysis, we mapped each sanctioned post/comment to the subreddit's rules and Reddit's Content Policy. In the process, we discovered norms as well as several ambiguous removals. Ambiguous removals are posts/comments that seemed rule-abiding, but annotators couldn't determine any rational reason for their removal. Out of the 6, 260 posts annotated, there were 3, 030 sanctions with no explanations for removal. Out of these, 965 (31.8%) instances were removed due to unstated norm violations and 454 (15%) instances were ambiguous removals. We discuss community norms in detail in Section 5.6. We argue that it would be equally difficult for the user who posted this content as well as a newcomer in the community to understand what kind of content is sanctioned from the community. Thus, it becomes essential that moderators specify the reasons for removals.

Figure 2 shows percentage of (unique) sanctioned posts/comments where authors could not map content to a rule (norms+ambiguous removals) for each of the five meta communities. The Pro free speech & anti-censorship meta community (MC2) and Academic meta community (MC5) have the highest percentage of such removals. Subreddits belonging to MC2 were mostly pro free speech communities that had zero to a handful of rules. Similarly, MC5 consists of one subreddit that had only one rule specified at the time of qualitative coding. Thus, most of its sanctions were either coded as norms or were ambiguous removals since coders could not map them to any site-wide or subreddit specific rules.

5.5 RQ1b: Do moderators provide reason for removals after removing user's content?

The second principle of the SCP (Notice) states that users should be informed *why* their content was removed from a social media platform. Recall that 3030 annotated posts/comments did not provide a reason for removal. This shows that the SCP Notice guideline is violated in many subreddits. Figure 2 shows percentage of sanctioned posts/comments where moderators do not provide an explanation for content removal for each of the five meta communities. The Pro free speech & anti-censorship meta community (MC2) and Academic meta community (MC5) again have a large amount of such removals indicating that moderators of these subreddits sanction several posts/comments because of norm violations but seldom provide any feedback. The Gaming, erotic and political meta community (MC1) also has significant number of sanctions without explanations. This meta community has several subreddits that provide a platform for healthy discussions and debates which often turn into flame wars. As a reactive measure, moderators of these subreddits nuke and lock threads containing such heated arguments without providing any explanations for content deletions. As a result of this action, several rule-abiding posts/comments are sacrificed increasing the count of both ambiguous removals as well as content without feedback.

Out of the 3030 posts that were sanctioned without providing a reason, human moderators and AutoModerator moderated 68.8% and 31.2% of the posts respectively. In the current Reddit design, moderators do not have a feature where they can provide feedback about content removal. Thus, they either have to manually provide feedback as a "reply" to the deleted post or send a personal message to the user. We discuss this limitation and design implications in Section 7.

5.6 RQ1c: Are rules in subreddits clearly defined and elucidated? Are they enforced and operationalized in a consistent and transparent manner by human and auto moderators?

Our qualitative analysis reveals several practices where rules were either not clearly elucidated or their enforcement was not transparent.

Hidden word filters: Manually reviewing every post/comment is laborious. The Gaming, erotic and political meta community (MC1) and the Cryptocurrency, news and special interest meta community (MC3) are high traffic meta communities with mean subscriber counts 62248.48 and 63305.9 respectively. Thus, moderators of these communities rely on AutoModerator to identify troublesome posts. They configure word filters using custom regexes to detect profanities and other offensive content. However, the exact details of the word filters are not revealed to the user. These offensive words could be pejorative terms for females (*feminazi*, *fucking bitch*, *pussies*) or males (*cuck*), abusive slangs and phrases (*fat fucks*, *shit lord*, *overweight*, *prick*, *shut the fuck up*, *fucking mouth*, *jack off*), disrespectful terms for homosexuals (*fags*) and racist slurs (*nigger*, *negro*, *nig*, *porch monkey*). Words and slangs like *jew*, *hymies* and *kike* are used to filter out posts/comments about Jews in an attempt to prevent antisemitism. In addition to offensive content, AutoModerator is also used to detect low quality posts by either setting a word count threshold or a word filter. For example, phrases like *shill me*, *rate my portfolio*, *censor*, *shill*, *and arbitrage* and *beer markets* are often removed from cryptocurrency related subreddits (MC3) as they are generally associated with low-quality content. All these banned and blacklisted words and slangs haven't been made public in any of the rules.

It is important to note that even after a violation, an offender seldom learns which particular word caused his content to be removed. For example, subreddit POTUSWatch does not disclose the banned word because of which the following comment was removed: *"This has dominated the 24 hours new cycle. I think it was done so Donny Two Scoops could give Melania a day in the limelight. You know, so he could transfer a little bit of heat from him to her."*. We suspect that the comment was sanctioned because of the presence of the slang *two scoops*. However, there was no way to confirm this since the "detail" field in the modlog contained the phrase *"word filter"* and no message explaining the reason for removal was found with the original comment.

Some comments which used the banned content innocently were also caught by these filters and removed. In some subreddits, these posts were re-reviewed by human moderators and were brought back while in others they were not. For example, consider the following two posts (1) *"...what heritage is that about? German, British, Dutch, Swedish, Italian, Spanish, Portuguese, Polish, European **Jewish**, Gypsy? Those cultures and heritages have nothing in common. 'White pride' is to those cultures and heritages the same thing an 'Asian restaurant' is to the cuisine of Asian nations: great by themselves...."* (2) *"What is rule 5? I mean **bitch** in an inadequately masculine gangster rap way, and I mean **fag** in a South Park contemptible person way."* The first post was removed because of the presence of the word *Jew* but was later brought back by the moderators. On the other hand, the second post was removed permanently even though words *bitch* and *fag* were used in a non-derogatory way. This practice indicates that posts removed by AutoModerator are not thoroughly reviewed by the human moderators. We plan to investigate this line of thought in future work.

Non-exhaustive rules: A few subreddits in the Pro free-speech and anti-censorship meta community (MC2) and the Cryptocurrency, news and special interests meta community (MC3) provide a list of blacklisted words and slurs but this list is by no means exhaustive. For example, the subreddit Socialism (present in MC3) is a community for socialists with a robust set of rules. The list of prohibited words have been publicly listed as a part of one of their rules. It includes both explicitly abusive terms (e.g. retarded, bitch, homo, etc.) as well as normalized abusive terms (e.g., stupid, crazy, bitching, etc.). However, this list of examples provided by the moderators is not comprehensive. For example, we suspect the following comment was removed because of the presence of word *brainless* which is not present in the subreddit's blacklisted word set. *"...removed a word that the brainless automoderated bot mistook for an insult since it's not able to read context proving that it's a dangerous and un-bright idea to automate censorship."*

Banned Topics: A few cryptocurrency related subreddits in MC3 do not permit the use of certain projects and platforms. For example, subreddit CryptoCurrency has banned coin projects and platforms like *VeChain, Skycoin, Foresting, Maecenas* etc. It also removes posts that mention any of the pump-and-dump[6] (PnD) cryptocurrency groups. We found no explicit rules providing an exhaustive list of such projects. We also found posts that were not specifically discussing these projects being removed. For example, the post *"I believe Teeka Tiwari prediction of $40k by end of year. Dude has been accurate the last 3 years"* was removed because of the use of PnD group *Teeka*, but the post talks about the financial advisor "Teeka Tiwari". Neither was the user notified about the removal, nor did the moderators provide any feedback. This example clearly shows that AutoModerator cannot determine the context in which a banned word is used. Thus, the involvement of human moderators is critical in content moderation.

Hidden karma and account age requirement: Karma is a very important concept in Reddit and was introduced to curb spam, off-topic and low quality content. Members with higher karma points have certain privileges that others don't. In almost all the subreddits present across MC1, MC3 and MC4, a minimum comment karma is required to post content, a rule enforced by AutoModerator. Surprisingly, many subreddits have not mentioned this practice in their rules. For example, subreddits KotakuInAction, Cuckold, Vinesauce and btc (MC3) filter posts using the karma filter but none of them mention this requirement in their rules. Subreddit Sugarbaby (MC1) requires accounts to be at least 7 days old before being allowed to create new posts, but this requirement has not

[6] https://en.wikipedia.org/wiki/Pump_and_dump

	Subreddit	Example post/comment removed	Remarks (*/**)
MC 1	POTUSWatch	*The investigation is over. It's Obama's fault.*	short comment**
	NeutralCryptoTalk	*As someone who doesn't truly understand the economics of the crypto market, can anyone inform me as whether this is a dip/crash or when/will it recover? Is there correlation with the traditional stock market? ... Any insight would be much appreciated*	noob comment*
MC 2	Ninjago	*PSA: Regarding the recent netflix issues Don't have Hulu :(*	off topic*
	dark_humor	*[Dunking on Helpless Boy](https://m.youtube.com/watch?v=yuIDC3cq6fU)*	ableist jokes*
MC 3	Socialism	*eu blandit Ut consectetur bibendum volutpat. amet non neque nisi adipiscing dolor amet, tempus Vestibulum consequat sit venenatis Aenean leo Sed a amet quis eros condimentum. in tempor sapien sit euismod*	non-english comment*
	RedditCensors	*Still absolutely no acknowledgement of the fact that u/*** is blatantly just banning anyone trying to discuss the fact that her friends get special treatment and anyone questioning authority is banned.*	attacking mods*
MC 4	ConspiracyII	*Please, do elaborate.. This sub you're in, harping on about my joke of a /u/; you do know that I and some pals created all of this, right? Do you feel silly now? You should.Worry not, I'll be abandoning this username soon. I'm tired of hearing nonsense from fucking delicate idiots with misunderstood shill hysteria. What is /u/AllThat? Are you a movie for tweens? How about I change my shit to /u/Hannibal? Am I now a serial killer? Come back when you have an apology at the ready and you pick your head up from the dirt, dumbass.*	lead by example*
	ConspiracyII	*Damn, very interesting information, thank you now I'm going to do a little digging myself.*	phatic talk*
MC 5	Indian_Academia	*My gawd, you Pretentious bastard.*	derisive content*
	Indian_Academia	*Hi Indu, I have been scavenging the internet to find info regarding the same. I found this video on YouTube and found it useful. Do check it out: http://bit.ly/2FAuT73 All the Best!*	URLs with incorrect format*

Table 4. Sample posts/comments from all meta communities that were removed due to norm violations. Remarks column contains the rationale for removal of the content. Reasons marked * are removed by human moderators and the ones marked ** are removed by the AutoModerator.

been explicitly specified anywhere. Even after a violation, the exact karma requirement was not revealed by the moderators. For example, subreddit MurderedByWords (MC1) removes posts by new members, leaving a message *"Your comment has been removed due to your account having low karma. This is done to combat spam. If you would like to participate on /r/MurderedByWords, you will need to earn karma on other subreddits."* The message does not reveal the minimum karma required. In some subreddits, comments with negative karma (number of downvotes > number of upvotes) are also removed. Again, neither are the users notified about the removal nor is this requirement specified in the subreddit's rules.

