Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

Exploring the Magnitude and Effects of Media Influence on Reddit Moderation

Hussam Habib, Rishab Nithyanand

University of Iowa
{hussam-habib, rishab-nithyanand}@uiowa.edu

Abstract

Most platforms, including Reddit, face a dilemma when applying interventions such as subreddit bans to toxic communities — do they risk angering their user base by proactively enforcing stricter controls on discourse or do they defer interventions at the risk of eventually triggering negative media reactions which might impact their advertising revenue? In this paper, we analyze Reddit's previous administrative interventions to understand one aspect of this dilemma: the relationship between the media and administrative interventions. More specifically, we make two primary contributions. First, using a mediation analysis framework, we find evidence that Reddit's interventions for violating their content policy for toxic content occur because of media pressure. Second, using interrupted time series analysis, we show that media attention on communities with toxic content only increases the problematic behavior associated with that community (both within the community itself and across the platform). However, we find no significant difference in the impact of administrative interventions on subreddits with and without media pressure. Taken all together, this study provides evidence of a media-driven moderation strategy at Reddit and also suggests that such a strategy may not have a significantly different impact than a more proactive strategy.

1 Introduction

Strict platform moderation is rarely a first-order priority for newly developed online platforms. After all, the early adopters are often homogenous with a shared goal of nurturing the community. However, as platforms become more mainstream and contend with a large and consistent influx of new users, each with their own ideals and agendas, effective and timely platform moderation becomes paramount to maintaining a civil community. Despite the absence of any legal consequences for not effectively moderating platforms, effective moderation is often tied to another goal of the platform – avoiding negative media attention so that the platform remains appealing to advertisers who ultimately are their primary revenue source. Complicating matters, as economically rational actors, platforms need to also account for the loss in users and popularity as a result of platform-wide moderation decisions. This suggests that the effectiveness of moderation on platforms might be tied to the media's coverage of their failures as well as the costs of moderation decisions on platform activity. The research presented in this paper investigates these relationships on Reddit.

Reddit's history with media-driven moderation decisions. The story of platform moderation on Reddit appears similar to the evolutionary trend described above. In its early days, Reddit was celebrated as the bastion of free speech due to its minimal moderation and interference. However, as its popularity grew over the years it found itself being criticized by outsiders and the media for its lack of effective moderation. There have been numerous examples of Reddit's moderation decisions being driven by media pressure including *r/The_Donald* which was only shutdown after widespread reporting in the media for the violent and incivil political discourse it facilitated, *r/TheFappening* which was shutdown only after reports of its role as the facilitator in the distribution of involuntary pornography involving celebrities, *r/CoonTown* which was not banned during Reddit's first purge of 'hateful' subreddits until criticism from mainstream media outlets, and most notably – *r/jailbait*. The *r/jailbait* subreddit was one of the earliest cases of Reddit moderation being performed only in reaction to media attention (Centivany 2016). The subreddit featured provocative pictures of minors and due to the lack of any rules against it, Reddit condoned its existence even awarding it the *voted best subreddit of 2008* (Chen 2012). In September 2011, in a segment on his show, Anderson Cooper of CNN brought *r/jailbait* to wider attention heavily criticizing Reddit on hosting such content. Following more negative attention, the subreddit was finally banned by administrators in October 2011. This extremely delayed intervention led many, including the creator of the subreddit, to speculate that the closing of the subreddit was only direct response to the negative attention (Tufekci 2012). This speculation was further validated by the lack of administrative action against other 'bait'-type subreddits such as *r/asianjailbait*. Taken together, these anecdotes suggest that media pressure does impact Reddit's moderation decisions. The extent of this impact is the subject of this research.

The consequences of media-driven moderation. Aid from the media, users, and outsiders helps platforms conduct effective moderation. By bringing attention to egregious content and highlighting gaps in its policies, such attention can

help platforms perform difficult administrative actions and evolve their content policy. However, over reliance on the media for moderation may lead to several problems including: inconsistent enforcement of policies owing to the medias own inconsistent coverage of problematic content, delayed moderation decisions due to the fact that action is taken only after a violation is egregious enough to warrant coverage by the media, and finally the normalization of problematic behaviours since media coverage may only focus on the egregious violations while ignoring the problematic behaviours leading up to it. Our work seeks to uncover whether these consequences are also experienced by Reddit.

Our hypotheses. This research seeks to highlight the extent to which media reporting drives Reddit moderation decisions and the consequences it subsequently faces. Specifically, we explore the following hypotheses.

