Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit

SHAGUN JHAVER, Georgia Institute of Technology
AMY BRUCKMAN, Georgia Institute of Technology
ERIC GILBERT, University of Michigan

When posts are removed on a social media platform, users may or may not receive an explanation. What kinds of explanations are provided? Do those explanations matter? Using a sample of 32 million Reddit posts, we characterize the removal explanations that are provided to Redditors, and link them to measures of subsequent user behaviors—including future post submissions and future post removals. Adopting a topic modeling approach, we show that removal explanations often provide information that educate users about the social norms of the community, thereby (theoretically) preparing them to become a productive member. We build regression models that show evidence of removal explanations playing a role in future user activity. Most importantly, we show that offering explanations for content moderation reduces the odds of future post removals. Additionally, explanations provided by human moderators did not have a significant advantage over explanations provided by bots for reducing future post removals. We propose design solutions that can promote the efficient use of explanation mechanisms, reflecting on how automated moderation tools can contribute to this space. Overall, our findings suggest that removal explanations may be under-utilized in moderation practices, and it is potentially worthwhile for community managers to invest time and resources into providing them.

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

Additional Key Words and Phrases: content moderation; content regulation; platform governance; post removals

ACM Reference Format:
Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 150 (November 2019), 27 pages. https://doi.org/10.1145/3359252

1 INTRODUCTION

Social media platforms like Facebook, Twitter, and Reddit have become enmeshed in a wide range of public activities, including politics [19], journalism [44], civic engagement [17], and cultural production [45]. As such, the decisions that these platforms make have a substantial impact on public culture and the social and political lives of their users [12, 18]. Unfortunately, the black-box nature of content moderation on most platforms means that few good data are available about how these platforms make moderation decisions [27, 58]. This makes it difficult for end users,

Authors' addresses: Shagun Jhaver, jhaver.shagun@gatech.edu, Georgia Institute of Technology, 85 5th Str. NW, Atlanta, GA, 30308; Amy Bruckman, asb@cc.gatech.edu, Georgia Institute of Technology, 85 5th Str. NW, Atlanta, GA, 30308; Eric Gilbert, eegg@umich.edu, University of Michigan, 105 S State St, Ann Arbor, MI, 48109.

2573-0142/2019/11-ART150 $15.00
https://doi.org/10.1145/3359252

particularly those with low technical expertise, to form an accurate mental model of how online content is curated. For example, most of the time on Reddit, content simply disappears without feedback. This lack of transparency of moderation decisions can diminish the comprehensibility of content regulation, which can decrease users' trust in social media platforms.

One strategy for improving transparency is to provide end users with explanations about why their content was removed. Prior research on explanations span a number of different fields such as cognitive science, psychology and philosophy [25]. The importance of explanations in providing system transparency and thereby increasing user acceptance has been demonstrated in many areas: e-commerce environments [46, 61], expert systems [34], medical decision support systems [2], and data exploration systems [7].

What effect does providing explanations have on content moderation? When equipped with the right explanation mechanisms, moderation systems have the potential to improve how users learn to be productive members of online communities. Explanations could provide individualized instructions on how to complete tasks such as making a successful submission or finding the right community for their post. However, this obviously comes with a cost: someone has to spend time crafting and delivering explanations to users whose content has been removed.

In this work, we focus on understanding transparency in content moderation on the popular social media platform Reddit. Reddit has more than a million subcommunities called subreddits, with each subreddit having its own independent content regulation system maintained by volunteer users. In this way, the Reddit platform provides a rich site for studying the diversity of explanations in content management systems and their effects on users.

Our analysis is guided by the following research questions:

- RQ1: What types of post removal explanations are typically provided to users?
- RQ2: How does providing explanations affect the future posting activity of users?
- RQ3: How does providing explanations affect the future post removals?

We break our analysis into two parts. First, we present a general characterization of removal explanations that are provided on Reddit communities. This characterization provides a descriptive sense of the types of information made available to users whose posts are moderated. Applying topic modeling techniques on a corpus of 22K removal explanations, we found that explanations not only provide information about why submissions are removed, they also reveal the mechanics of how moderation decisions are made, and they attempt to mitigate the frustrations resulting from content removals. We also characterize the differences between explanation messages offered through different modes (comments v/s flairs, described in Section 2), which we further inspect in our subsequent analyses. Next, we explore quantitative relationships between removal explanations and subsequent user activity. We also analyze how different elements of explanation such as the length of explanation, the mode through which it is provided, and whether it is offered by a human moderator or an automated tool affect user behavior.

Our findings show that provision of removal explanations is associated with lower odds of future submissions and future removals. We also find that offering explanations through replying to the submission is more effective at improving user activity than simply tagging the submission with a removal explanation. We build on our findings to provide data-driven guidelines for moderators and community managers in designing moderation strategies that may foster healthy communities. Our results also suggest opportunities for moderation systems to incorporate education (over punishment), and we discuss how such a shift may help communities manage content at scale.

We begin by describing Reddit moderation, the context of our study. Next, we situate our research in a body of literature that focuses on content moderation and transparency in moderation systems. Following this, we discuss how we collected and prepared data to answer our research questions

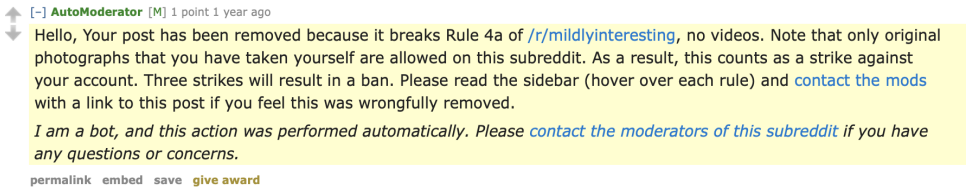


Fig. 1. An example explanation message provided through a comment to a removed submission.

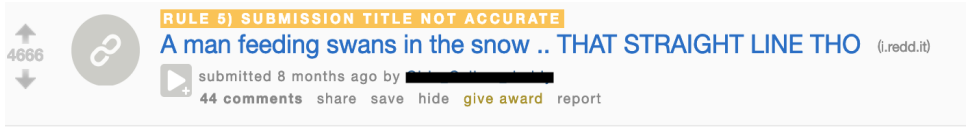


Fig. 2. An example explanation message provided by flairing the removed submission. Username has been scrubbed to preserve the anonymity of the submitter.

for this study. We then present an overview of explanations for content removals that are provided on Reddit communities using topic modeling and n-gram analyses. Next, we describe our methods, before detailing our findings on how explanations are associated with future activity of Reddit users. Finally, we use the insights gained from our overview of removal explanations to ground our quantitative results, and articulate the lessons learned from our research for the benefit of site managers, moderators, and designers of moderation systems.

2 STUDY CONTEXT: REDDIT MODERATION

Our study focuses on the popular social media platform Reddit[1]. Founded in 2005, Reddit is a social news aggregation website that hosts user-generated content. This content is organized into thousands of active, independent, user-created communities called *subreddits*, with each subreddit focused on a separate area of interest. We use the terms *communities* and *subreddits* interchangeably in this paper. Registered Reddit users can *subscribe* to each subreddit in order to see the popular content from that subreddit on their personalized front pages.

On every subreddit, users can post submissions that include text-based posts, images and videos. These posts are usually referred to as *submissions* or *posts*. After these submissions are posted on a subreddit, they are commented upon by other members of the community. Users can also *upvote* or *downvote* each submission, and the aggregations of these votes, referred to as the *score* of the submission, determine its visibility.

Each subreddit is regulated by a volunteer group of users called *moderators* who participate in a variety of tasks that include creating subreddit rules, removing content that violates rules, and responding to user inquiries and complaints. Moderators also use automated tools or bots that assist them in enacting a variety of moderation tasks [27, 39]. Part of moderators' work includes configuring automated tools for moderation, and verifying whether these tools are operating as expected [27].

Our paper is concerned with the removals of submissions and the subsequent actions taken by the moderation teams. We focus only on moderation of submissions but not comments in this study because our interviews with Reddit moderators in previous studies [27, 31] led us to believe that explanations for comment removals are provided extremely rarely on Reddit. When a submission is

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

removed on Reddit, moderators can choose to provide the submitter with an explanation for why this removal occurred. This can be done in a variety of ways. For example, moderators can comment on the removed post with a message that describes the reason for removal (Figure 1). Alternatively, they can flair[2] the removed post (Figure 2), or send a private message to the submitter. Moderators can either choose to compose the removal explanation themselves, or they can configure automated tools (e.g., AutoModerator [27]) to provide such explanations when the submission violates a community guideline.

Our analysis focuses on how content removals affect future user behaviors on Reddit. We quantify the user behaviors using two measures: (1) whether the user posts a submission, and (2) whether the user's posted submission gets removed. We also explore how providing explanations and the different attributes of explanations affect these measures of user behavior.

3 RELATED WORK

3.1 Content Moderation

Content moderation determines which posts are allowed to stay online and which are removed, how prominently the allowed posts are displayed, and which actions accompany content removals [20]. It is important for platforms to enact content moderation efficiently so as to ensure that low-quality posts don't drown out useful content and exhaust the limited attention of users [32]. Perhaps more importantly, content moderation is critical to determining which groups' voices get heard, and whether minorities and other vulnerable groups are able to participate in online public spheres [28, 29].

Today, millions of individuals use social media platforms like Facebook, Twitter, and Reddit. With a rapid increase in the numbers of users who use these sites, platforms have had to quickly devise ways to process, examine, and curate content at a scale that was previously unimaginable. For example, according to an estimate, Facebook now processes 4 new petabytes of data per day [5]. This has led to the development of complex, multi-layered content moderation systems that include "visual interfaces, sociotechnical computational systems and communication practices" [63].