5.7 RQ1d: What are the norms prevalent in various subreddit communities?

In this section, we present the norms that we uncovered in the meta communities. Table 4 presents sample posts/comments from all the five meta communities that were removed because of norm violations.

Threads containing flame wars nuked (MC1): Discussions on controversial topics like politics and important public figures often turn into heated personal arguments in MC1. It is a common practice followed by a few subreddits in this meta community to delete entire threads containing flame wars and slap fights, often sacrificing a few innocent comments. Other subreddits take a step further and lock the threads stopping community members to participate in the discussion.

Threads containing arguments nuked (MC4): The moderators of this meta community nuke entire threads containing arguments. As a result several rule abiding (valid and civil) comments were removed too. For example, a user's civil comment *"If it's upsetting to you please, appeal"* was removed when a thread was nuked. The users are neither notified about the removal nor do the moderators provide reasons for such removals.

Short posts and comments removed (MC1): Posts/comments with low character/word count are considered low quality and are removed. Few of the subreddits have configured AutoModerator to perform this task. Neither this practice nor the word/character count used to perform the filtering has been made public in any of the rules.

Derisive content removed (MC1, MC5): Sarcasm and snide remarks are not appreciated in meta communities 1 and 5. Few subreddits had an explicit rule, while in others, it's a practice. Moderators in these meta communities believe that being snarky doesn't contribute much in the discussion.

Content without corroboration removed (MC1): Arguments containing conspiracy theories and posts/comments without substantiation from trustworthy sources are often removed from meta community 1.

Noob comments and posts removed (MC1): In some subreddits, basic questions (that one could have searched online) asked by new users are removed. There are no rules specifying this practice. Table 4 contains an example of such a post.

Shadow bans (MC1): Some comments in the content we analyzed were shadow banned. When a user is shadow banned, his posted content is only visible to the user but not to the community.

Content posted by repeat offenders and trolls removed (MC1): Some subreddits in this meta community have a list of Redditors who are known trolls or repeat offenders in their communities. Content posted by such users are put in the moderator queue by the AutoModerator and are reviewed by human moderators. We found instances where posts (not violating any rule) posted by repeat offenders were removed and no reason for removal was provided.

Posts with disguised links, banned domains and incorrect format removed (MC1, MC3, MC5): A few moderators expect the posts published on their subreddits to follow a certain format. They also disallow use of certain domains and shortened URLs. Most of the times human moderators rely on AutoModerator to perform this task. For example, moderators of subreddit RBI filter posts containing Facebook links. Subreddit SugarBaby requires the post to be in a certain format *([age] [online] or [2 character state abbreviation for irl and online] catchy header)]*. But this requirement has not been specified in the sidebar rules. A newcomer can learn this norm either after a violation or by observing the subreddit for a certain period of time.

Sexist and ableist jokes removed (MC2): Subreddit i_irl, a meme aggregator, allows people to cross post a variety of memes from all over Reddit. It asks users to submit *"dark jokes and memes without white knighting and general faggotry"*. But, the subreddit removes jokes with hints of ableism and sexual connotation.

Off topic content removed (MC2): Off topic content is removed from almost every subreddit under our scrutiny. Many communities all over Reddit have an explicit rule stating this practice. But, we found a handful of growing communities (with less number of subscribers) in this meta community with no rules, where the practice of removing irrelevant content is followed. There are also some communities where description of the subreddit is not enough to know what kind of content is welcomed by the moderators and members. For example, a TV show based subreddit removed the post where the user complains that he/she can't see the series online [*"Post: PSA: Regarding the recent Netflix issues Don't have Hulu :("*].

Attacking Mod (MC3): Name calling the moderators and protesting against censorship is not encouraged. We found several posts attacking the authority getting removed from this meta community by the moderators.

Non-English posts (MC3): Most of the subreddits in this meta community 3 have a norm to remove non-English posts.

Unexplained removal of posts containing Facebook links (MC4): Posts containing Facebook links were removed by AutoModerator. None of the sanctioned posts were accompanied by reason for removal.

Phatic Talk (MC4): Several comments that add little value to the conversation were removed by the AutoModerator. However, we were unable to identify the exact basis of classification of content as phatic and low quality. The AutoModerator neither leaves the reason for removal as a reply to the deleted comment nor did we find any accompanying description in the modlog dataset.

Lead by Example: Moderators posting and deleting comments that violate subreddit's rules (MC4): We found moderators removing their own comments. These comments were demeaning in nature and clearly violated the rules of the subreddit. For example, consider the comment *"you do know that I and some pals created all of this, right? Do you feel silly now? You should. Worry not, I'll be abandoning this username soon. I'm tired of hearing nonsense from fucking delicate idiots with misunderstood shill hysteria. What is All That? Are you a movie for tweens? How about I change my shit to Hannibal? Am I now a serial killer? Come back when you have an apology at the ready and you pick your head up from the dirt, dumbass."*

The findings from our qualitative coding raised several questions about the widespread lack of transparency in subreddit communities. Several aspects of transparency are violated. For example, posts and comments are silently removed without notifying the user about the reason for the removal. Rules' enforcement is hidden from the community in most cases. We summarize the SCP guidelines and the corresponding violations in Table 3.

6 RQ2: HOW DO MODERATORS VIEW THE VARIOUS FACETS OF TRANSPARENCY?

By delving into the moderators' side of the story, RQ2 helps in understanding the reasons behind the widespread transparency violations.

6.1 Method: Interview with moderators

We conducted semi-structured interviews with 13 Reddit moderators between January, 2019 and May, 2019. 11 of them are moderating the communities present in our public modlog dataset. A semi structured interview script was designed to understand the rationale behind the widespread transparency violations summarized in Table 3. The script was revised multiple times based on the feedback received from four researchers. In order to get additional insights about moderation practices, we first asked moderators about their background, moderation process that they followed in their respective subreddits as well as the rules and norms of those subreddits. We probed them about the design of certain rules. We inquired about the use of AutoModerator and the appeal process against content removals and bans. We also discussed how a user is notified about a rule violation. Finally, we asked them the ways Reddit can help them in terms of policy changes, interface

Interviewee	Subreddit Topic	Moderation Experience (in months)	Gender	Country
P1*	censorship	12	M	USA
P2*	archive	84	M	USA
P3	school admissions	12	M	USA
P4*	entertainment, niche interests	156	M	USA
P5	news sharing	48	F	USA
P6*	memes	20	F	FRA
P7*	country-related	24	M	AUS
P8*	cryptocurrency-related	72	M	USA
P9*	health	60	M	USA
P10*	sexual fetish	12	M	USA
P11*	data & analytics, technology	48	M	USA
P12*	memes	53	F	USA
P13*	tv show	36	F	USA

Table 5. Moderators' characteristics. A * in the *Interviewee* column indicates that the moderator is moderating a subreddit present in our public modlog dataset. We refrain from specifying the subreddit names in order to protect the identity of the moderators. Instead we present high-level community topics that describe the subreddits.

and tools to make their job easier and effective. It is important to note that to elicit more detailed responses, we did not include direct questions about "transparency" and "SCP". The transparency theme automatically emerged from moderator's detailed narratives in response to the following interview questions: *Do you have a list of words whose use is prohibited in the subreddit? Have you made it public, i.e specified it as a part of any of the rules, why/why not? How does a user get notified that his comment has been deleted or he has been banned? How does the user learn why his content was deleted? etc.* Appendix C lists the complete interview protocol.

We adopted convenience sampling to recruit our subjects. We posted recruitment messages on several subreddits that accept surveys, polls, and interview calls, such as r/SampleSize and r/Favors. We also posted these messages on subreddits that are run specifically for moderators, for example, r/modhelp, r/modclub, and r/AskModerators. We also sent out personal messages to moderators who are moderating subreddits that are part of our modlog dataset.