H1: In communities with toxic content, Reddit's administrative interventions for violating the content policy related to toxicity occur because of media pressure. (§2) We test the validity of this hypothesis by checking if media pressure generated by a subreddit (quantified from negative media coverage of a subreddit) mediates the relationship between its measured levels of toxicity and administrative interventions for violating the content policy related to toxic content. Our analysis shows that measures of media pressure and internal pressure completely explains any relationship between measured levels of toxicity and administrative interventions for violating the content policy related to toxic content. This suggests a reactionary moderation strategy.

H2: Prior media attention on communities which receive interventions for toxic content: (1) increases the prevalence of problematic activity on the platform and (2) reduces the effectiveness of the issued interventions. (§3) We now focus on subreddits which: (1) received an administrative intervention for violating the content policy regarding toxic content and (2) received negative media attention prior to the administrative intervention. For these subreddits, we conduct an interrupted time series analysis to understand the platform-wide increase of problematic activity related to the toxic community as a consequence of: (1) the media pressure they receive and (2) the administrative intervention. Our analysis shows that media pressure and interventions both increase the levels of problematic activity. However, we find that the effects of the intervention are not statistically different from the effects observed by the communities which received no media pressure prior to their intervention — i.e., interventions are not less effective when they are preceded by media attention on the targeted community.

2 Are Reddit's Administrative Interventions Influenced by Media Pressure?

Overview. In this section, we explore the relationship between media pressure and administrative interventions in the context of toxic Reddit communities. Our focus is solely on subreddits which were banned or quarantined for violating the content policy related to toxicity[1]. Our hypothesis is that: *(H1) In communities with toxic content, Reddit's administrative interventions for violating the content policy related to toxicity occur because of media pressure.*

Put another way, we wish to test: when the toxicity of two subreddits are controlled for does the subreddit garnering more negative media attention become more likely to receive an administrative intervention for violating the content policy related to toxic content? If this hypothesis is valid, it would suggest that Reddit employs a reactionary administrative strategy which delays administrative interventions for toxic communities until media pressure forces action. To validate our hypothesis, we conduct three observational analysis. First, we explore any significant characteristics of intervened subreddits. Significant differences in distribution of media attention between intervened and active subreddits within the 3k most popular subreddits would suggest media attention as an important characteristic for intervened subreddits. More so, significantly higher media attention towards intervened subreddits compared with active subreddits within a set of subreddits controlled for toxicity would suggest media attention to be a strong characteristic and predictor for interventions and therefore validate the basis of our hypothesis. Next, we test whether there is a mediation relationship between toxicity and interventions. We test the validity of our hypothesis by proposing a mediation model. We propose media attention as a mediator between toxicity and interventions. If our mediation model yields significant relationships we can validate our hypothesis of a relationship between media pressure and interventions. Finally, we expand our mediation model in an attempt to construct a more holistic portrayal of the administrative interventions taken on Reddit. By including influence and attention from internal and external online users we are able to strengthen the relationships and further validate our hypothesis.

2.1 Methods and Datasets

Quantifying subreddit toxicity. We quantify the toxicity of a subreddit as the percentage of toxic content (posts and comments) present in a subreddit. The use of this metric is supported by comments from Reddit administrators. For example, in response to a question demanding transparency in their administrative interventions for violations of Rule 1, *u/spez* (an administrator and co-founder of Reddit) indicated that "high ratio" of hateful content was a major criteria for interventions. [2] This also motivates our study of the relationship between toxicity, media pressure, and administrative interventions for toxic content. In order to identify toxic content, we leverage the Perspective API[3] — a Google-owned

[1] "Rule 1: Remember the human. Reddit is a place for creating community and belonging, not for attacking marginalized or vulnerable groups of people. Everyone has a right to use Reddit free of harassment, bullying, and threats of violence. Communities and users that promote hate based on identity or vulnerability will be banned."

[2] https://www.reddit.com/r/announcements/comments/hi3oht/update_to_our_content_policy/fwe83at/

[3] https://www.perspectiveapi.com/

tool for identifying online toxic content. We use the Perspective API to identify the percentage of all toxic comments and posts on a subreddit. Perspective API provides the probability of a text being percieved as toxic. For our analysis, we select probability of 0.5 as the threshold for a text being considered toxic. This threshold has been used by prior studies validating Perspective API (Pavlopoulos et al. 2019) and has also been used by Perspective's team to evaluate the API (Obadimu et al. 2021). Using this threshold we identify toxic submissions within a community and count their occurrences. We quantify the toxicity of a subreddit as $T(s) = \frac{\#\text{ toxic comments}\in s + \#\text{ toxic posts}\in s}{\#\text{ comments}\in s + \#\text{ posts}\in s}$. We note that the Perspective API has been validated for use with Reddit and has been leveraged to quantify subreddit toxicity in several previous studies (Mittos et al. 2020; Zannettou et al. 2020) and has also been used as a plugin to aid moderation [4].