The complexity of content moderation infrastructure and the opacity with which social media platforms operate make it difficult to examine how moderation systems work. Yet, over the last few years, researchers have made important forays into understanding the different aspects of content moderation, often using theoretical or qualitative approaches. For example, Grimmelmann [23] contributed multiple taxonomies of content moderation, showcasing the wide variety of ways in which moderation mechanisms can be implemented in a community. Roberts studied the work of digital laborers tasked with enforcing moderation policies and presented rich understandings of governance from their perspectives [50, 51]. Crawford and Gillespie [11] analyzed how platforms rely on regular users to flag content that is offensive or that violates the community rules. Lampe and collaborators investigated distributed content moderation, which involves relying on aggregation of user ratings to evaluate a comment or post [35–37]. Many researchers have explored how Wikipedia Talk pages are used to clarify and discuss content curation decisions [1, 38, 52].

Although this prior literature has started to unpack the complex, opaque system of content moderation, there still exists a gap in our understanding of how transparency in moderation at different levels affects user attitudes and behavior. Even though there is limited research on this topic, many scholars have reflected on its importance and pointed out that this concept needs deeper investigation [55, 58, 63]. Our paper begins to fill this gap by examining the concepts of transparency in moderation through a large-scale analysis of Reddit data.

[2]Flairs are short tags that can be attached to users' submissions. Only the moderators on each subreddit have access to assign removal explanation flairs to the posts on that subreddit.

Our research is related to the line of work that focuses on understanding the impact of content moderation on end-users. For instance, Seering et al. analyzed Twitch messages to explore how different approaches to moderation affect the spread of antisocial behaviors [54]. Jhaver et al. studied how the use of third-party blocking mechanisms on Twitter allowed previously harassed users to meet their moderation needs and participate more comfortably on Twitter, but posed challenges for users who were mistakenly blocked [29]. Jhaver et al. also conducted a large-scale survey of Reddit users whose posts were removed so as to understand the end users' perceptions of what constitutes fairness in content moderation [26]. Lampe et al. identified the benefits and limitations of distributed moderation [35, 36], showing that distributed moderation can enable civil participation on online forums [37]. Our work adds to this research by investigating the effects of content removals on user behaviors in the distributed moderation system [23] of Reddit. We also bring to scrutiny the explanation mechanisms, an important albeit often under-utilized part of moderation systems. Our analysis presents the effects of providing explanations for content removals on the future posting activity of users. In this way, we contribute to a growing body of research that is exploring strategies beyond simply the sanctioning of problematic content or bad actors to improve the health of online spaces [31, 40].

In recent years, researchers have begun to analyze the use of automated tools for content moderation [27, 39]. As online communities grow large, such tools become increasingly important for handling the heavy traffic of posts [20]. Jhaver et al. studied the use of Reddit Automoderator, a popular automated tool provided to all Reddit moderators and found that using this tool not only helps moderators deal with the challenges of scale but it also helps reduce their emotional labor by automatically removing some of the most disturbing content [27]. Geiger and Ribes studied the use of software tools in the English-language Wikipedia, focusing on how autonomous editing programs enforce policies and standards on Wikipedia [16]. More recently, CSCW researchers have explored a variety of automated tools and mechanisms to help users make successful contributions on Wikipedia [3, 41, 62, 64]. For example, Asthana and Halfaker [3] developed an automated topic modeling approach that improves the efficiency of reviewing new articles contributed by Wikipedians. We add to this literature by highlighting the role that automated tools play in providing removal explanations on Reddit. We also scrutinize whether explanations provided by automated tools impact user behaviors differently than explanations provided by human moderators.

3.2 Transparency in Moderation Systems

Over the past few years, as social media sites have become increasingly popular, researchers have begun asking questions about the role that these sites play in the realization of important public values like freedom of speech, transparency, diversity, and socio-economic equality [20, 21, 24, 57, 59, 63]. Suzor et al. proposed evaluating the legitimacy of governance of online platforms based on how their design, policies and practices impact human rights values, which include fundamental rights and freedoms such as privacy, freedom of expression, and cultural and linguistic diversity, as well as procedural values such as due process, and transparency and openness [58]. Sloval et al. noted that moderation systems are often grounded in a punitive authoritarian paradigm but opportunities exist to emphasize an empowering paradigm in site design and policy so that users may learn valuable social skills from their experiences of conflict that naturally arise in online settings [56]. In this paper, we focus on one salient human rights value — transparency, and evaluate how Reddit moderation promotes transparency in its procedures.

Cornelia Moser defines transparency as opening up "the working procedures not immediately visible to those not directly involved in order to demonstrate the good working of an institution" [42]. Although transparency is not a new idea in governance, it has recently drawn a new surge of interest because of the transforming powers of digital technologies [15]. Internet and mobile

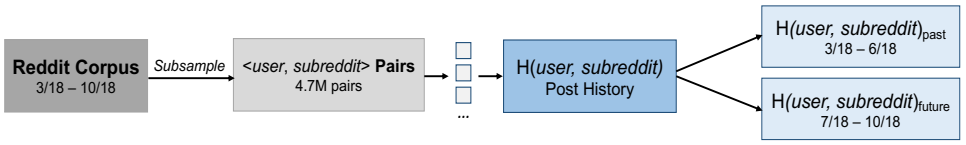


Fig. 3. Flowchart depicting the data preparation. We collected posting history for a sample of *<user, subreddit>* pairs between March and October 2018. Next, we split this posting history for each pair and aggregated posts to create H_{past} and H_{future} datasets.

technologies have reduced the information asymmetries between organizations and customers by facilitating instant dissemination of knowledge. Consequently, end-users increasingly expect to be well-informed [22]. Many media articles and non-profit organizations have advocated the virtues of transparency, and connected it to trust, corporate social responsibility, and ethics [48]. As a result, many organizations are increasingly declaring themselves as transparent in order to gain the trust of their customers.

While transparency can be seen as a means to ensure social accountability [6], the process of adopting a transparency strategy is far from trivial for organizations. Organizations need to account for the "complex dependencies, trade-offs, and indirect effects" of disclosing each informational element to their customers and competitors [22]. Some HCI scholars have also raised questions about the limits and effectiveness of transparency based strategies in sociotechnical systems [14, 30, 33, 47]. In the context of content moderation systems, social media platforms have to consider the effects of transparency not just on individual users but also on news media that are increasingly critical of moderation processes [53].

We focus on a specific aspect of transparency in moderation - the messages that provide users an explanation for why their posts was removed. In their work on Twitch platform, Seering et al. showed that banning a certain type of undesirable behavior had the effect of reducing the frequency of that behavior in the future [54]. However, whether and in what ways the reasoned explanations for removals affect future behaviors remains unclear. We see this work as one of the first steps in understanding transparency and explanations in content moderation. We aim to provide clear insights into how we can design transparent systems that encourage active and healthy participation.

4 DATA PREPARATION

We first collected a dataset D of all (allowed as well as removed) Reddit submissions that were posted over the eight months period March 2018 to October 2018. We downloaded this data using the pushshift.io service[3]. As we mentioned in Section 2, we only focus on moderation of submissions but not comments in our analyses. Following the ethical recommendations from prior research [8], we did not collect submissions that were deleted by their posters in this data. This dataset contained 79.92 million submissions, out of which 17.40 million submissions (21.77%) submissions were removed.

We wanted to explore how moderation decisions and removal explanations on prior posts of a user in a Reddit community affects the future posting behavior of that user in that community. For this, we began with identifying and removing the submissions made by bots in our data. First, we identified Reddit bot accounts by collecting a list of known bot accounts on Reddit [49] which included "AutoModerator." Analyzing the patterns of bot user-names on this list, we also considered

[3]https://files.pushshift.io/reddit/submissions/

accounts whose user-names ended with "Bot", "_bot", "–bot" or "Modbot" to be bot accounts. We also manually reviewed the user profile and posting history of accounts that posted more than 10,000 submissions, and identified accounts that were clearly bots. As there is no way to be fully certain whether a given Reddit account is a human or bot account, we acknowledge that our method only approximates distinguishing between human and bot accounts.

After identifying the bot accounts, we removed all the submissions posted by these accounts from our original dataset *D*. Next, we sampled a set of 4,705,048 <user, subreddit> pairs by retrieving all the unique <*u*, *s*> pairs where user *u* posted a submission *s* in the month of July, 2018. Following this, for each <*u*, *s*> pair, we retrieved the entire posting history H(*u*, *s*) of *u* in *s* between the period March 2018 and October 2018 (Figure 3). In total, this data sample, *S*, consisted of 32,331,120 submissions.

We split the posting history H(*u*, *s*) for each <*u*, *s*> pair in two groups - $H(u, s)_{\text{past}}$ and $H(u, s)_{\text{future}}$. The $H(u, s)_{\text{past}}$ group contains all submissions prior to and including the first submission made by *u* in *s* since the start of July 1, 2018 (mid-point of our dataset), and the $H(u, s)_{\text{future}}$ group contains all the remaining submissions made by *u* in *s*. We aggregated all submissions in $H(u, s)_{\text{past}}$ and $H(u, s)_{\text{future}}$ into the datasets H_{past} and H_{future} respectively.

4.1 Collecting Removal Explanations

To analyze the effects of past removal explanations on future behaviors, we next collected the removal explanations for removed posts in H_{past}. For each removed post in H_{past}, we first collected all the comments posted as direct replies to that post as well as the flairs assigned to the post. To distinguish removal explanation comments from other comments, we first examined all comments to a random sample of 500 removed submissions. We manually identified the comments that provided a removal explanation and were authored by one of the moderators of the corresponding subreddit or an automated moderation bot. Through inspection of these comments, we obtained an initial *removal phrases list* of 24 phrases that frequently occurred in removal explanation comments but not in other comments. These phrases included "doesn't follow rule", "submission has been removed" and "feel free to repost."