The interviews lasted between 30 to 180 minutes with variance according to medium of interview and how active and how strictly moderated the community is. Interviews conducted through text based chats took longer time and moderators governing highly active subreddits with stricter moderation principles had more to say. It also depended on how many subreddits a moderator was managing. Some moderators brought insights from multiple subreddits. The minimum subscriber count of a subreddit moderated by our interviewees was 79 and the maximum subscriber count was 21.3 million as of May, 2019. Table 5 summarizes our study participants, the nature of the communities they moderate, length of their moderation experience, gender and country where they belong. In order to keep the privacy of moderators, we do not reveal the names of their corresponding subreddits. We also use the term *subreddit_name* whenever a moderator refers to his subreddit in any of his quotes. All interviewees received $15 as compensation for their participation. We used three modes to conduct the interviews according to the interviewee's preference - audio/video, chat and email. The audio/video interviews were recorded using the inbuilt features on Skype/Zoom or the recording features on our mobile phones. After transcribing the interviews, the authors manually coded them to determine common themes. All conflicts were resolved through discussions. Once the codes were agreed-upon by both the authors, we performed axial coding to deduce relationships among themes.

6.2 RQ2 Results

Interviewing moderators helped us unpack the various facets of transparency, including silent removal of content, removals without specifying reasons, enforcement and operationalization of rules, and the appeal process.

6.2.1 Silent Removals: No notifications about post/comment removals. After talking to moderators, we discovered that Reddit, as a platform, does not notify users when their comment gets deleted. Moderators told us that users can figure out the deletion once they either reload the page containing the deleted content (P5), log out and log in to Reddit again (P2) or check the public modlogs (P4, P13). They informed us that it is up to the human moderators to either send user a private message or leave a comment as a "Reply" to the deleted content detailing the reason for removal. Configuring the AutoModerator to notify a user about content removal and stating the reason for that removal is left to the discretion of the human moderators of that subreddit. User bans, on the other hand, are always issued with a message that includes a link to ModMail—a messaging system used to communicate with the moderator team—along with instructions about the appeal process.

> *"Comment removals don't generate a removal notice in any of the subreddits. Bans always are issued with a message that includes a link to ModMail and instructions. " - P10*

> *"I don't think people can get customized notifications for comments until they configure the auto moderator system. " - P10*

> *"Reddit doesn't even have a system in place for notifying either. They (subreddit members) notice it (post/comment) disappeared if they log out and don't see the comment. Bans are very different. Notifications of a ban are built into reddit. the exact details/reasons are for us to provide in that process. " - P2*

Moderators did not have a consensus on silent removals. An equal number of moderators in our dataset (n=5) were for and against (n=5) silent removals. A handful of moderators (n=3) believed that silent removals must be applied depending on the situation.

Few moderators shared that users would be irked by multiple notifications and private messages.

> *"People would hate getting that many PMs (private messages) about removed comments, most of Reddit operates the way we do." - P12*

> *" And I don't really think that (notifications) would be very necessary, right, because people leave hundreds of comments a day, So it kind of makes Reddit a very bloated system." - P10*

One moderator shared that explicitly pointing out removals could make users belligerent towards that moderator and could lead to users exhibiting more bad behavior.

> *"So I find that silent removal, um, both visually and policy wise is the best course of action. If you negatively reinforce people, they tend to be more pigheaded about it. They tend to be more stubborn and they will out of spite, try and continue this negative behavior." - P11*

Moderators who opposed silent removals claimed that this practice is pro censorship and can decrease the willingness of users to participate in the subreddits. For example,

> *"We message that user and tell them, Hey, we had to remove this. Here is why, or we removed the comment. We have to do that because the subreddit is devoted to censorship and pointing that out." - P1*

Few moderators were on the fence about notifying users. One moderator indicated that he notified users whenever he could as the process was labor-intensive.

> *"I notify people like here is exactly why you're being punished. But other times I'm just completely silent and just let them be confused. I'm generally not very consistent cause I'm just like one person, you know. So if I had to write down a special message for every single person, I banned it would take up a lot of time. " - P10*

Another point of view that emerged was that silent removals were suitable in certain cases. For example, one moderator shared that he selectively notifies law abiding users.

> *"I think it depends if it's somebody who's clearly like a participant in the community, and maybe they, like, break the rules once or twice. Then I think it's very appropriate to say, Hey, I removed your comment with a public post. Hey, I removed your comment for breaking this rule. Um, if it's things like cleaning up spam links, I don't really feel any obligation, to reply to all of those and say I removed it for being spam." - P8*

In summary, while the practice of silently removing content is in direct violation of the SCP transparency guidelines, which ask social media platforms to notify users when their content is taken down from the platform, interviews revealed the other side of the picture. Lack of proper infrastructure coupled with millions of users leaving millions of comments makes it impossible for moderators to notify each and every user about the content removal. As a result, few moderators start providing feedback selectively to rule-abiding participants.

6.2.2 Unexplained Censorship: Removing posts/comments without specifying a rule/reason. While not providing notification to users about content removal is one transparency violation, not including specific reasons for removals in those notifications is another violation. The SCP clearly states that the minimum level of information required for a notice to be considered adequate includes the *"specific clause of the guidelines that the content was found to violate"* [49]. During the qualitative analysis, we discovered that moderators seldom leave reasons or point to community rules while removing content. Our interviews showed that moderators' stance on unexplained censorship was divided.

Five moderators believed that specifying a rule/reason while removing content is helpful for the community. They pointed out that informing users about the rules they have previously broken serve as a learning opportunity to avoid such behavior in the future.

> *"We remove and we inform userbase by leaving a comment on why it was removed, it sorts of educates other users who go through the comments that this is what the moderators remove." - P7*

> *"We always post a comment specifying which rule was broken and distinguish+sticky it after we're done with the removal. It is incredibly helpful. Other communities just remove posts 'learned the lesson or not'. By showing our users which rule they've exactly broken, they can learn from it, and avoid specific behavior next time." - P6*

One moderator stated that explanations will improve the image of the moderators and the subreddit.

> *"I think it's helpful, and it makes us look better, right." - P3*

Another moderator recounted that his past experience as a Reddit user shaped his actions as a moderator. Silent removals without reasons decreased his willingness to participate in the subreddits. Thus, as a moderator he notified users of his subreddit about content removal and also specified the rules that they had violated.

> *"Well, those kind of practices, they extremely decreased my willingness to participate in those subs. Now that I found out.... like I didn't even realize it was happening because you never know your content is being removed. So after I figured I just stopped participating in subreddits that silently remove content, which is a lot of them." - P9*

Five moderators were on the other end of the spectrum and believed that specifying a rule or reason for removal is not helpful for the community. Some believed that people who sincerely exchange ideas do not need to be reminded of the rules.

> *"When dealing with these people, that value their account, and sincerely want to exchange ideas... people that are of value to us...they don't need to be told to act civilized...the rules lawyers that demand a list of forbidden words to avoid are there to use our rules as a game" - P2*

Other moderators shared similar sentiments about posting rules. They reported that miscreants will not change their behavior and will continue to break subreddit's rules despite transparent moderation practices.

> *"Because the people who are going to violate the rules, they're going to violate the rules, no matter what you tell them. And those who really are making a mistake and are in good faith are going to appeal. Moderators explain why the message was removed very rarely because the type of comments we remove are only those which are bigoted and that almost ninety nine percent of the time comes with an instant permanent ban" - P4*

One moderator informed us that most of the traffic on Reddit is coming from the smartphone app. While the sidebar is host to subreddit rules on Reddit's desktop client, it is not visible on Reddit's mobile application. Thus, he did not see any value in mentioning the violated rules.

> *"Well... funny thing about the sidebar... 80% of the traffic never sees the sidebar (rules) anyway. mobile users don't" - P2*

Another participant (P5) told us that negative reinforcement, sometimes, makes the user aggressive which leads to arguments with the moderator and further removals of the user's content. At other times, community members supporting the moderator's decision also start down-voting an offender's other posted content.

> *"A comment removal reasons are generally unhelpful because they would be posted in the same thread as the removed comment, not sent separately. I've modded one sub where some of the mods posted comment removal comments. It never worked out well. Either the user was belligerent about having one comment removed and proceeded to fight with the removing mod leading to more removed comments. Or other users agree with the mod and then find the person with the removed comments' other submissions and downvote them." - P5*

Both P5 and P10 asserted that it is too much of a manual effort to provide reasons for each and every removal since hundreds of posts and comments get posted on the subreddits. Similarly P11 also shared that he does not find writing reasons worthwhile since majority of posts removed from his subreddit are spam.

> *"Mods are free to send an individual removal PM (personal message) but I don't know that many do this. Removal reasons in threads are not something that can be handed smoothly now that Reddit has millions of users." - P5*

> *"There are only a handful of genuine posts that get removed and those people do not learn the reason. But such posts are in minority and it's not worth the effort to write reasons and notify users about every removal because most of them are spam or actually bad posts" - P11*

Other moderators selectively provide reasons to users they believe were participating in good faith. One participant (P8) told us that he provides reasons for removals only to community members who have broken rules either once or twice. Two moderators revealed that posts are more important than comments and thus, moderators of that subreddit provide reasons only for post removals. For example,

> *"The post are the meat of the subreddit, right? The posts are what's going to get pushed up to the front page or what's going to get seen. We're less concerned about the comments. " - P1*

Although few moderators favored either selective or no explanations for content removal, some of them admitted that they were open to the idea of providing reasons if the process is automated by Reddit. For example,

> *" I just don't do that because I would have to, like, manually type message. Um, but if it was just, like a pop up when I clicked, removed post and said, Please select why you're removing this comment, and then it automatically posted a message. I would use that. " - P8*

6.2.3 Vague Rules. Previous studies have shown that community policies and rules can be unclear and vaguely worded [20, 47] making it difficult for users to understand them. While mapping posts/comments to rules during our qualitative coding process (RQ1 phase), we discovered several vaguely worded rules that were open to interpretation. For example, "Be nice. Participate in good faith". Upon talking to moderators, we discovered that few moderators follow this practice to ensure that people don't find their way around these rules.