Quantifying negative media attention as media pressure. We seek to quantify negative attention towards subreddits from popular media outlets. We start by identifying the number of published media articles that mention a subreddit in a negative or critical tone. We refer to each of these articles as a 'negative media mention'. To measure the negative media mentions for a subreddit, we use the MIT media cloud API[5] to obtain articles mentioning the subreddit's name. We restrict our analysis to articles from US 'top sources' and 'mainstream media' sites as categorized by the MIT media cloud. We focus the remainder of our analysis only on articles published between 01/2015 and 04/2020. Furthermore, for subreddits which received an intervention we only include pre-intervention articles (i.e., those published up to the month prior to the intervention). We do this to ensure the exclusion of articles which report the occurrence of an intervention. Next, for each article, we use the entity-level sentiment analysis API from the Google NLP platform[6] to measure the sentiment towards the subreddit. Articles which include negative sentiments towards the subreddit are counted as negative media mentions. We quantify the 'media pressure' towards a subreddit s as $P_{media}(s) = \frac{\text{negative media mentions of } s}{\text{total media mentions of } s+L}$, where L is the Laplace smoothing constant and is set to 10. This metric captures the frequency of negative media mentions relative to all media mentions received by a subreddit. The presence of 'total media mentions' and the smoothing constant L ensures that the quantified media pressure (P_{media}): (1) is not identical for two subreddits a and b, where a and b have similarly high ratio of negative:total media mentions but differ significantly in their raw number of total media mentions and (2) is not identical for two subreddits a and b, where a and b have the same number of negative media mentions but significantly different total media mentions. We specifically selected $L = 10$ since after manual verification it appeared to achieve our above goals without introducing noise that would erase any differences between the attention received by subreddits. Our rationale for using percentage of negative media mentions to represent media pressure is that it removes popularity as a confounding factor. Our analysis using raw count of negative media mentions as media pressure showed media pressure to be directly proportional to popularity. Rather than confounding popularity and negative media mentions in a single variable (i.e. raw negative media mentions count) we separate these two variables and use them separately in our analysis.

Identifying subreddits receiving administrative interventions for violating the content policy related to toxicity. Reddit's content policy requires communities (i.e., subreddits) to adhere to eight rules [7]. Violation of these rules are meant to result in administrative interventions by Reddit. In this paper, we focus on the communities found to be in violation of *Rule 1* (commonly referred to as the anti-toxicity policy). We focus on this rule specifically because it was the subject of the most administrative interventions during the period of this study (from 01/2015 to 04/2020). Further, anecdotes of media-driven interventions appear to occur most frequently for communities found to be violating this policy, perhaps due to its subjective nature.

Reddit's administration has one of two administrative actions they can take on a violating subreddit: banning or quarantining. Bans result in the closure and deletion of the subreddit and all associated posts. Quarantines are less severe and result in the removal of the subreddit from the search results and other efforts to limit the growth and visibility of the subreddit. In our work, given our goal of identifying the role of media pressure in any administrative intervention, we do not distinguish between the two. In order to identify subreddits banned/quarantined for violations related to the anti-toxicity content policy, we scraped the homepages of all subreddits and identified the ones marked as banned or quarantined for violations of the policy [8]. In total, 120 of the 535 subreddits which received an administrative intervention from 01/2015 to 04/2020 were targeted for the violation of this policy. In the remainder of this paper we broadly use the term 'administrative interventions' to refer to administrative interventions whose stated reason was a violation of the anti-toxicity content policy.

Next, we identify the date of interventions. To this end, we use the method used by Habib et al. to obtain the dates of interventions (Habib et al. 2022). First, to get the banning date of a subreddit, we search for the last submitted comment/post on the subreddit. Since banning results in complete closure of the subreddit, we consider the date of the last submission as a proxy for the date of banning. Next, to determine the date of quarantine we use a combination of methods. First we scrape *r/reclassified* for any mention of the subreddit. *r/reclassified* is a crowd sourced collection of interventions and their dates on Reddit. If a submission mentioning the quarantine of the subreddit we consider date of the submission the quarantine date. Additionally, we also search for any posts pinned in the subreddit mentioning the

[4]https://www.perspectiveapi.com/case-studies/

[5]https://mediacloud.org/

[6]https://cloud.google.com/natural-language

[7]https://www.redditinc.com/policies/content-policy

[8]Reddit provides specific violations in the subreddit homepage when a ban occurs. See www.reddit.com/r/The_Donald as an example.