Following this, based on a snowball sampling approach, we filtered all the comments to the removed submissions in H_{past} containing any of these seed phrases and added more phrases incrementally through manual inspection of those comments. With addition of each phrase to the *removal phrases list*, we retrieved a sample of comments to removed submissions containing only that phrase and verified that the obtained comments were removal explanations. This process expanded our *removal phrases list* to 93 phrases. We searched for comments containing any of these phrases to retrieve a list of removal explanation comments.

We adopted a similar approach to create a *removal phrases list* for flairs. This list contained 32 phrases, which included "removed," "karma," and "low effort." We used this list to distinguish removal explanation flairs from other flairs. We also distinguished removal explanation comments that were provided by bot accounts from those authored by human moderators. We identified bot accounts for this step using the same approach as described at the beginning on this section.

This process resulted in a collection of 212,748 explanation messages. To evaluate the data quality, we randomly sampled 200 of these messages and manually reviewed them to see whether they provided information about post removals. This analysis found only 3 messages that were not explanations for post removals, which indicates that our approach to identify explanation messages has a high precision.

We note that some moderators may have provided removal explanations through private messages to post submitters. Because we did not have access to private messages, our data is missing these explanations. Therefore, our results should be interpreted taking this limitation into account.

5 AN OVERVIEW OF REMOVAL EXPLANATIONS

We begin by characterizing the removal explanations in our dataset and then provide an overview of how explanations provided through different modes differ from one another.

As described in Section 4, we extracted all the messages that explained why submissions were removed in H_{past}. Figure 1 shows an example of removal explanation posted by AutoModerator, an automated moderation bot popular on Reddit, as a reply comment to the removed submission. This message describes in detail the specific rule the submitter seems to have violated, the negative consequences of future rule violations, and the steps the user can take to appeal against the moderation decision. Figure 2 shows a removed submission that has been flaired with the message: "Rule 5) Submission Title not Accurate." This message specifies the community rule the submitter has broken, but doesn't provide any other contextual information.

Overall, we found that 207,689 removed submissions in H_{past} were provided removal explanations. 10.72% (N = 22,269) of these submissions received explanations only through a comment, and 86.84% (N = 180,357) received explanations only through a flair. 2.44% of removed submissions (N =5,059) received explanations through a comment as well as a flair. This shows that Reddit communities use flairs much more frequently than comments as a mechanism to present the reasoning behind their moderation decisions. The average length of removal explanations provided through comments was 728.81 characters (median = 572, SD = 510.12), whereas the average length of flair explanations was 17.34 characters (median = 16, SD = 10.44). This indicates that explanations supplied through comments are usually much more detailed than explanations provided using flairs.

Next, we sought to separate explanations provided by human moderators from those provided by automated tools. Because the Reddit API does not provide information on which Reddit accounts are responsible for flairing any submission, we could not calculate how many of the explanation flairs were produced by automated tools and how many were produced by human moderators. However, explanations provided through comments contained information about which Reddit account posted them (e.g., see Figure 1). Analyzing these comments, we found that 58.18% of all comment explanations were provided by known bots. This shows that automated tools have come to play a critical role in moderation systems — not just for removing inappropriate content but also for associated tasks such as explaining moderation decisions. We also found that a great majority (94.62%) of these automated explanations were provided by "AutoModerator," a moderation tool offered to all subreddits [27].

Next, we analyzed the removal explanation messages in order to get a descriptive understanding of the content of these messages. We began by applying standard text-processing steps such as lowercasing all characters, removing special characters, and excluding stop words from comments and flairs. We also discarded all hyperlinks that appeared in these data.

We adopted the Latent Dirichlet Allocation (LDA) technique [4] to extract the range of different types of explanations provided through comments. LDA is a widely used statistical model to discover latent topics in a collection of documents in which each topic consists of a set of keywords that defines it, and text tokens are distributed over latent topics throughout each document. We treated each explanation as a document and applied LDA on all comment explanations. We chose the number of topics, k, for this model based on perplexity scores [60], a useful measure for comparing and selecting models and adjusting parameters [43]. Testing for different values of k, we found that the perplexity score for the model dropped significantly when k increased from 5 to 28, but did not change much from 28 to 50. We also looked at the topics themselves and the highest probability words associated with each topic when using different values of k to consider if the structure made sense. Through this process, we determined to use 28 topics for comment explanations.

Table 1. Top Topics from LDA model of comment explanations with snippets of example messages. All user names in examples have been converted to 'username' to preserve the anonymity of users.

Lexical Group and Topic Terms	Examples
Removal is automatic (6.72%): "automatically", "compose", "performed", "contact", "bot", "action", "concerns"	"Your submission has been automatically removed. *I am a bot, and this action was performed automatically. Please contact the moderators of this subreddit if you have any questions or concerns.*"
Low karma (6.48%): *karma, threshold, spammers, banned, subreddits, note, automatically*	"Hello /u/username! Unfortunately, your post was automatically removed because you do not exceed our karma threshold. This has nothing to do with rule violations, it just means that your account is either too new, or doesn't have enough karma. We have a threshold to prevent spammers from posting on /r/dankmemes."
Flair post before submitting (6.18%): *"flair", "science", "forum", "post", "medical", "wait", "questions"*	"Hi username, thank you for submitting to /r/Askscience. **If your post is not flaired it will not be reviewed.** Please add flair to your post. Your post will be removed permanently if flair is not added within one hour. You can flair this post by replying to this message with your flair choice.
Ask questions in post title (5.64%): "question", "title", "answers", "post", "please", "edited", "mark"	"Your post has been removed as it violated [Rule 1] because it did not end with a question mark.* You must post a clear and direct question, **and only the question**, in your title. * Do not include answers or examples in the post title. You can post answers as comment replies when you've reposted.* Please combine clarifying sentences into the question itself."
Title must describe content (5.12%): "content", "original", "allowed", "outline", "esque", "indicating", "opening"	"Hi username, thank you for posting on /r/oddlysatisfying. Unfortunately, your post has been removed for the following reason:* **Rule 5)** The title of the submission must describe the content it shows."
Removal is unfortunate (4.93%): *submission, unfortunately, removal, contact, action, concerns, please, questions*	"Thank you for your submission! Unfortunately, your submission has been automatically removed because it contains the phrase ELI5, so it is possible you are looking for /r/explainlikeimfive. "
Don't post easily searchable questions (4.85%): "thread", "easily", "question", "questions", "daily", "topic", "reach"	""Hey there, /u/username! Thanks for your submission, but unfortunately we've had to remove your post as it doesn't follow Rule 3 - No limited scope or easily searchable questions. These types of question belongs in our [Daily Question Thread] and are not allowed as standalone posts"
Check rules in the sidebar (4.24%): "sidebar", "check", "thinking", "appreciate", "search", "quite", "rules"	"Hey there, friendo u/...! Thanks for submitting to r/wholesomememes. We loved your submission, *r/askquija can be wholesome*, but it has been removed because it doesn't quite abide by our rules, which are located in the sidebar."
Rule number that has been violated (4.11%): "rule", "removed", "violating", "breaking", "thank", "following", "months"	"Removed, rule 1."; "Removed for rule 5"
Submission must be a direct image link (3.93%): *imgur, jpg, gif, png, links, albums, longer*	"We are no longer accepting imgur albums as submissions. Please re-submit each individual picture in the album using a direct image link (must end in jpg, gif, png, etc). Thanks. [These instructions might help.](http://i.imgur.com/RjrqakK.gifv) "

Table 1 lists the top ten topics for explanations provided through comments. We manually labeled the topics and measured the relative frequency with which each topic occurs in the data. Specifically, given $\theta(topic_i) = \Sigma p(topic_i|comment)$ over all comments, the values in the parentheses in the first column correspond to $\theta(topic_i)/\Sigma_j\theta(topic_j)$, expressed as a percentage. This table also shows the corresponding keywords as well as explanation examples from each topic. The remaining topics in our analysis reflected a variety of themes. For examples, we identified topics such as "a submission with the same title has been posted before," "the submission is 'low-effort'," and "the submission is unrelated to the subject of the subreddit."

Explanations offered through comments often provide information about the reason why the submitter's post was removed. For example, topics like "Low karma" and "Flair posts before

Table 2. Frequent phrases from Removal Explanation Flairs

Unigram		Bigram		Trigram	
Phrases	**Frequency**	**Phrases**	**Frequency**	**Phrases**	**Frequency**
removed	65817	removed rule	20149	non whitelisted domain	2671
rule	53902	low karma	7038	rule overposted content	2252
fluff	24773	removed repost	3119	rule non gore	1350
repost	17485	fluff question	2896	r14 social media	1179
low	10737	non whitelisted	2671	social media sms	1179
submitted	10238	whitelisted domain	2671	media sms removed	1179
karma	7138	low effort	2593	removed crappy design	1151
title	6816	repost removed	2548	use approved host	1110
content	5406	rule overposted	2252	approved host removed	1110
post	4397	overposted content	2252	assign flair post	979
non	4299	appropriate subreddit	1629	low effort meme	902
shitpost	3812	rule repost	1602	removed restricted content	875
question	3774	social media	1594	removed location missing	849
domain	3387	rule animeme	1528	removed low quality	715
r1	3254	rule non	1491	r3 repost removed	637

submitting" suggest attempts to explain to the users why their post warranted expulsion. However, we also found topics like "Removal is automatic" and "Submission must be a direct image link" which suggest efforts by moderators to make the process of automated moderation and its limitations more explicit to the users. Topics such as "Removal is unfortunate" indicate an effort to gain the confidence of the users and to cushion against the dissatisfaction resulting from the removal. We also found topics on normative guidelines such as "Check rules in the sidebar" that go beyond just the specific post in question and educate users on how to become more central members of the community.