> *" We have a general rule that mods can use their discretion to do whatever they need to. That covers pretty much everything that you might be referring to." - P12*

> *"Participate in good faith is a rule but it is also vague because you just have to just be on the subreddit to understand what is participation in good faith and what isn't..... We have a rule against abuse too and kept it a little vague. You may call it unethical and not very transparent, but this is the best we can make things work.......We also just added a rule that says if there is no rule that covers a certain issue, exactly or completely, then mod team will have a unanimous decision and do what is best needed." - P7*

> *"We have to be a little vague about precisely what it is that all the moderators blocking so that people don't know how to get around it " - P3*

6.2.4 Non-Exhaustive Rules and their Hidden Enforcement. In a previous interview based study, scholars found that moderators avoid making rule changes transparent to avoid conflicts with the community [57]. In RQ1 analysis, we found that even rule operationalization and enforcement is not transparent. Multiple subreddits employ AutoModerator to filter content based on karma threshold, account age, presence of swear words and racial slurs. For example, subreddit MurderedByWords remove posts/comments that contain the term *feminazi*. Some subreddits maintain *whitelists*— a list of approved sources of news media outlets that are considered legitimate. The whitelist is voted on by all mods. A source is added to the list if majority of moderators agree upon it. Any submission containing source that is not whitelisted goes to the spam queue and is looked at by moderators with full permissions. Some subreddits also have blacklisted domains and projects, talking about which will get the content deleted. Except in a few subreddits, none of these lists have been publicly revealed in entirety as part of any rules. Again, the views of moderators were divided on this issue.

Seven moderators favored hidden implementation of karma, account age requirements and word filters. Majority of them argued that if more information is given to the users regarding the rules, trolls will be able to game the system. Most of the blacklisted words are specified in AutoModerator configuration using regexes. Moderators believe that people can get around these filters if they make such words publicly available. For example, people can supplement a character in a word with a number or a symbol and the AutoModerator won't be able to catch it. Similarly, miscreants can exploit the (public) karma threshold by creating fake and duplicate accounts that have minimum required karma to post in the subreddit.

> *"It would be silly to draw attention to something like that. Almost like daring users to write those things." - P5*

> *"We don't reveal the details of that publicly to prevent spammers from manipulating the karma threshold. We wouldn't state it as a rule because we wouldn't want to give any insight to those looking to exploit the karma requirement. We want to make that part hard to find, because then it can't be manipulated." - P4*

Two moderators revealed that they do not make AutoModerator configuration public since the words they remove are commonly known racist and swear words.

> *"It's common sense that our users shouldn't use any slurs, sexism, trans-phobia or racism over our community." - P6*

> *"I think everyone is aware that "nigger" isn't ok." - P12*

One moderator admitted that even though she publicly shares the details of AutoModerator configuration, she regrets her decision.

> *"As for if we tell them we've banned certain words?, the answer is yes, but it never goes well. It leads to people trying to be cute or clever to skirt the rule. For example, when we added the word 'flowers' to the AutoMod and explained that harassing this user would be considered doxing, some people started saying roses or spelling it fl0w3rs. So, looking back, maybe it would have been better if we kept mum about banning the word, then let the AutoModerator notify us when they caught it so we could messaging people individually. It probably would have been less of a circus that way." - P13*

Moderators who are against (n=3) hidden implementation of the aforementioned rules said that they did so in the interest of transparency. For example,

> *"Lists are public. We also have regular surveys to ask what words should be added or removed" - P7*

> *"There was a large shake up on the moderator team when admins had to come in and change things up because all moderators were removing things that they didn't personally like or had a financial incentive against. After there was that change, there was kind of a push to be more transparent. [Therefore we made our AutoModerator configuration public.]" - P11*

Two moderators did not use AutoModerator since they moderate small communities and thus were able to manually moderate the posts/comments.

6.2.5 Complaints, Protests and Appeals Against Content Removal. One of the SCP guidelines ask social media platforms to be transparent about the appeal process. But, if users are not notified of the removal, do they even appeal? In our qualitative analysis, we discovered that comments and posts containing rants against moderators and their actions are heavily moderated. Why is it so? What is the complaint etiquette/appeal process one should follow in order to protest or complain against content removal? Responses from the moderators offer answers to these questions.

Do people protest, complain and appeal against content removal? Majority of moderators (n=10) informed us that members of the subreddit community rarely protest against deleted content. Few moderators shared that users do not notice that their post/comment was removed until they are banned. For example,

> *"People don't usually notice that their comment was removed unless they are also banned....it's extremely rare that anyone has a problem with this (content deletion). We sometimes remove entire comment chains if the parent comment is removed to avoid confusion. Should every person in a 150 comment chain that started with a parent comment doxing someone be alerted that their comment telling that person to go fuck themselves has been deleted? Nah." - P12*

Pro free speech subreddits rarely remove content, thus, protests are infrequent.

> *"Uh, I mean, we don't delete a lot of content, so that's very infrequent." - P8*

The same reason applies to inactive subreddits where either community size is small or frequency of new content getting posted is very less, therefore, less removals.

> *"Ah, I wouldn't say it happens.... because the sub reddit is I'm sad to say it's fairly quiet." - P1*

Lastly, few subreddits follow a trend where moderators re-post the deleted post/comment once a user complains about the removal. The deleted content usually contains profane language, racial slurs and arguments. The community then jumps in and informs the user that they erred. To avoid this backlash from the community, people do not complain against removals.

> *"People complain about deleted comments a handful of times... usually it's some, you know, crude angry, foul mouthed, you know, user screeching about why racial slur was removed. The moderator will re-post it for the world to see. And then basically, everyone gets to kind of, you know, circle around and laugh at the idiot who is mad that his racial, slur laden tirade was removed or whatever. So for the most part, people don't complain because it's usually reasons like that the comment was removed. Also, people just generally don't notice that their comment was removed" - P11*

> *"People don't usually protest against a deleted comment. they know I let them get away with a lot and when I finally act they've gone way past crossing a line. if they do decide to protest they'll usually just comment a protest right then and there. that usually means that the rest of the community will jump in and tell them they are wrong and looking foolish now. subreddit_name users are pretty good at keeping each other in check." - P2*

Two moderators admitted that they receive multiple complaints and protests against deleted content. Their community is very interested in learning why their content was deleted. For example,

> *"People protest against deleted content ALL THE TIME. They typically write into ModMail demanding to know why their comments were deleted or their posts removed. They are usually removed because they contain ad hominem attacks against other users or because they have little conversational value. The users do not like having their content removed. They will typically ask us if we also removed so-and-so's comments since 'they are doing the same thing,', or we're accused of working for the government or law enforcement. People definitely go a little nuts." - P13*

How do users behave while protesting and appealing against moderator's decision? Moderators reveal that people belittle them, get abusive and disrespectful while complaining about bans and content removal. This behavior does not help their cause since moderators believe that after being subjected to such behavior they do not feel liable to reinstate a comment or undo a ban. Right attitude and use of right channel to communicate is essential if one wants a moderator to re-consider mod actions. Currently, Reddit has no dedicated communication channel to appeal against content removal. Thus, some users use ModMail; some send private messages to moderators; while others post publicly on the subreddit.

> *"Mostly no one says sorry on subreddit_name" - P7*
>
> *"They make it basically impossible for us to back down of ah, removal of content because they, instead of actually disputing the reasons that we removed it, say something about how we're terrible people." - P3*
>
> *"People are abusive: Here's one from four days ago where a guy was banned for being abusive. And then he replied to the automatic private message. He said, "What the fuck?" He sent more messages. "So what's wrong with you? I was giving my fucking opinion." So he's cursing at us. " - P8*

What is the complaint etiquette/appeal process one should follow in order to protest or complain against an account ban or post/comment removal? Reddit does not have a standardized appeal process for content removal. Even Reddiquette does not provide any informal guidelines about the appeal process and etiquette. Moderators revealed that users appeal and complain against comment removal through different channels namely ModMail, personal messages or public posts. But moderators' expectations can be different with respect to receiving appeals. Our qualitative analysis suggested that rants, protests and appeals, if publicly posted as posts or comments, will be sanctioned by the moderators. Hence, it is important for a user to appeal in a way that their case gets heard. Based on the interview responses, we have come up with a complaint etiquette—guidelines that Reddit users should follow if they want to complain or protest against a content removal.

- *Approach the moderator in good faith:* Moderators expect the users to approach them politely, in good faith with a little understanding as to why their content was removed.

  "I definitely think that someone should write into ModMail, but do so in a calm manner. Ask exactly which rules were broken, but be respectful. We have definitely overturned a lot of bans this way." - P13
- *Use the right communication channel:* Publicly posting complaints on the subreddit or sending private message to every moderator is not a good idea.

  "Don't post on sub if u have a complaint. Use ModMail or message mod privately." - P3

  "They should emphatically not send every mod on the subreddit mod team a private message regarding the ban." - P5

Users should either use ModMail to message the entire moderator team or send a private message to one of the moderators depending upon the situation. For example, users should send a private message if she fears her identity is going to be gossiped about in the moderator team.

> *"Use ModMail or message mod privately: They (users) have a mod mail message button available to them in the sidebar. We would recommend, we would prefer that they reach out to us by using that button instead of by posting about it again in the subreddit." - P3*
> *"If they (users) fear that their identity is going to be gossiped about in the mod team, contact a single moderator directly to try solve the issue." - P6*

- *Last resort for appeals:* As a last resort, moderators suggest one can contact Reddit's admins to file a complaint against a moderator or open a new subreddit of their own. But before resorting to these options, one should definitely contact a moderator or the entire moderator team. This shows respect towards authority. It also reveals that the user has tried to follow the chain of authority.

> *"they should, of course, try talking to administration if they feel that a mod is breaking Reddit in their actions. that's never effective, but it's nice to show you -tried- to use the chain of command. Or, my favorite solution, make a new subreddit of your own and see if you can do things better." - P2*

> *" So it's like if you don't like what they're (moderators) doing, you can try to argue with them. You know, you can try to get a ready administrator involved. You can try anything under the sun. But if you really want to have power over your own content, you should just found your own subreddit." - P10*

6.2.6 Releasing Public Moderation Logs: A good idea? RQ1 analysis showed that several moderators' practises cannot be considered transparent. But as a step towards taking accountability of their actions and making the content moderation practices transparent, moderators of a few subreddits have released their moderation action log to the public.

Out of the 13 moderators we interviewed, 11 have at least one subreddit whose moderation logs have been made public. As explained in Section 4, public modlogs are one of the several ways using which the moderators make their modactions publicly available. All these moderators admitted that the driving force behind making moderation logs public is to be transparent.

> *"For transparency, and the idea that it'd minimize people getting mad over removals because we could easily justify them. - P6"*

Few moderators shared that public modlogs keep them accountable for their actions (see P7 for example). Few others shared that the modlogs act as a reply to users who libel the moderators, accuse them for censoring content and call them unjust and biased towards certain groups of people (see P13).

> *" The public moderation logs helps keep us accountable. " - P7*

> *"Moderation logs were made public in order to dispel the myth that moderators were favoring one group of users over the other. We wanted them to see that there was no bias, and that people from both sides were having content removed and getting punished for the same things (and for the same amounts) " - P13*

> *"Because we used to have a lot of users that demanded complete transparency and questioned every little thing, all day - every day. it was a huge distraction. " - P2*

One moderator added that the public modlogs also help in making the existence of subreddit known to the world—another unique incentive behind making modlogs public.

> *"The whole idea of that is that we want people to know that we're simply there, too." - P1*

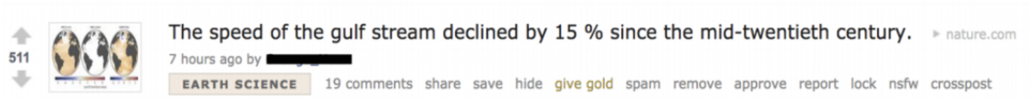
While there are a lot of positive aspects of making the moderation logs public, few moderators bear the brunt of this action. Two moderators reveal that miscreants use these logs to target moderators.

> *"So we still have quite a bit of inconsistency with what is considered an insult, or how far is too far when it comes to sarcasm, etc. We have people who check our mod log every day and really study it, then write in and confront us about punishing this person but not that person, or asking us why this comment was offensive, but not that comment. We've also run into situations where users see that an unpopular commenter is one or two bans away from their permanent ban, so they start reporting every single thing they say, spamming our report queue." - P13*
>
> *"No, because the type of people that were removing are mostly just really bad, bigoted, racist, and a public mod log would just allow them to target our moderators." - P4*

7 DISCUSSION

Transparency Violations and Moderators' Perspectives: Our findings reveal three broad trends in Reddit moderation practices that violate the transparency guidelines recommended by SCP. The first trend relates to the existence of a variety of norms in content moderation practices. The second corresponds to the lack of feedback to users about reasons for content removal. The third equates to the lack of transparency in the operationalization of rules and norms. Among the several norms that emerged in our analysis, we believe norms where moderators remove criticism of their own actions and violations are the most problematic. Such practices reveal the uncommon power that moderators exercise over the users of a community. It is very important for community's health that moderators maintain users' trust in their authority and in their fair enforcement of the rules. Some other prevalent norms that we uncovered include removal of snarky comments, conspiracy theory posts and basic naive questions about the subreddit. How can we educate users, especially newcomers about these norms? We propose that every subreddit should encourage newcomers to engage in open discussions with veteran Redditors and moderators to learn the existing norms. Such systems have been successfully used in the past when rules proved to be too ambiguous for the community. For example, players on the popular online multiplayer game, League of Legends used such forums to engage in discussions to clarify rules and learn norms. [33].

Deep content analysis on our sample of subreddits revealed that many moderators do not notify users as to *what* part of the posted content triggered the removal. Neither do they tell them *why* their posts were removed. Interviews with moderators helped us unpack these practices from the moderator's point of view. Moderators believe miscreants do not change their behavior and thus notifying them about content removal could be counter to the health of the subreddit. This suspicion is concerning because previous literature has shown that moderator feedback is important for community development and can increase users' participation and rule compliance [40]. It reduces the likelihood of users' future posts getting removed [29]. Furthermore, lack of feedback—for e.g. not providing reasons behind content removal—can adversely impact newcomers' participation in the community. Our interviews also revealed that moderators are selective in providing feedback to users. Some prefer providing feedback for post removal over comment removal since posts are lesser in number. Others may provide feedback to long time community members who have a reputation of participating in good faith discussions. Yet others selectively prefer to provide feedback only to newcomers. These practices involving selective transparency are not perfect, especially in large communities, with millions of subscribers, since keeping track of reputed old-timers and newcomers is a challenge. But they can work well in small communities where selective feedback on censored content can provide a middle ground for ensuring transparency without being gamed by trolls and miscreants. As a workaround to deal with the lack of transparency in content moderation practices, some moderators suggest that a newcomer should spend considerable time in the subreddit community to get a sense of its norms, culture and rule enforcement before actively participating in it.



The speed of the gulf stream declined by 15 % since the mid-twentieth century. ▸ nature.com

7 hours ago by [redacted]

Fig. 3. Moderator's view of a submitted post or comment. Moderators can take action on the posted content by clicking the pre-filled button labels - *spam, remove, approve etc.*. But there are no buttons that moderator can click to provide reason behind content removal. Figure has been reproduced from previous work [28].

> *"I think Reddit in general is designed to be hostile to newcomers. Not hostile with a negative intention. Every subreddit, Reddit in general to a newcomer will intentionally give off a vibe that you should learn how this works before you try to participate. To get a sense of what the community likes, what the community talks about, how it reacts to a title, a submission, a piece of content, a comment you need about six to twelve months to really pick up on that hexis" - P4*

Our study also revealed that several of the subreddits' rules are vaguely worded and their operationalization and enforcement is not transparent. By qualitatively analyzing ~6000 instances of removals made by both human moderators and AutoModerator, we identified several blacklisted words, slangs and slurs whose presence can lead to content removal. None of these blacklisted words have been publicly revealed to the users in entirety. Posts/comments also get removed if they do not satisfy karma and account age requirement. These practices are also hidden from the users. Most of these removals are performed by AutoModerator—whose automated scripts are incapable of determining the context in which a particular word was used. Thus, some law abiding posts and comments become victims of its moderation. Moderators reveal that such practices are followed to prevent miscreants from reverse-engineering the AutoModerator's configuration. This revelation raises several open questions: To what extent community moderators and social media companies hosting such communities should reveal how automated detection is used for content removal in an online community? How much transparency in moderation practices is too-much or too-little? Is there a need to provide reason for every removal? What level of granularity is required in those reasons? The answers to these questions can help in further enhancing and improving the current SCP and making social media platforms accountable and responsible in their sanctioning practices.

Design Implications: Prior studies have shown that the design of a social media platform plays an important role in promoting transparency [39]. Our study reveals how the current design of Reddit acts as a hindrance to the moderators to be transparent.

The qualitative analysis we performed showed that 69% of posts and comments removed by the human moderators were not accompanied by feedback. At present, Reddit's design does not have any feature that will automatically notify a user when her content is removed along with stating reasons for the removal. Figure 3 shows moderators' view of a post/comment. While there are several pre-filled button labels that allow moderators to take action on the content, there are no labels that allow them to provide reason for content removal. As a result, the moderator has to go to the deleted content and manually write the reason as a "reply" to the deleted post/comment. Thus, human moderators shy away from specifying reasons for every removal. Imagine a design where there is an additional button label, called *rules*, which opens a drop-down list listing subreddit's rules whenever a moderator clicks on it. Moderator just has to select the rule that the content violates. This design will considerably reduce moderator's manual effort and will enable them to be more transparent. Moreover, it also provides an option for moderators to explicitly state the unsaid norms as rules.

The design of the Reddit's mobile app is also not conducive to adhering to transparency principles. Moderators claim that most of the traffic on Reddit comes from smartphone apps where subreddit's rules are not visible on the sidebar. Lack of visibility of rules make moderators believe that users may never refer to them while using apps to post content. Thus, it makes specifying rules while providing reason for removal pointless. This situation can be rectified by introducing a pop-up which reminds users to check the community's rules while they are composing their post or responding with a comment.

Subreddit communities do not have a standardized, well-written appeal process or guidelines for content removal. Our study shows that a lack of well-defined appeal process, results in users using various communication channels (ModMail, private message, public post) to complain about removals. Based on four interview data, we compiled feedback from moderators and presented a *complaint etiquette* that users can follow to appeal against moderators' actions. The challenge of making this appeal process standardized and accessible through Reddit's platform and app still remains. We foresee that our work can inform the development of new features on both Reddit's smartphone app and website interface which would assist moderators to be more transparent.

Future Directions: Our work can take several important directions. How important is transparency (and all its facets) to the Reddit users? How important are subreddit's rules to a user? How easy and efficient do users find the current appeal procedure? Does transparent content moderation practices improve the quality and quantity of a user's (newcomer as well as old-timers) participation in subreddit communities? Investigating these questions could be fruitful avenues for future research.