We did not apply LDA on explanations provided through flairs because flair explanations were often too short (median length = 16 characters) to obtain valid insights using LDA modeling. In lieu of this, we extracted unique unigrams, bigrams, and trigrams from removal explanation flairs and counted their frequencies in the corpus of flair explanations. n-gram refers to a contiguous sequence of n words from text. Table 2 list the most frequent unigrams, bigrams, and trigrams for flairs. This table suggests that explanations provided through flairs do not seem to employ hedging phrases as frequently as comment explanations. They appear to be much more direct and to the point, with many common phrases like "non whitelisted domain," "overposted content," and "r3 repost removed" referring to the subreddit rule the submitter seems to have broken.

We will build upon the differences between comment and flair explanations identified in this section to analyze later in Section 7 whether these differences are associated with variations in future activity of moderated users.

6 RELATIONSHIP WITH FUTURE ACTIVITY

Building on the descriptive characteristics of explanations in the previous section, we turn to the relationship between explanations and relevant user activity measures. In this section, we describe how we developed our analytic models on *S*, the dataset containing the posting history of our 4.7 million <user, subreddit> pairs, to answer our research questions. We also discuss the simplifying assumptions we made for these analyses.

We applied logistic regression analyses on *S* for their ease of interpretability after checking for the underlying assumptions. We built these models in such a way that the independent variables derive from characteristics of submissions in the H_{past} group and the dependent variables derive

from information about submissions in the H_{future} group. In this way, we are able to analyze the relationship between moderation actions on past submissions and future user activity.

Our aim was to use these statistical models to investigate the different aspects of removals and explanations, and present results on how they relate to future user submissions and content removals. By splitting the post history H(u, s) for each <u, s> pair into H(u, s)$_{past}$ and H(u, s)$_{future}$ at the same time, our analyses aimed to control for the external events and temporal factors that may have affected future user behaviors across different <user, subreddit> pairs. For removed submissions that received explanation through a comment as well as a flair, we chose to ignore the flair explanation and considered only the comment explanation because comments are usually much longer and more informative than flairs. We do not make causal claims that the moderation practices we explore in this study leads to improved user behavior. Rather, we are providing evidence that explanations play some role in determining the users' future activity on Reddit communities.

We note that many Reddit users post on multiple subreddits, and moderation actions in one subreddit may affect user behavior in other subreddits in the future. For example, if a user's submission on the r/science subreddit is removed with the explanation message asking that user to read the subreddit rules before posting, this incident is likely to influence the user to read the community rules when posting on any other subreddit too. However, we make a simplifying assumption of treating different <user, subreddit> pairs for the same user as statistically independent in our analyses.

We explored in a separate analysis how filtering the dataset further to include only the subreddits that are active[4] would affect our results and found that the regression analyses on this filtered dataset produced very similar results. Therefore, we only present our results on the dataset without the additional filter for active subreddits. Next, we list the variables that we use in our analyses for each <user u, subreddit s> pair.

6.1 Dependent Variables

Our descriptive analyses of the data showed that the future number of submissions had a mean of 3.25 and a median of 0. We also found that the future number of removals across our dataset had a mean of 1.77 and a median of 0. Since median is a robust statistical measure of the data, we chose to focus our analyses on whether the future submissions and removals exceed their median value of 0:

(1) **Future Submission**: This is a binary variable that indicates for each <u, s> pair whether the user u has a submission in H(u, s)$_{future}$.
(2) **Future Removal**: This binary variable indicates for each <u, s> pair whether the user u has a submission in H(u, s)$_{future}$ that was removed.

6.2 Control Variables

6.2.1 Subreddit Variables. For each subreddit s in our sample, we measured these subreddit features in the month of July, the midpoint of our dataset:

(1) **Subreddit Subscribers**: Number of subscribers in subreddit s .
(2) **Subreddit Submissions**: Total number of submissions posted in subreddit s.
(3) **Net Subreddit Removal Rate**: Percentage of all submissions posted in subreddit s that were removed.

These variables are indicative of the size, activity level, and moderation rate of each Reddit community. We control for these variables because we suspected that they are likely to have an effect on user activity. We note that Reddit communities differ among one another on many other

[4]For these analyses, we considered subreddits that received more than one submission per day on average over the four months period March - June 2018 as active subreddits.

important variables which are likely to have an impact on user behaviors. For example, subreddits have different community guidelines, behavioral norms [9, 10], topics, rates of user activity, and age, among other factors, all of which are likely to influence user responses to content moderation. Since we do not account for these variations in our large-scale analyses, our statistical models are simplifications of the community dynamics on Reddit.

6.2.2 Post History Variables. Post history variables include:

(1) **Past Submissions**: Number of submissions in $H(u, s)_{past}$.
(2) **Average Past Score**: Average score (determined by the number of upvotes and downvotes) received by the submissions in $H(u, s)_{past}$.
(3) **Average Past Comments**: Average number of comments received by the submissions in $H(u, s)_{past}$.

These variables measure the number of submissions made by user u in subreddit s and the average community response measured through the number of upvotes and number of comments received by those submissions. Distributed moderation attained through community response is critical to determining how prominently each post appears on Reddit [36]. Therefore, we suspected that these variables are likely to have an effect on future user activity. We distinguished community response from centralized moderation actions because we wanted to focus on the role of explicit regulation decisions made by the moderation team.

We note that although these variables capture some basic features of community responses, we do not account for the nuances of feedback in user-to-user messages. Such feedback may also affect user attitudes about future postings.

6.3 Independent Variables

We operationalized a set of independent variables to capture different aspects of content moderation and measure their impact on users. We discuss these variables below:

(1) **Past Removal Rate**: Percentage of submissions in $H(u, s)_{past}$ that were removed. Intuitively, as the proportions of post removals in a community increase, users are less likely to post in the future. We also suspect that with increasing removals, users may learn from their mistakes and are less likely to post submissions that will be removed. Therefore, we predict that past removal rate will have a negative association with both whether the user submits posts in the future and whether the submitted posts are removed.
(2) **Explanation Rate**: Percentage of removed submissions in $H(u, s)_{past}$ that were provided a removal explanation. Our hypothesis is that if users receive explanations for a greater proportion of their removed submissions, it can provide them an understanding of the ways in which they falter in their postings, and help them become more productive in the future. We expect that explanations, as opposed to silent removals, indicate to the moderated users that the community is dedicated to providing transparency in its regulation, and the moderators are willing to engage and work with them. Thus, we predict that explanation rate will be associated with both future postings and future removals. We note that this variable is defined only for $<u, s>$ pairs where user u had at least one removed post in $H(u, s)_{past}$.
(3) **Average Explanation Length**: Average length of all explanations offered to user u in $H(u, s)_{past}$. We expect that longer explanations are likely to be more comprehensive and provide more details to the moderated user on why they were moderated and what steps the user can take to better attend to the social norms of the community. Thus, we hypothesize that an increase in explanation length will be linked to the future activity of users. We measured this length by using the number of characters in the explanation messages. This variable has

meaningful values only for <u, s> pairs where user u had at least one removed post in H(u, s)$_{\text{past}}$ that was provided an explanation.

(4) **Explanation through Comments Rate**: Percentage of removal explanations that were provided through a comment to the removed submission. Section 5 highlights some of the differences between the explanations provided through comments and through flairs. We use this measure to test whether providing explanations through comments as opposed to using only a flair has a significant relationship with future activity of users. We note that this variable is defined only for <u, s> pairs where user u had at least one removed post in H(u, s)$_{\text{past}}$ that was provided an explanation.

(5) **Explanation by Bot Rate**: Percentage of removal explanations provided by Reddit bots. We expect that when human moderators provide an explanation as opposed to a bot, the explanations are likely to be more accurate and specific to the context of the post [27]. We also suppose that users are likely to appreciate the individualized attention of human moderators more than an automatic message by a bot. It is also possible that users may consider explanation messages more seriously if they are reprimanded by a real person instead of a bot. Therefore, we hypothesize that a decrease in this rate or a corresponding increase in the rate of explanations provided by human moderators will be linked to an increase in the future activity of users and reduce instances of post removals. Note that Reddit API does not provide any information on which user account flaired a post. Therefore, we have calculated this rate only for explanations provided through comments. As a result, this variable has meaningful values only for <u, s> pairs where user u had at least one removed post in H(u, s)$_{\text{past}}$ that was provided an explanation through a comment.

We note that although the independent variables discussed above capture many important aspects of Reddit moderation that may affect user behavior, there are other factors that we do not control for in our analyses. For example, we could not account for how moderated users may be affected by their private conversations with moderators in cases where they appeal to reverse the moderation decisions because we do not have access to these conversations. Further, we could not control for how users' demographic characteristics such as their gender, race, age, and education affect their responses to content moderation. Therefore, we see our models as reasonable simplifications of the complex sociotechnical system of Reddit.

It should also be noted that community managers may provide removal explanations for reasons that go beyond providing transparency about moderation decisions. For example, this may be a signaling mechanism for the moderators to communicate to the community members that the subreddit is actively monitored, or this may indicate to the fellow moderators that a specific post has already been reviewed. Regardless of the specific motivations that drive different moderators to provide explanation messages, these messages provide users greater insight into the moderation processes. Therefore, our analyses seek to explore the effects of these messages on user behaviors.

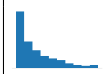
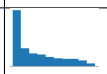
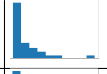
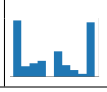
7 RESULTS

In this section, we use logistic regression models to examine the influence of independent variables on the dependent variables identified in the last section.

7.1 Descriptive Statistics

To begin, we report in Table 3 the descriptive statistics for all the variables we have introduced in the previous section, before entering into the regression models. As we mentioned in section 6.3, some of these variables do not have any meaningful value in many instances because of the way that they are defined. For example, "Explanation Rate" does not have any meaningful value for

Table 3. Descriptive statistics (min, max, mean, median, frequency distribution, number of valid entries) of all introduced variables. The distributions of post history and independent variables are shown at a logarithmic scale on y axis.

Variable Group	Variable Name	Min	Max	Mean	Median	Distribution	Valid entries
Dependent variables	Future Submission (binary)	0	1	0.37	0		4.7M (100%)
	Future Removal (binary)	0	1	0.31	0		1.8M (37.5%)
Subreddit variables	Subreddit Subscribers	0	31.8M	1.7M	91.5K		4.7M (100%)
	Subreddit Submissions	1	187.5K	12.7K	1,471		4.7M (100%)
	Net Subreddit Removal Rate (in %)	0	100	23.38	14.82		4.7M (100%)
Post history variables	Past Submissions	1	19.71K	3.62	1		4.7M (100%)
	Average Past Score	0	266.2K	100.8	3.5		4.7M (100%)
	Average Past Comments	0	72K	10.19	3		4.7M (100%)
Independent variables	Past Removal Rate	0	1	0.25	0		4.7M (100%)
	Explanation Rate	0	1	0.09	0		1.4M (29.3%)
	Average Explanation Length	2	9.9K	152.4	20		147.8K (3.1%)
	Explanation through Comments Rate	0	1	0.20	0		147.8K (3.1%)
	Explanation by Bot Rate	0	1	0.41	0		31K (0.7%)

<u, s> pairs where user u did not have any removed posts in $H(u, s)_{past}$. Thus, we create separate models for evaluating different sets of independent variables with each model containing only the valid entries for the variables considered. Table 3 lists the number of valid entries for each variable.

We found that across all (u, s) pairs, users posted an average of 3.62 submissions (median = 1) in the corresponding subreddit in $H(u, s)_{past}$. Past submissions received a median score of 3.5 (mean = 100.81) and a median of 3 (mean = 10.19) comments. Our analysis shows that in 37.5% of all cases (N = 1.73M), user u had at least one future submission in subreddit s. We also saw that for instances where users posted on the corresponding subreddit in the future, a future removal occurred in 31.2% (N = 550.5K) of the cases. The median number of subreddit subscribers is 91.5K and the median net count of subreddit posts is 1,471. This suggests that a majority of the users submit posts in large, active subreddits. Past submissions were posted in subreddits that removed a median of 14.82% of all submissions.

Table 4. Descriptions of statistical models used in our analyses. For each model, the input and output variables, criterion for including data, and the number of valid data entries are shown.

Output Variable	Model	Input variables	Inclusion criteria	Valid entries
Future Submission	A.1	Subreddit variables + Post history variables + Past Removal Rate	All <user, subreddit> pairs	4.7M
	A.2	+ Explanation Rate	Past Removal Rate > 0	1.4M
	A.3	+ Average Explanation Length + Explanation through Comments Rate	Explanation Rate > 0	147.8K
	A.4	+ Explanation by Bot Rate	Explanation through Comments Rate > 0	31K
Future Removal	B.1	Subreddit variables + Post history variables + Past Removal Rate	Future Submissions > 0	1.8M
	B.2	+ Explanation Rate	Future Submissions > 0 AND Past Removal Rate > 0	548.7K
	B.3	+ Average Explanation Length + Explanation through Comments Rate	Future Submissions > 0 AND Explanation Rate > 0	64.8K
	B.4	+ Explanation by Bot Rate	Future Submissions > 0 AND Explanation through Comments Rate > 0	15.2K

We created four regression models for each of the two dependent variables. We began creating each new model by first discarding all the cases with any missing data for the variables in the model. This was done to analyze the role of the additional variables in each subsequent model by focusing only on the cases where the variable value is meaningful. Table 4 describes what variables and data are included in each model and the number of data points for that model. Sections 7.2 and 7.3 will describe each of these models in more detail. Tables 5 and 6 show the results of these regression models.

For ease of comparing the relative importance of the explanatory variables, we standardized all the predictor variables in our models so that each variable had a mean of zero and a standard deviation of one. We report the results of our analyses as odds ratios (OR), the change in the odds of posting a submission or experiencing a removal in the future when an input variable is increased by one standard deviation. Odds ratios greater than one indicate an increase in the odds of the corresponding dependent variable, while odds ratios less than one indicate a decrease in the odds. For each model, we verified that multicollinearity was not a major problem as none of the correlations were higher than 0.5. We note that direct comparisons between the Nagelkerke R Square of different models in Tables 5 and 6 are not possible as each model is composed of a separate subset of the entire data.

7.2 Future Submission

In this section, we discuss the results of several regression models for the dependent variable, "Future Submission," a binary variable that indicates for each <u, s> pair whether the user u has a submission in $H(u, s)_{\text{future}}$ (see Tables 4 and 5).

Observation 1: High past removal rate for the user is associated with lower odds of posting in the future.

We first created a model A.1 using all the control variables (the subreddit variables as well as the post history variables) and past removal rate (Table 4). Model A.1 reports the main effects of the control variables and past removal rate on future submissions. It shows that past number of

Table 5. Odds ratio of predicting whether the user will post in the future. Here, p<0.001: ***; p<0.01:**; p<0.05:*. Each model was constructed on the corpus for which all the included variables have valid entries. The results show that higher *past removal rate* and higher *explanation rate* are associated with a decrease in the odds of users posting in the future. In contrast, higher *explanations through comments rate* and higher *explanations by bot rate* are linked to an increase in the odds of users posting in the future. Subreddit variables and post history variables are used as controls.

Group	Variables	Model A.1	Model A.2	Model A.3	Model A.4
Subreddit variables	Subreddit Subscribers	0.968***	0.988***	0.984*	0.981
	Subreddit Submissions	1.118***	1.164***	1.146***	1.151***
	Net Subreddit Removal Rate	0.940***	0.938***	0.900***	0.86***
Post history variables	Past Submissions	7.4E+10***	4.5E+6***	687.6***	16.628***
	Average Past Score	0.995 ***	0.999	0.988	0.972
	Average Past Comments	1.037 ***	1.014**	1.043***	1.057**
Independent variables	Past Removal Rate	0.978 ***	0.638***	0.563***	0.520***
	Explanation Rate		0.988***	0.686***	0.636***
	Average Explanation Length			1.003	0.990
	Explanation through Comments Rate			1.035***	0.824***
	Explanation by Bot Rate				1.264***
	# Obs	4.7M	1.4M	147.8K	31K
	Intercept	1.135***	1.154***	1.218***	1.355***
	Nagelkerke R Square	0.191	0.267	0.337	0.327
	Omnibus Tests of Multiple Coefficients	p <.001	p <.001	p <.001	p <.001

submissions overwhelmingly determines (OR = 7.4E+10) whether the user posts in the future. This is as we would expect—people who are in the habit of posting often will probably continue to post. Beyond this, since the odds ratio for past removal rate is 0.978, one standard deviation (SD) increase in the past removal rate for u in s was associated with 2.2% (= 100 * (1 - .978)) lower odds of u posting in the future. Intuitively, when users' posts continue to get removed on a subreddit, they may feel that their contributions are not welcome and stop posting, or in some cases, even leave the subreddit.

The odds ratio for the net subreddit removal rate is 0.94. This suggests that an overly strict moderation policy may have a chilling effect on users and inhibit their future postings. We also found that the odds that user u posts in subreddit s in the future increases by 11.8% (100 * (1.118 - 1)) with each standard deviation increase in the net number of submissions that s receives. This shows that regardless of other factors, users are likely to continue posting in active communities. Our results also indicate that community engagement with the user posts has a positive effect on future submissions. For example, since the odds ratio for past comments is 1.037, users who received one standard deviation increase in comments on their past submissions are 3.7% more likely to post in the future. Surprisingly, the odds of future posting reduced with increase in the number of subreddit subscribers (OR = 0.968). The average past score had a much smaller effect on future submissions (OR = 0.995).

Observation 2: Greater explanation rates characterize reduced odds of posting in the future.

Next, we created model A.2 to test the relationships between provisions of explanations and the occurrence of future submissions. This model makes a simplifying assumption that the users who received the explanation messages noticed and read them. We only considered cases where the user u had at least one post removal in the past to build this model. We found that explanation rate adds

significantly to the model even after controlling for subreddit characteristics, post history variables and past removal rate. Since the odds ratio is 0.988, one standard deviation increase in explanation rate was associated with 1.2% decrease in the odds of future submissions. One explanation for this association is that receiving removal reason messages makes users realize that their posts are being carefully reviewed, and this may make users become more cautious in their posting behavior.

We note that this result has different implications for different types of communities. For example, consider a small community that receives 100 posts a month. Assuming that the relationship between explanations rate and future posts in model A.2 applies to this community, if explanations rate is increased by one standard deviations, this community may have 1.2% fewer posts or about 99 posts a month in the future. In contrast, the same increase in explanations rate would cause a large community that usually receives 10,000 posts a month to have 120 fewer posts a month in the future. Thus, communities must consider how much decrease in traffic they can withstand when determining whether to provide explanations.

Observation 3: Having a higher fraction of explanations offered through comments, rather than through flairs, is associated with an increased likelihood of users posting in the future.

Following this, we built model A.3 to evaluate how different attributes of removal explanations affect user behavior. We only used cases where the user u received at least one removal explanation for his or her past removal to build this model. Our results in Table 5 show that explanation length did not add significantly to the model for the occurrence of future submissions (OR = 1.003). Thus, our hypothesis that longer explanations are more comprehensive and are therefore more likely to influence greater user engagement was not supported. This model, however, showed that given a fixed number of explanations, providing explanations through comments rather than through flairs is likely to cause an increase in the occurrence of future submissions (OR = 1.035).