8 LIMITATIONS

Our work is not without limitations. First, our analyses is limited to the moderation logs of 204 subreddits. Hence, it is unclear how representative our results are with respect to all subreddit communities in Reddit. Second, our interviews were conducted with a handful of Reddit moderators (13). While the number is small, the participants spanned a variety of subreddits, had reasonable gender representation (9 males and 4 females) and had varying moderation experience (12 to 156 months). Many of these moderators are moderating multiple subreddits and brought insights from all communities that they are moderating. Therefore, even with 13 moderators we were able to observe diverse opinions on various aspects of transparency. Third, our dataset comes from subreddits that have willingly subscribed to the publicmodlog. This indicates a readiness to share their moderation practices, implying that the moderators of these subreddits already believe in transparent practices. This limits our work in that subreddits that might be notorious for their non-transparent practices are not being studied as they do not share their moderation logs. We also excluded all the non-english subreddits from our dataset as the study involves qualitative analysis and the coders are only fluent in English. Although SCP are widely adopted by companies, there is no common standard of transparency that is universally accepted. Thus, our transparency analysis using SCP as a framing lens might not be applicable to other platforms that are not adhering to these principles. Furthermore, our study has focused on Reddit and we do not claim that results are generalizable to moderation practices in other social media platforms, like Facebook or Twitter.

Also, it is important to note that the rule mapping and qualitative coding process happened between February and March 2019 and that we do not consider changes to subreddits' rules after this period. Reddit also went through a drastic redesign of its website during this time. Some moderators were still in the process of moving their subreddit, including description, rules, wiki from the old[7] to the new site[8]. Therefore, during the rule mapping process, we checked the sidebar rules from both the old and the new versions of the website. During our qualitative coding analysis, we only examined the reasons for removal that were either posted in the thread by a moderator or

[7] https://old.reddit.com/

[8] https://www.reddit.com/

mentioned in the modlog. It is possible that moderators of some subreddits use private messaging to notify their users about post or comment removals. We were unable to consider these cases. In addition, we also consider posts and comments equally during our analysis. However, subreddits may be enforcing rules differently on posts and comments. Future work should consider these nuanced distinctions. Moreover, just like any other interview study, our interview data might suffer from social desirability bias [14]—a scenario where a participant tends to respond to questions in a way that is thought of favorably in a society. Although our interview recruitment was purely voluntary, the moderators who interviewed with us chose to enter the study, indicating the possible presence of self-selection bias [25].

9 CONCLUSION

We examined moderation practices in subreddits through a lens of transparency by analyzing 0.5M instances of posts and comments sanctioned by moderators (both human moderators and AutoModerator) from subreddit communities. Through our qualitative analysis, we discovered several prevalent norms in subreddits. We also identified several moderation practices that violate the SCP. We then interviewed Reddit moderators to understand their view on different facets of transparency. In the process, we understood why few moderators shy away from being transparent while removing content. Taken together, our study highlights the need to determine a middle ground where communities are transparent about content moderation practices but not at the cost of disruptions caused by deliberate transgressors.

REFERENCES

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[3] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1101–1110.

[4] Alissa Centivany. 2016. Values, ethics and participatory policymaking in online communities. In *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives through Information & Technology*. American Society for Information Science, 58.

[5] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 31.

[6] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 32.

[7] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

[8] Adrian Chen. 2014. The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed. (2014). https://www.wired.com/2014/10/content-moderation/

[9] Kate Crawford and Tarleton Gillespie. 2014. What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint. *New Media & Society* 18 (07 2014). https://doi.org/10.1177/1461444814543163

[10] Srayan Datta and Eytan Adar. 2019. Extracting Inter-Community Conflicts in Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 146–157.

[11] John P Davis. 2002. The experience of 'bad'behavior in online social spaces: A survey of online users. *Social Computing Group, Microsoft Research* (2002).

[12] Paul B de Laat. 2012. Coercion or empowerment? Moderation of content in Wikipedia as 'essentially contested' bureaucratic rules. *Ethics and information technology* 14, 2 (2012), 123–135.

[13] Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (2016), 56–62.

[14] Allen L Edwards. 1957. The social desirability variable in personality assessment and research. (1957).

[15] Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the article: Recognizing dialog acts in Wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*.

Association for Computational Linguistics, 777–786.

[16] Casey Fiesler, Cliff Lampe, and Amy S Bruckman. 2016. Reality and perception of copyright terms of service for online content creation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1450–1461.

[17] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.

[18] Ann Florini. 2007. Introduction: The battle over transparency. *The Right to Know: Transparency for an Open World* (01 2007), 1–16.

[19] Genny Gebhart. 2019. Social Media Platforms Increase Transparency About Content Removal Requests, But Many Keep Users in the Dark When Their Speech Is Censored, EFF Report Shows. (2019). https://www.eff.org/press/releases/social-media-platforms-increase-transparency-about-content-removal-requests-many-keep

[20] Tarleton Gillespie. 2010. The politics of 'platforms'. *New media & society* 12, 3 (2010), 347–364.

[21] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

[22] David Gilmour. 2019. Reddit mods restore Tiananmen Square image after censorship claims. (2019). https://www.dailydot.com/layer8/tiananmen-square-image-reddit-takedown/

[23] Nelson Granados and Alok Gupta. 2013. Transparency strategy: competing with information in a digital world. *MIS quarterly* (2013), 637–641.

[24] The Guardian. 2019. Revealed: how TikTok censors videos that do not please Beijing. (2019). https://www.theguardian.com/technology/2019/sep/25/revealed-how-tiktok-censors-videos-that-do-not-please-beijing

[25] James J Heckman. 1979. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society* (1979).

[26] Steve (submitter) Huffman. 2018. Reddit's Transparency Report. (2018). https://www.redditinc.com/policies/transparency-report-2018

[27] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 1, 1 (2019).

[28] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *TOCHI* (2019).

[29] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 2 (2019).

[30] Brian Keegan and Casey Fiesler. 2017. The Evolution and Consequences of Peer Producing Wikipedia's Rules. In *Eleventh International AAAI Conference on Web and Social Media*.

[31] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design* (2012), 125–178.

[32] Jason Koebler and Joseph Cox. 2018. The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People. (2018). https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works

[33] Yubo Kou and Bonnie A Nardi. 2014. Governance in League of Legends: A hybrid system.. In *FDG*.

[34] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1675–1684.

[35] Himabindu Lakkaraju and Cynthia Rudin. 2017. Learning cost-effective and interpretable treatment regimes. In *Artificial Intelligence and Statistics*. 166–175.

[36] K. Leetaru. 2018. Without transparency, democracy dies in the darkness of social media. Forbes. (2018). https://www.forbes.com/sites/kalevleetaru/2018/01/25/without-transparency-democracy-dies-in-the-darkness-of-social-media/#694732567221

[37] Lawrence Lessig. 2009. *Code: And other laws of cyberspace*. ReadHowYouWant. com.

[38] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research* 5, Apr (2004), 361–397.

[39] Sorin Adam Matei, Martha G Russell, and Elisa Bertino. 2015. *Transparency in social media*. Springer.

[40] J Nathan Matias. 2016. Posting rules in online discussions prevents problems & increases participation.

[41] J Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media+ Society* 5, 2 (2019), 2056305119836778.

[42] Albert Meijer. 2009. Understanding modern transparency. *International Review of Administrative Sciences* 75, 2 (2009), 255–269.

[43] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 262–272.

[44] Internet Policy Observatory. 2017. The Santa Clara Principles on Transparency and Accountability of Content Moderation Practices. http://globalnetpolicy.org/research/the-santa-clara-principles-on-transparency-and-accountability-of-content-moderation-practices/. (2017).

[45] Elinor Ostrom. 1990. *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.

[46] Albert Park, Mike Conway, and Annie T Chen. 2018. Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: a text mining and visualization approach. *Computers in human behavior* 78 (2018), 98–112.

[47] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*. ACM, 369–374.

[48] Radha Iyengar Plumb. 2019. Exploring feedback from data and governance experts: A research-based response to the Data Transparency Advisory Group report. https://research.fb.com/exploring-feedback-from-data-and-governance-experts-a-research-based-response-to-the-data-transparency-advisory-group-report/. (2019).

[49] Santa Clara Principles. 2018. The Santa Clara Principles: On Transparency and Accountability in Content Moderation. (2018). https://santaclaraprinciples.org/

[50] Reddit. 2019. Moderator guidelines for healthy communities. (2019). https://www.redditinc.com/policies/moderator-guidelines-for-healthy-communities

[51] Reddit. 2019. Reddit Content Policy. (2019). https://www.redditinc.com/policies/content-policy

[52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[53] Sarah Roberts. 2016. *Commercial Content Moderation: Digital Laborers' Dirty Work*.

[54] Sarah T Roberts. 2016. Commercial content moderation: digital laborers' dirty work. (2016).

[55] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 487–494.

[56] Michael Schudson. 2015. *The rise of the right to know: Politics and the culture of transparency, 1945-1975*. Harvard University Press.

[57] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* (2019), 1461444818821316.

[58] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 952–961.

[59] Cynthia Stohl, Michael Stohl, and Paul M Leonardi. 2016. Digital age| managing opacity: Information visibility and the paradox of transparency in the digital age. *International Journal of Communication* 10 (2016), 15.

[60] Nicolas Suzor, Tess Van Geelen, and Sarah Myers West. 2018. Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette* 80, 4 (2018), 385–400.

[61] Nicolas P Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication* 13 (2019), 18.

[62] Twitter. 2019. Twitter's Platform manipulation and spam policy. (2019). https://help.twitter.com/en/rules-and-policies/platform-manipulation

[63] Hanna M Wallach, David M Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Advances in neural information processing systems*. 1973–1981.