Observation 4: Explanations provided by human moderators, rather than by automated tools, are associated with lower odds of moderated users posting in the future.

Finally, we created model A.4 to test the effects of sources of removal explanations. We only used instances where users were provided at least one explanation through a comment to the removed submission to build this model. Because the odds ratio for explanations by bot rate is 1.264 (Table 5), this model showed that one standard deviation increase in the rate of explanations provided by bots was associated with 26.4% increase in the occurrence of future submissions. Equivalently, explanations provided through human moderators are linked to reduced odds of users submitting posts in the future.

7.3 Future Removals

In this section, we analyze which factors are associated with whether a post removal occurred for submissions made in $H(u, s)_{\text{future}}$. For these analyses, we only consider the data points where user u posted at least one submission in the subreddit s in $H(u, s)_{\text{future}}$ since our focus was on distinguishing cases where removals occur from cases where there are no removals.

Observation 5: High past removal rate for a user is associated with higher odds of that user experiencing a post removal in the future.

Table 6 reports the results of several binomial regression models predicting whether a removal will occur. We began by creating a model B.1 that includes all the subreddit and post history

Table 6. Odds ratio of predicting whether the user's post will get removed in the future. Here, p<0.001: ***; p<0.01:**; p<0.05:*. Each model was constructed on the corpus for which all the included variables have valid entries. The results show that higher *past removal rate* is associated with an increase in the odds of user experiencing a post removal in the future. In contrast, higher *explanation rate* and higher *explanations through comments rate* are linked to a decrease in the odds of user experiencing a post removal in the future. Subreddit variables and post history variables are used as controls.

Group	Variables	Model B.1	Model B.2	Model B.3	Model B.4
Subreddit variables	Subreddit Subscribers	0.910***	0.934***	0.891***	0.93**
	Subreddit Submissions	1.168***	1.033***	1.11***	1.079**
	Net Subreddit Removal Rate	2.461***	2.215***	2.058***	2.021***
Post history variables	Past Submissions	1.164***	2.6***	5.636***	2.443***
	Average Past Score	0.991***	0.981***	0.96**	0.975
	Average Past Comments	1.02***	1.000	1.027	1.0
Independent variables	Past Removal Rate	1.968***	1.366***	1.236***	1.286***
	Explanation Rate		0.935***	0.701***	0.649***
	Average Explanation Length			1.003	1.002
	Explanation through Comments Rate			0.905***	0.774***
	Explanation by Bot Rate				1.019
	# Obs	1.8M	548.7K	64.8K	15.2K
	Intercept	0.392***	2.148***	2.287***	2.044***
	Nagelkerke R Square	0.378	0.187	0.199	0.231
	Omnibus Tests of Multiple Coefficients	p <.001	p <.001	p <.001	p <.001

variables as well as the past removal rate (Table 4). This model shows that the net subreddit removal rate is associated with higher odds of future removals (OR = 2.461). This suggests the expected association that subreddits that are stricter in their moderation are more likely to remove future postings regardless of the user in consideration. Our results also show that a standard deviation increase in the specific past removal rate for each user u in subreddit s leads to a two-fold increase in the odds of future removals (OR = 1.968). Thus, users who have faced more prior removals are likely to have a higher chance of facing a removal again.

Users were more likely to have their posts removed if they submitted in a subreddit that receives more submissions in total (OR = 1.168). One explanation for this is that subreddits that receive many submissions are likely to have a greater number of overall removals. However, posting in a subreddit with a higher number of subscribers was associated with lower odds of future post removals (OR = 0.910).

We found a positive Pearson correlation of statistical significance ($r = 0.366$, $p < .001$) between the number of past submissions and future submissions. This positive correlation suggests that as the number of past submissions increases, users are also more likely to submit more posts in the future, increasing the likelihood that a future removal will occur if at least one of those future posts is removed. Indeed, we found an odds ratio of 1.164 for past submissions, indicating that a standard deviation increase in the number of past submissions by a user in a subreddit was associated with 16.4% higher odds (100 * (1.164 - 1)) of future removals for the user in that subreddit. Other control variables had much smaller effects on future removals.

Observation 6: Greater explanation rates characterize reduced odds of post removals in the future.

Model B.2 adds the influence of explanation rate to future removals. This model includes only the cases where the user has received at least one post removal in the past and has submitted at least one post in the future. It shows the encouraging result that the odds of the occurrence of future removals lower by 6.5% (OR = 0.935) with each standard deviation increase in the explanation rate. This suggests that explanations help users understand their mistakes and learn the social norms of the community, enabling them to subsequently post submissions that are less likely to get removed.

This result has different implications for different types of communities. For example, consider a small community that experiences 100 post removals a month. Assuming that the odds ratios of model B.2 apply to this community, if explanations rate is increased by two standard deviations, this community may have 2 * 6.5 = 13% fewer post removals or about 87 post removals per month in the future. In contrast, the same increase in explanations rate would cause a large community that usually experiences 10,000 post removals a month to have 1,300 fewer post removals a month in the future. Therefore, moderators on different communities must judge whether the reduction in post removals are worth the investments made in providing removal explanations on their community.

Observation 7: Having a higher fraction of explanations offered through comments, rather than through flairs, is associated with a decreased likelihood of users experiencing a post removal in the future.

Next, we developed a model B.3 to understand the effects of different aspects of explanations on future removals. We found that the average explanation length did not have any significant effect on the occurrence of future removals. One possibility is that as long as explanations provide users the specific information that helps them understand why the removal occurred, the comprehensiveness of explanations do not add to their usefulness. However, we found an odds ratio of 0.905 for explanations through comments rate, indicating that a one unit increase in the rate of explanations provided through comments, rather than through flairs, resulted in a 9.5% decrease (100 * (1 - 0.905)) in the odds of future removals.

Finally, we developed a model B.4 to analyze the impact of explanation source. We found that explanations by bot rate did not have any statistically significant effect on future removals (OR = 1.019). This indicates that the source of removal explanations does not seem to have any substantial effect on the quality of subsequent posts. Our ongoing work on interviews with Reddit moderators provides one possible explanation for this. We have found that many moderators use pre-configured removal explanations in order to expedite moderation tasks. Thus, the text outputs for explanations look quite similar, whether they are provided by a human moderator or an automated tool. This may be the reason why the users seem to have similar responses to both human and bot explanations.

8 MEASURING EFFECTS OF EXPLANATIONS IN INDIVIDUAL COMMUNITIES

While the analyses in Section 7 evaluate the effects of moderation and explanations on user behaviors across all subreddits, these analyses do not sufficiently take into account the heterogeneity of different Reddit communities because our 'Subreddit variables' only control for a few important factors that distinguish subreddits. In an effort to address this, we used the approach described above to evaluate the effects of explanation rates on future user behaviors in a few individual subreddits. Through these analyses, we also demonstrate how moderators of any community can adopt our approach to evaluate the effects of explanations on that community.

For this, we filtered large, active subreddits (# subscribers > 1M, # submissions > 10K) where explanations are frequently provided (average explanation rate > 0.2). We found four subreddits that satisfied these criteria - *r/politics*, *r/pics*, *r/mildlyinteresting*, and *r/buildapc*. Taking each of these four subreddits one at a time, we built a corpus that only contained <user, subreddit> pairs

Table 7. Odds ratio of predicting (1) whether the user will post in the future and (2) whether the user's post will get removed in the future on four large, active subreddits. Here, p<0.001: ***; p<0.01:**; p<0.05:*. Each model was constructed on the corpus of the corresponding subreddit for which all the included variables have valid entries. The results show that on each subreddit, higher *explanation rate* are linked to a decrease in the odds of user experiencing a post removal in the future. Other variables are used as controls.

Subreddit	**r/politics**		**r/pics**		**r/mildlyinteresting**		**r/buildapc**	
Dependent Var.	**Future subm.**	**Future removal**	**Future subm.**	**Future removal**	**Future subm.**	**Future removal**	**Future subm.**	**Future removal**
Past Submissions	5022 ***	13.682 ***	2.181 ***	1.217 ***	3.108 ***	1.778 **	1.119	1.048
Avg Past score	0.972	0.889	1.287	1.229	1.12	0.902	0.684	1.133
Avg Past comments	1.04	1.139	0.846	0.834	0.926	1.28	1.56	1.224
Past Removal Rate	0.538 ***	1.521 ***	0.586 ***	1.756 ***	0.771 ***	1.36 ***	0.779 ***	1.331 **
Explanation Rate	0.997	0.877 *	0.943	0.561 ***	1.02	0.795 ***	0.774 *	0.48 ***
Intercept	7.304 ***	7.367 ***	1.317	0.481 ***	0.536 ***	0.431 ***	2.329 *	0.052 ***
Nag. R Square	0.382	0.084	0.204	0.21	0.104	0.049	0.188	0.224
# Obs	4357	2537	4105	1404	4670	1368	419	174

belonging to that subreddit and developed regression models that test the effects of explanation rates on future postings and future removals on the subreddit. Table 7 shows the results of these analyses. Note that we do not include subreddit variables in these analyses as all the data used in each model belong to the same subreddit. These results show that while increases in explanation rates do not significantly affect future submissions on every subreddit, they characterize reduced odds of post removals in the future in every case. This again suggests the important role that explanation mechanisms can play in improving the quality of user contributions.

9 DISCUSSION

Online communities thrive on user-generated content. However, inappropriate posts distract from useful content and result in a poor user-experience. Therefore, moderation systems usually desire to increase the number of overall contributions while lowering the number of posts that need to be removed [23, 32]. Our analyses in the previous sections explored how moderation decisions affect the occurrence of future submissions (Sections 7.2, 8). We also investigate how moderation actions shape the level of future removals (Sections 7.3, 8). In this section, we discuss the implications of our results for moderators, site managers, and designers of moderation tools.

9.1 Removal Explanations Help Users Learn Social Norms

In prior research, Kiesler et al. have suggested that people learn the norms of a community by (1) posting and directly receiving feedback, (2) seeing community guidelines, and (3) observing how other people behave and the consequences of that behavior [32]. We contend that explanations serve as learning resources for Reddit users in each of these three ways.

First, posters who receive explanation messages receive direct feedback from the moderator team in these messages. This can help them realize how their submission did not align with the norms of the community. Therefore, receiving explanations can be a moment of learning for the post submitters.

Second, as our topic modeling and n-gram analyses show, explanations messages usually mention the rule that the submitter has broken (Tables 1 and 2). For example, topics for explanation comments like "Ask questions in post title" and high frequency of flairs like "non whitelisted domain" and "low effort meme" signify a focus on educating users about the community guidelines. Explanation messages often contain a link to the wiki page for the subreddit rules. Therefore, receiving these

explanations increases the likelihood that the moderated users will attend to the community guidelines, and it would help them better understand the explicit social norms of the community[5].

Third, a noteworthy aspect of explanation messages is that they are posted publicly. Although submissions that are removed stop appearing on the front page of the subreddit, they are still accessible to the users who have already engaged with them, for example, through replying via a comment to those submissions. Therefore, many users can still see the removal explanations provided by the moderators. Observing the removal of the post and a reasoned explanation for that removal can inform these bystanders why certain types of posts are unacceptable on the community. In this way, such interactions can help these bystanders become better content submitters themselves in the future.

9.2 Removal Explanations are Linked to Reduction in Post Removals

Our regression analyses (Sections 7, 8) show that when moderated users are provided explanations, their subsequent post removal rate decreases (Observation 6, Model B.2, Table 6). As we show in Table 7, this relationship holds even in individual subreddits. These are encouraging results, as they indicate that explanations play a role in improving users' posting behaviors. This also raises an interesting question: what would happen to the quality of posted content if 100% of removals were provided explanations? We calculated a rough estimate for this based on our regression results, assuming that the relationship between explanation rate and future removals (Table 6) remains linear over a long interval, and noting that an explanation rate of 100% is about 3.20 standard deviations away from mean explanation rate. Our calculation shows that the odds of future post removals would reduce by 20.8% if explanations were required to be provided for all removals. Thus, offering explanations could result in a much reduced workload for the moderators.

Our LDA analysis of explanation comments (Section 5) shows that removal explanations are not just a mechanism to inform users about why the current removal occurred, they are also a means through which moderators can begin to develop a relationship with moderated users. We frequently saw these explanation messages thanking the submitter for posting on the community or expressing regret that the submission had to be removed. This suggests that some of these explanations may have been designed to reduce the moderated users' displeasure about the post removals. Such attempts to engage with the user, in addition to the knowledge about social norms that explanation messages provide, could explain why users who are offered removal explanations submit improved posts in the future.

Prior research has found that in the absence of authoritative explanations, users make sense of content moderation processes by developing "folk theories" about how and why their content was removed [13, 26]. These folk theories often pinpoint to human intervention, including the perceived political biases of moderators, as the primary cause of content removals [26, 63]. Our findings suggest that removal explanations can address some of these problems by providing transparency about the moderation mechanisms that shape content removals. For example, the occurrence of LDA topics like "Removal is automatic" suggest an attempt by the moderators to clarify that the post removal was made through the use of automated moderation tools and did not involve a direct human intervention. This increased transparency may improve user attitudes about the community and motivate users to submit valuable contributions.

We also found that only 1,421 subreddits, a small proportion (0.6%) of all Reddit communities in our data, chose to provide removal reason messages. Thus, explanations are an underutilized

[5]Related to this, it is important to consider whether, where, and how prominently community guidelines are posted in discussion spaces. Because certain Reddit interfaces (e.g., mobile website and some third-party Reddit apps) obscure the presence of these guidelines, they may interfere with users' ability to learn the social norms.

moderation mechanism, and site managers should encourage moderators to offer explanations for content removals. Providing explanations may also communicate to the users that the moderator team is committed to providing transparency and being just in their removals.

9.3 How should Removal Explanations be Provided?

As we discussed, Reddit moderators can provide explanation messages in a variety of ways. They can comment on a removed submission or flair it. They may compose an explanation message themselves or they may configure a bot to do it. Do these differences matter? Is one approach better than the others in improving future outcomes?

9.3.1 Comments v/s Flairs. Our analyses suggest that offering explanation through comments, rather than through flairs, is associated with a decreased likelihood of users experiencing a post removal in the future (Observation 7, Model B.3, Table 6). In a similar vein, in Observation 3 (Model A.3, Table 5), we note that controlling for explanation rate among other variables, explanation through comments rate is associated with increased odds of future posting.

Our findings in Section 5 provide clues to interpret these results. Explanation comments differ from flairs in that they cushion against the dissatisfaction resulting from post removals. They often provide information that is future-oriented and go beyond the context of the current removal. Frequently, explanation comments contain information about how users can appeal to reverse the moderation decisions in case the users consider the post removal a mistake. Explanation flairs, on the other hand, are usually very direct, do not employ hedging phrases as frequently, and only pertain to the current removal. These differences may contribute to the relationship between higher levels of explanations through comments and lower levels of future removals.

Although our regression analyses establish the effectiveness of explanation comments over explanation flairs, our data show that flairs are used much more frequently than comments to provide explanations (Section 5). This may be because the flairs are much shorter, and therefore, easier for the moderators to provide than comments. Yet, our findings suggest that it may be worthwhile for Reddit moderators to take the time to provide elegant explanations for content removals through comments rather than tagging the post with a short flair. At a broader level, these results indicate that conducting amiable, individualized correspondence with moderated users about their removed posts may be an effective approach for content moderators to nurture potential contributors.

9.3.2 Human moderators v/s automated tools. Our results show that controlling for other factors, explanations provided by automated tools or bots are associated with higher odds of moderated users posting in the future (Observation 4, Model A.4, Table 5). Additionally, explanations provided by human moderators did not have a significant advantage over explanations provided by bots (Model B.4, Table 6) for reducing future post removals. These results suggest an opportunity for deploying automated tools at a higher rate for the purpose of providing explanations. We expect that the field of explainable AI can provide valuable insights for improving the quality of explanations provided by automated tools.

Using these tools can also help address the challenges of scale. When communities grow large quickly and the moderation resources run scarce, it may be difficult for moderators to focus on providing explanations as they are instead engaged in the primary task of firefighting against bad posts. However, if the moderators set up automated tools to provide removal reasons automatically, those tools can continue to provide explanations to users even in high-traffic circumstances.

At the same time, we caution that automated tools should be used with care for the purpose of offering explanations. In cases where the removal reasons are unclear, human moderators should continue to provide such explanations. Prior research shows that an overuse of automated tools

for content moderation results in cases where these tools make mistakes, thereby causing users to become dissatisfied with the moderation processes [27]. We expect that inaccurate removal explanations are likely to increase resentment among the moderated users rather than improve their attitudes about the community. Therefore, automated tools for providing explanations should be carefully designed and deployed, and their performance should be regularly examined.

9.4 When should Removal Explanations be Provided?

Our Observation 2 (Model A.2, Table 5) states that greater explanation rates are associated with reduced odds of posting in the future. One possible reason for this is that explanations may bring users' attention to the fact that their post has been removed, which they otherwise may not have known about, owing to the frequent silent removals on Reddit. Thus, drawing attention to the removals by providing explanations may irritate users and reduce their user activity. On the other hand, Observation 6 (Model B.2, Table 6) notes that providing removal explanations is linked to lower number of future removals. Thus, although offering removal explanations may alienate some users and reduce the likelihood of their future contributions on the community, it may improve the quality of future submissions that *do* get submitted. Therefore, in determining explanation policies, moderators may need to consider whether having high traffic is more important to them than having quality content on their community.

Related to this, it is necessary to consider: In which cases is it worthwhile to provide removal explanations? Should moderators offer an explanation message for every post removal? Or should submissions or post submitters be categorized such that explanations are provided only for certain categories but not others?

It is unclear how providing explanation messages for removing content that is blatantly offensive or trollish would affect the activity of its submitters. As Observation 5 of our Findings (Model B.1, Table 6) notes, high past removal rate for a user is associated with higher odds of that user experiencing a post removal in the future, regardless of other factors. It may very well be possible that some bad actors may thrive on the attention they receive for their bad posts from the moderators, and further increase the rate at which they post unacceptable content. Yet, it is difficult to draw the line between inappropriate content that deserves explanations and blatantly offensive content that does not merit providing an explanation. This boundary may also vary between different communities, depending on their social norms, topic, and size, among other factors. Furthermore, it may be problematic to classify certain users as irredeemable and unworthy of providing explanation. Thus, significant challenges remain in determining when to use the limited moderated resources in offering explanation mechanisms. We suggest that future research should explore how distinguishing between good actors and bad actors (along a number of dimensions) when providing explanations affects the user-activity and post-removal outcomes.

9.5 Limitations and Future Work

As we discussed throughout Section 6, we have made many assumptions and simplifications to arrive at the statistical models used in our analyses. We hope that future research in this space starts to inspect these assumptions and explore the role that other factors in moderation systems play in mediating user behaviors. We have only looked at responses to removals that were publicly posted on Reddit communities. It is, however, possible that some subreddits notify users about their content removal through private messages. We only focused on analyzing transparency in regulation of submissions. However, subreddits may also be implementing different levels of transparency in comment removals. It would be useful to focus on moderation of comments in future research.