[64] S.M West. 2018. Policing the Digital Semicommons: Researching Content Moderation Practices by Social Media Companies. *Paper presented at the International Communication Association Conference, San Diego, CA* (2018).

[65] S. M West. 2018. Policing the digital semicommons: Researching content moderation practices by social media companies. *Paper presented at the International Communication Association Conference, San Diego, CA* (05 2018).

[66] Dennis M Wilkinson and Bernardo A Huberman. 2007. Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on Wikis*. ACM, 157–164.

A APPENDIX: DESCRIPTION OF QUANTITATIVE METHODS

In this section we describe in detail the quantitative methods employed in our study.

A.0.1 Pre-processing. Since both the authors are fluent only in English, we removed all non-English subreddits to ease our subsequent qualitative investigations. In order to have a sizable input from each subreddit, we also removed all subreddits that had fewer than 10 entries in our dataset. After these steps, we were left with 195 subreddits and 478,081 moderated posts/comments. Topic modelling algorithms are highly sensitive to noisy data. Hence, rigorous pre-processing of the dataset is necessary in order to obtain interpretable topics. We started our pre-processing by first converting markdown texts to plain text. Following this, we used custom regexes to remove all URLs present in the posts along with special characters and escape sequence characters. Function words sometimes dominate topics and render them meaningless [63]. Therefore, we removed these words from each of the post/comment by using python's NLTK[9] and spacy's[10] English stopword list along with the standard SMART [38] stopword dictionary. One of the dangers of using common/non-important words that are not relevant to a document is that the algorithm can erroneously identify co-occurring terms. Therefore, removal of standard stop words is not enough. Hence, we further filtered out common tokens (occurring more than 200 times) and rare tokens (occurring less than 4 times in total). We used spacy's NLP pipeline to remove punctuation and non-alphabetic characters. Next, we performed tokenization and lemmatization of these tokens. We expanded our vocabulary of tokens (unigrams) by adding frequent bigrams—adjacent words that occur 20 times or more. We empirically tested with multiple combinations of pre-processing steps and found that unigrams combined with bigrams resulted in the most interpretable set of topics as output from our topic model. Finally, we fed the bag-of-words representation of these tokenized pre-processed comments to the ATM algorithm.

A.0.2 Modelling topics using Author LDA. Author Topic Modelling (ATM) allows us to learn topic representations of authors in a corpus. As mentioned before, in our implementation, each subreddit acted as an author and all posts/comments sanctioned from that subreddit acted as the documents for that author. We used the implementation provided in the gensim package with the default parameters (γ threshold=0.001, decay=0.5, offset=1.0, α='symmetric', passes=1). We tuned and increased the default number of *iterations* from 50 to 100. This parameter denotes the maximum number of times the model loops over each document. We also need to input the number of topics to be extracted from the training corpus. We varied this parameter (n) from 10 to 150 with a step size of 5 while keeping the other parameters constant. We used both quantitative metric-based ranking and qualitative human judgment ranking to evaluate the quality of our topic model and determine the optimal number of topics. First, we evaluated all the models using *u_mass coherence* [43], a metric used to determine semantically interpretable topics [58]. We then selected three models (n=45, 55 and 70) whose coherence values were the highest. Two authors then ranked these models on the basis of how interpretable they appeared (Cohen's Kappa score of 1). Finally, we selected the model with 55 topics as it was ranked highest by both the authors. Next, we qualitatively analyzed each of the 55 topics by studying the top 25 posts/comments that ranked highly in that topic.

A.0.3 Extracting high ranking posts/comments from each topic. To code and interpret each of the 55 topics we need to study sanctioned content that is highly representative of a given topic. To achieve this, we mapped every moderated post/comment to a topic. Each topic, $topic_x$ is a collection of words ordered by their likelihood or probability of occurrence of the word in that topic. For this work, we considered the top 500 words in each topic since after that the probability associated with each word becomes insignificant.

[9] https://www.nltk.org/
[10] https://spacy.io/

MC 1	AVGN, BestOfNoPolitics, Coinex, CuckoldPregnancy, DNCleaks, Diepio_, ElderScrolls, EthTrader_Test, Italian, Kelloggs, Labour, MontanaPolitics, MurderedByWords, NSFW_Snapchat, NSFWarframe, NeutralCryptoTalk, Orego_Politics, POTUSWatch, Pennsylvania_Politics, ShitClashRoyaleSays, Stuff, SugarBaby, TheCinemassacre, UpliftingKhabre, ZeroCoin, batonrouge, bonehurtingjuice, chickengifs, horny, iotchain, iranian, moderatepolitics, neogaming
MC 2	AllModsAreBastards, AltcoinBeginners, AnswersFromHistorians, AsianFeet, BadRedditNoDonut, BioshockInfinite, BitcoinCashLol, Bitcoin_Exposed, Blackout2015, CardanoMarkets, CuckoldCommunity, Dcrtrader, ElonMuskTweets, EverythingFoxes, FMTClinics, GGinSF, GitInaction, HoMM, HyperSpace, JustNews, MeanJokes, MemoCash, ModerationLog, Morrowind, Ninjago, OUR_PUBLIC_ACCOUNT, Offensive_Speech, OpenFacts, POLITIC, PicEra, PrivacyCoinMatrix, ProjectMDiedForThis, SRC_Meta, ScarletSquad, Skeletal, Stranger_Things, TrumpSalt, UncensoredPolitics, WarFrameCirclejerk, WatchRedditDie, YourOnlySubreddit, askSteinSupporters, autogynephilia, btcfork, cryptotaxation, cyubeVR, dark_humor, dnl, evergreenstate, fuckthealtfurry, healthdiscussion, paradoxpolitics, picsUL, pushshift, swcouncil, trueaustralia, verylostredditors
MC3 3	ArkEcosystem, Automate, Bellingham, BitcoinDiscussion, BitcoinSerious, Bitcoincash, BytecoinBCN, CAMSP, CardanoCoin, Corridor, CryptoCurrency, CryptoCurrencyMeta, CryptoMarkets, CryptoTechnology, CryptoWikis, Cuckold, Dirtybomb, EVEX, Ellenpaoinaction, EthereumClassic, FoxesInSnow, Gangstalking, Hotwife, HumanMicrobiome, IndiaNonPolitical, IndiaSpeaks, Iowa, KotakuInAction, Libertarian, Lightbulb, Lisk, LitecoinTraders, MakingaMurderer, MassEffectAndromeda, Oppression, PRPS2, PhantomForces, PhillyPA, Playdate, RBI, RedditCensors, ReportTheBadModerator, Ripple, SRSsucks, SocialistRA, SpaceStationThirteen, SubredditSentinals, TIL_Uncensored, The_Cabal, TotalWarArena, TrueSPH, Vinesauce, WeAreTheMusicalMakers, WhereIsAssange, XRP, animenocontext, arizonapolitics, btc, cardano, cfs, chrisolivertimes, conspiracy, decred, ethereum, ethtrader, gamers, i_irl, information, knives, liberalgunowners, ndp, neutralnews, nyancoins, olympia, pivx, pussypassdenied, pythoncoding, racistpassdenied, radeon, recycling, reverseanimalrescue, seedboxes, siacoin, smallboobproblems, socialism, speedrun, subredditcancer, talkcrypto, tanlines, tezos, thewitcher3, torrentlinks, uber, uberdrivers, viacoin, virgin
MC 4	ConspiracyII
MC 5	Indian_Academia

Table 6. Meta communities and constituting subreddits

For every post/comment, we calculated $p(target\_body|topic_x)$—probability that a post/comment belongs to topic x. This value is the sum of probabilities of occurrence of words present in the post/comment, in topic x. Finally, the topic for which the calculated sum of probabilities is the highest (max_sum) is assigned to the post/comment. Step 4 in Figure 1 illustrates this approach with an example. Then, we sorted all posts/comments belonging to topic x in decreasing order of their max_sum and extracted the top ones. We don't employ a tie breaking strategy. If several posts/comments have same max_sum, we randomly selected the posts/comments for analysis. We used the top 25 highest ranked posts/comments to interpret each of the 55 topics obtained after ATM step. We present the method of coding these topics as well as the codes briefly in Appendix B. It is important to note that we study all these topics in detail from each of the meta communities as well. If we study top posts/comments directly from topics, few huge subreddits might dominate. To ensure that subreddits are equally represented in the qualitative analysis, we study them in meta communities.

A.0.4 Community Detection using Louvain. We used python's `community` package's implementation of Louvain's community detection algorithm. To empirically find "meta communities", this algorithm requires a graph input representing distance between data points. Therefore, we built a graph in the form of an adjacency matrix containing distance between subreddits. Each subreddit is represented by its topic distribution obtained from the Author LDA step. In other words, each subreddit is represented by a vector of length 55 where the i_{th} entry in the vector corresponds to the probability of occurrence of i_{th} topic in that subreddit. Several similarity measures can be used to quantify the distance between two probability distributions, such as Kullback_Leibler divergence, Wasserstein distance, Bhattacharyya distance or the Hellinger distance. We chose Hellinger distance metric, a probabilistic equivalent of Euclidean distance that returns similarity value in the range of [0,1]. Values closer to 0 indicate that probability distributions are more similar. We calculated distance between subreddit pairs using this metric, filled in the adjacency matrix and fed the matrix as input to Louvain. After the application of this algorithm, we obtained 5 clusters. Each cluster is considered a "meta community" that disallows the same kinds of infractions. We describe each cluster below and present its constituting subreddits in Table 6.

Meta Community 1: Gaming, erotic and political communities: These set of communities consist of 33 small and medium sized subreddits with mean and median subscriber count of 62248.48 and 4412 respectively. It consists of communities discussing political (r/POTUSWatch, r/moderatepolitics, r/iranian) , gaming (r/ElderScrolls, r/TheCinemassacre, r/AVGN) and erotic (r/CuckoldPregnancy, r/SugarBaby) themes. Some of these subreddits like r/iranian, r/MurderedByWords and r/POTUSWatch provide a platform for open discourse. While r/MurderedByWords is a community for sharing well-constructed take-downs or counter-arguments on a myriad of topics ranging for anti-vaccination to pop-culture, r/POTUSWatch is a community that discusses actions and statements of the POTUS (President of the United States) and his administration, for example, gun and immigration laws. We annotated 2746 posts from this cohort.

Meta Community 2: Pro free speech and anti-censorship communities This meta community consists of 57 small and medium sized subreddits with mean and median subscriber count as 6, 720.7 and 599.5 respectively. Its subreddits have few to no rules. We annotated 157 posts from this meta community. We had fewer moderated comments from these communities because the majority of subreddits in this cohort are pro free speech and provide a censorship free platform to the users. For example, one subreddit (r/Dark_Humor) has this rule: *"Dont be a whiny faggot. Do not complain about racism, sexism, homophobia and the like".*

Meta Community 3: Cryptocurrency and special interest communities This meta community consists of 96 medium and large sized subreddits with mean and median subscriber count as 63305.9 and 13498 respectively. We annotated 2750 posts from this cohort. Most of the posts that we study from this meta community were dominated by the larger subreddits - CryptoCurrency, ethereum, ethtrade, KotakuInAction, Conspiracy, Socialism, NeutralNews and IndiaSpeaks. All of these subreddits have a set of exhaustive and well defined rules. Most of them have explicitly mentioned the karma requirement, minimum character count of comments required and defined what constitutes a spam and a low quality content. However, we notice that the communities are not fully transparent in the execution of these rules. They haven't revealed the exhaustive list of blacklisted words, domains and projects which we think can keep a user guessing.

Meta Community 4: Conspiratorial community This meta community consists of a singleton subreddit - ConspiracyII. It is dedicated to discussions about alternative history, esoteric concepts and occultism. It had 15, 000 subscribers and well defined set of rules. We annotated 591 posts from this meta community.

Meta Community 5: Academic community This meta community also consists of a single subreddit - Indian_Academia. It is a community dedicated to discussions about Indian higher education, research, admissions process and similar topics. It had 2, 100 subscribers and only a single rule directing users to make the title of the posts informative. We annotated 16 posts from this meta community out of which only 5 posts had reasons for removal associated with it. The moderators of this subreddit remove comments and posts containing shortened URLs, having demeaning language, obscenities, sarcastic comments, spam bots, advertisements and promotional content despite having no corresponding rules.

B APPENDIX: INTERPRETATION OF THE TOPIC MODEL

To code and interpret each of the 55 topics obtained from the Author Topic Modelling step, two authors independently coded each of the topics by taking into account the top 25 posts/comments that ranked highly in each of the topics (see Appendix A.0.3). The coding was done in an inductive and iterative fashion [7]. The authors studied the content of the posts/comments as well as the description of the subreddits where they were posted in while coding the topics. Each author independently assigned a category to a topic. In the end, both authors came together to compare and adjudicate the categories that they coded independently. The disagreements and conflicts were resolved by discussions and by re-iterating over the codes. In the end, we came up with 33 unique topic codes. In Figure 4 the 33 topic codes obtained after the qualitative coding are grouped into six categories. Please note that this grouping is purely qualitative and sometimes each topic had several themes. The topic code was assigned the theme that was present in majority of the posts analyzed. The number of topic codes are less than the total number of topics. This is because several topics were coded similarly. For example, more than 10 topics were coded as spam. Even though we discarded modactions 'spamlink' and 'spamcomments', we found that several topics were dominated by spam. It shows that sometimes moderators remove spam post/comment like regular content and do not specifically mark it as *spam*. We briefly discuss the categories of topic codes below:

Topics about Conspiracy theories: We found 3 topic codes on hoaxes, rumors and conspiracy theories. These theories were pertaining to several controversial topics like air travel, nuclear warfare and climate change.

Topics about controversial political and geopolitical themes: Ten codes were grouped in this category. They contain themes ranging from the controversial conflict between Israel and Palestine to the equally controversial 2016 US elections. These codes were obtained from conversations about Iranian leaders and popular figures like Khomeini, Shah and Rajavi, Antifa's violence, theories claiming Democratic National Party manufactured the Russian Campaign during US elections 2016 and content remembering John McCain's political career after his death.

Topics about spam and promotional content: Announcement, promotion and shilling of new technology and projects was also subjected to removal. 4 topic codes were grouped into this category.

Topics about low quality content: This category contains 9 codes. Comments and posts where users seek and offer help, personal anecdotes, exchange of pleasantries (phatic talk) and fiery arguments were sanctioned by the moderators. Mockery and short posts are also not considered as quality content.

Topics about erotic content: This category contains topics where erotica, sexual experiences, cuckold and hotwife lifestyles were being discussed.

Miscellaneous: This category includes topics containing non-english posts, facebook links, discussions about environment (e.g pros and cons of recycling plastic) and tv series.

C APPENDIX: INTERVIEW PROTOCOL

The interview was semi-structured and some questions were adapted from the conversation when needed. But in general, the following script was used to interview Reddit's moderators.

C.1 Primary Questions

(1) How long have you been active on Reddit?
(2) When and where did you first start moderation on Reddit?
(3) Can you describe the subreddit *subreddit_name* to us.
(4) How much time do you devote to your moderation duties?
(5) Can you take us through the entire moderation process followed by the subreddit?
(6) Can you describe all the rules of your subreddit.
  (a) *[Follow-up example...]* How do you define a low quality comment? Are there any general guidelines you follow while determining the quality of content in a comment or post? How do you treat low quality comments?
  (b) Which rules are most violated by newcomers?
  (c) Which rules are most violated in general?
(7) How stringently is reddiquette enforced?

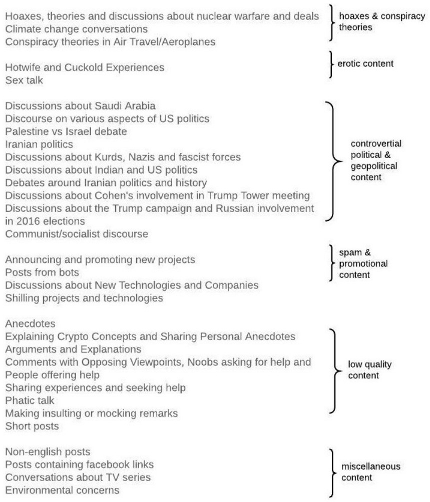

Hoaxes, theories and discussions about nuclear warfare and deals
Climate change conversations
Conspiracy theories in Air Travel/Aeroplanes
hoaxes & conspiracy theories
Hotwife and Cuckold Experiences
Sex talk
erotic content
Discussions about Saudi Arabia
Discourse on various aspects of US politics
Palestine vs Israel debate
Iranian politics
Discussions about Kurds, Nazis and fascist forces
Discussions about Indian and US politics
Debates around Iranian politics and history
Discussions about Cohen's involvement in Trump Tower meeting
Discussions about the Trump campaign and Russian involvement in 2016 elections
Communist/socialist discourse
controvertial political & geopolitical content
Announcing and promoting new projects
Posts from bots
Discussions about New Technologies and Companies
Shilling projects and technologies
spam & promotional content
Anecdotes
Explaining Crypto Concepts and Sharing Personal Anecdotes
Arguments and Explanations
Comments with Opposing Viewpoints, Noobs asking for help and
People offering help
Sharing experiences and seeking help
Phatic talk
Making insulting or mocking remarks
Short posts
low quality content
Non-english posts
Posts containing facebook links
Conversations about TV series
Environmental concerns
miscellaneous content

Fig. 4. Topics obtained from moderated posts and comments

(8) Can you remember a time when you had to moderate something but there was no explicit rule specified?
(9) Are there any practices in your subreddit that you have to be a long term member to understand? In other words, are there any community norms/values that are followed throughout the subreddit and are understood by the existing members but haven't been formally specified as a Rule?
(10) What parts of your job have you automated?
  (a) Are there some Rules that have been completely automated such that you completely rely on AutoModerator to find the violations for that rule?
(11) Are the comments removed by the AutoModerator re-reviewed by human moderators?
(12) Do you have a list of words whose use is prohibited in the subreddit?
  (a) *[If yes..]* Have you made it public i.e specified it as a part of any of the rules?
(13) How often do you encounter racial slurs and other profane content? How do you act on such comments?
(14) Can you recall a time when a person protested against a deleted content?
  (a) How often does this happen?
  (b) How do you respond to such behavior?
(15) What do you believe a reddittor should do if he/she feels that their comment has been wrongly removed or they have been unjustly banned?
  (a) Where and how does one report a moderator? And how are these reports handled?
(16) How does a user gets notified when his comment is deleted or he is banned?
(17) How does the user learn why his content was deleted?
  (a) Are community members interested in learning why their content was removed by the moderators?
(18) Do you think specifying the rules detailing why the post/comment was deleted is helpful for the community? Why or why not?
(19) How do you think moderation practices affect user participation in a subreddit?
(20) What was the purpose of making moderation logs public?

(21) Can you think of any significant experience that you had with the subreddit in the recent past as a moderator?
(22) Do you think Reddit can help you in any way in terms of policies, interface and tools to make your job easier or more effective?
(23) Are there any interesting questions that I have failed to ask you?

Received June 2019; revised October 2019; accepted November 2019