It is a limitation that this research does not divide users into people we want to post again (well-meaning users who need to be educated in the rules of the community) and people we don't

want to post again (users who are being deliberately disruptive, i.e. trolls). Of course, determining who is a troll is subjective and difficult to operationalize fairly [28]. However, in future work, we would like to separate them if possible to determine what aspects of removal explanations encourage trolls to go away and others to come back. In a similar vein, it would be useful to divide explanations into different categories based on what moderators intended to achieve through those explanations. The topic analyses we presented in Section 5 could be a valuable guide to categorize explanations and pursue this direction. We cannot be sure whether users actually read the removal explanations they are given. In future work, we would like to control for this variable.

Our large-scale data analysis provides useful insights into how removal and explanation decisions affect future user activity. However, it is critical to investigate the in-situ practical concerns and constraints under which content moderators work. We call for researchers to study how and why moderators currently provide removal explanations and the conditions under which they work. Understanding the perspectives of moderators and building upon current work, researchers can provide design recommendations that are not just valuable for the communities but also feasible for the moderators to implement.

10 CONCLUSION

The sheer volume of content that gets posted on social media platforms makes it necessary for these platforms to rely on moderation mechanisms that are cheap and efficient. However, at this scale and speed, these mechanisms are bound to make many mistakes. Currently, platforms largely make content moderation decisions in an opaque fashion. This secretiveness causes speculations among end-users who suspect that the platforms are biased in some ways [26, 63]. Would it help platforms to instead be transparent about their processes? Would it improve community outcomes if platforms engage with users and explain the reasoning behind their moderation decisions?

In this paper, we contribute one of the first studies that explore the effects of transparency in moderation decisions on user behavior. Our research focuses on one important aspect of transparency in content moderation — the explanations about why users' submissions are removed. Our findings show that provision of removal explanations is associated with a reduction in future removals, suggesting that taking an educational, rather than a punitive, approach to content moderation can improve community outcomes. Our analysis also indicates that using automated tools to provide removal explanations is a promising approach to design for transparency without unduly increasing the work load of moderators.

ACKNOWLEDGMENTS

We thank Benjamin Sugar, Koustuv Saha, and Stevie Chancellor for their insights. We also thank the AC and reviewers of this paper for their thoughtful feedback that helped shaped this work. Jhaver and Gilbert were supported by the National Science Foundation under grant IIS-1553376.

REFERENCES

[1] Ofer Arazy, Felipe Ortega, Oded Nov, Lisa Yeo, and Adam Balila. 2015. Functional roles and career paths in Wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1092–1105.

[2] Eva Armengol, Albert Palaudaries, and Enric Plaza. 2001. Individual prognosis of diabetes long-term risks: A CBR approach. *Methods of Information in Medicine-Methodik der Information in der Medizin* 40, 1 (2001), 46–51.

[3] Sumit Asthana and Aaron Halfaker. 2018. With Few Eyes, All Hoaxes Are Deep. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 21 (Nov. 2018), 18 pages. https://doi.org/10.1145/3274290

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

[5] Brandwatch. 2019. 53 Incredible Facebook Statistics and Facts. https://www.brandwatch.com/blog/facebook-statistics/

[6] Albert Breton. 2007. *The economics of transparency in politics.* Ashgate Publishing, Ltd.

[7] Giuseppe Carenini and Johanna Moore. 1998. Multimedia explanations in IDEA decision support system. In *Working Notes of the AAAI Spring Symposium on Interactive and Mixed-Initiative Decision Theoretic Systems*. 16–22.

[8] Stevie Chancellor, Zhiyuan Jerry Lin, and Munmun De Choudhury. 2016. This Post Will Just Get Taken Down: Characterizing Removed Pro-Eating Disorder Social Media Content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1157–1162.

[9] 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. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 31 (Dec. 2017), 22 pages. https://doi.org/10.1145/3134666

[10] 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.

[11] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. 18, 3 (2016), 410–428. https://doi.org/10.1177/1461444814543163

[12] Laura DeNardis and Andrea M Hackl. 2015. Internet governance by social media platforms. *Telecommunications Policy* 39, 9 (2015), 761–770.

[13] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 153–162.

[14] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 494.

[15] Archon Fung, Mary Graham, and David Weil. 2007. *Full disclosure: The perils and promise of transparency*. Cambridge University Press.

[16] R. Stuart Geiger and David Ribes. 2010. The Work of Sustaining Order in Wikipedia: The Banning of a Vandal. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. ACM, New York, NY, USA, 117–126. https://doi.org/10.1145/1718918.1718941

[17] Homero Gil de Zúñiga, Nakwon Jung, and Sebastián Valenzuela. 2012. Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of computer-mediated communication* 17, 3 (2012), 319–336.

[18] Tarleton Gillespie. 2015. Platforms intervene. *Social Media+ Society* 1, 1 (2015), 2056305115580479.

[19] Tarleton Gillespie. 2017. Governance of and by platforms. *Sage handbook of social media. London: Sage* (2017).

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

[21] Robert Gorwa. 2019. What is platform governance? *Information, Communication & Society* (2019), 1–18.

[22] Nelson Granados and Alok Gupta. 2013. Transparency strategy: Competing with information in a digital world. *MIS quarterly* 37, 2 (2013).

[23] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.

[24] Natali Helberger, Jo Pierson, and Thomas Poell. 2018. Governing online platforms: From contested to cooperative responsibility. *The information society* 34, 1 (2018), 1–14.

[25] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[26] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2018. "Did You Suspect the Post Would be Removed?": User Reactions to Content Removals on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 33.

[27] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 31 (July 2019), 35 pages. https://doi.org/10.1145/3338243

[28] Shagun Jhaver, Larry Chan, and Amy Bruckman. 2018. The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action. *First Monday* 23, 2 (2018). http://firstmonday.org/ojs/index.php/fm/article/view/8232

[29] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 12 (March 2018), 33 pages. https://doi.org/10.1145/3185593

[30] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems* (2018).

[31] Shagun Jhaver, Pranil Vora, and Amy Bruckman. 2017. *Designing for Civil Conversations: Lessons Learned from ChangeMyView.* Technical Report. Georgia Institute of Technology.

[32] Sara Kiesler, Robert Kraut, and Paul Resnick. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design* (2012).

[33] René F Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2390–2395.

[34] David A Klein and Edward H Shortliffe. 1994. A framework for explaining decision-theoretic advice. *Artificial Intelligence* 67, 2 (1994), 201–243.

[35] Cliff Lampe, Erik Johnston, and Paul Resnick. 2007. Follow the Reader: Filtering Comments on Slashdot. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 1253–1262. https://doi.org/10.1145/1240624.1240815

[36] Cliff Lampe and Paul Resnick. 2004. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2004).

[37] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31, 2 (2014), 317 – 326. https://doi.org/10.1016/j.giq.2013.11.005

[38] David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. 2011. When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In *Fifth International AAAI Conference on Weblogs and Social Media*.

[39] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. 2017. "Could You Define That in Bot Terms"?: Requesting, Creating and Using Bots on Reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3488–3500. https://doi.org/10.1145/3025453.3025830

[40] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 369–380.

[41] Jonathan T. Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. 2013. Tea and Sympathy: Crafting Positive New User Experiences on Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 839–848. https://doi.org/10.1145/2441776.2441871

[42] Cornelia Moser. 2001. How open is' open as possible'?: three different approaches to transparency and openness in regulating access to EU documents. (2001).

[43] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 100–108.

[44] Nic Newman. 2009. The rise of social media and its impact on mainstream journalism. (2009).

[45] David B Nieborg and Thomas Poell. 2018. The platformization of cultural production: Theorizing the contingent cultural commodity. *new media & society* 20, 11 (2018), 4275–4292.

[46] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM, 93–100.

[47] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 103.

[48] Brad Rawlins. 2008. Give the emperor a mirror: Toward developing a stakeholder measurement of organizational transparency. *Journal of Public Relations Research* 21, 1 (2008), 71–99.

[49] RedditBots. 2019. autowikibot. https://www.reddit.com/r/autowikibot/wiki/redditbots

[50] Sarah Roberts. 2016. Commercial Content Moderation: Digital Laborers' Dirty Work. *Media Studies Publications* (jan 2016). https://ir.lib.uwo.ca/commpub/12

[51] Sarah T. Roberts. 2014. *Behind the screen: the hidden digital labor of commercial content moderation.* Ph.D. Dissertation. University of Illinois at Urbana-Champaign. https://www.ideals.illinois.edu/handle/2142/50401

[52] Jodi Schneider, John G Breslin, and Alexandre Passant. 2010. A content analysis: How Wikipedia talk pages are used. *Web Science* (2010).

[53] Mark Scott and Mike Isaac. 2016. Facebook restores iconic Vietnam War photo it censored for nudity. *The New York Times* (2016).

[54] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277

[55] 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.

[56] Petr Slovak, Katie Salen, Stephanie Ta, and Geraldine Fitzpatrick. 2018. Mediating Conflicts in Minecraft: Empowering Learning in Online Multiplayer Games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 595, 13 pages. https://doi.org/10.1145/3173574.3174169

[57] Nicolas Suzor. 2018. Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media + Society* 4, 3 (2018), 2056305118787812.

[58] 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.

[59] 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.

[60] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*. ACM, 1105–1112.

[61] Weiquan Wang and Izak Benbasat. 2007. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems* 23, 4 (2007), 217–246.

[62] Morten Warncke-Wang, Vladislav R Ayukaev, Brent Hecht, and Loren G Terveen. 2015. The success and failure of quality improvement projects in peer production communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 743–756.

[63] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* (2018).

[64] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 194 (Nov. 2018), 23 pages. https://doi.org/10.1145/3274463

Received April 2019; revised June 2019; accepted August 2019