Dataset	Label	Subreddits	Avg. Toxicity (T)	Negative Media Mentions	Avg. Media Pressure % (P_{media})
$\mathfrak{D}_{3K}$	Intervened	29	**23%**	407	**18.7**
	Active	2971	**8%**	644	**0.9**
$\mathfrak{D}_P$	Intervened	120	**24%**	463	**5.6**
	Active	120	**6%**	8	**0.2**
$\mathfrak{D}_T$	Intervened	120	24%	463	**5.6**
	Active	120	24%	31	**1.9**
All	Intervened	120	**24%**	463	**5.6**
	Active	3211	**9%**	683	**0.9**

Table 1: Characteristics of the datasets used in this study. Bold values indicate a statistically significant ($p < .05$) difference between the attributes for the intervened and active groups in the corresponding dataset.

quarantining of the subreddit. If found, the date of the post is considered to be the date of quarantining. Finally, due to the low number of interventions, we also manually validate the dates and find them to be accurate.

Datasets. Our data was gathered using Pushshift (Baumgartner et al. 2020) and comprised of all the comments and posts made on Reddit during the period from 01/a015 to 04/2020. In total, this included 5B comments and 684M posts from 39M unique users. For the analysis presented in this section, we use this data to construct three different datasets that are described below. The characteristics of each dataset are illustrated in Table 1.

Dataset of most active subreddits ($\mathfrak{D}_{3K}$): This dataset contains all the content (comments, posts, and media mentions) associated with the 3000 most active subreddits between 01/2015 and 04/2020. We define activity as the average number of monthly comments and posts made on the subreddit. For subreddits which receive an administration intervention (referred to as 'intervened subreddits'), this average is computed only over the post-creation and pre-intervention months that occurred within the period from 01/2015 to 04/2020. For subreddits without an administrative intervention (referred to as 'active subreddits'), this average is computed over all the post-creation months that occurred between 01/2015 and 04/2020. In total, this dataset contained 29 intervened subreddits and 2971 active subreddits.

Dataset of popularity-controlled subreddits ($\mathfrak{D}_P$): This dataset contains all the content associated with all 120 subreddits which received an intervention for violating the 'anti-toxicity' policy between 01/2015 and 04/2020. For each of these intervened subreddits, we also include content associated with an active subreddit that has the most similar popularity. Popularity is measured by the average number of active users on the subreddit each month (i.e., the number of unique users making posts or comments on the subreddit). As above, this average is only computed over the subreddit's post-creation and pre-intervention period between 01/2015 and 04/2020. The Kolomogorov-Smirnoff goodness-of-fit test is a non-parametric test of equality between two sets. The null hypothesis states that the samples were sampled from the same distribution. Our test on the popularity of active and intervened subreddits in $\mathfrak{D}_P$ fails to reject the null hypothesis, therefore, we consider them to have similar distribution of popularity.

Dataset of toxicity-controlled subreddits ($\mathfrak{D}_T$): This dataset also contains all the content associated with our 120 intervened subreddits. However, the active subreddits in this dataset are obtained by matching each intervened subreddit with the non-intervened subreddit having the most similar toxicity (T) score. Similar to the previous datasets, toxicity scores were only computed over the post-creation and pre-intervention period between 01/2015 and 04/2020. In our dataset, we observe intervened subreddits such as *r/TheRedPill* (T=45), *r/Incels* (T=28), *r/uncensorednews* (T=20), were matched with *r/asktrp* (T=46), *r/terfisaslur* (T=28), *r/TumblrInAction* (T=20). A Kolmogorov-Smirnoff goodness-of-fit test shows that the distributions of toxicity scores observed within the two groups of subreddits in this dataset are similar.

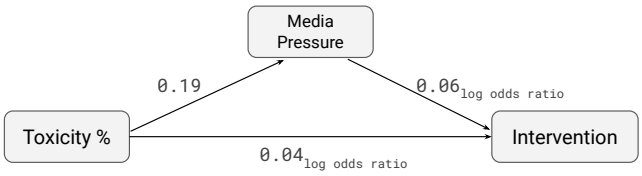
Validity of results and dataset choice. We note that since we have abundance of active subreddits to match our intervened subreddits with, to ensure validity of our results, we repeated our analysis a total of three times with different samples of $\mathfrak{D}_P$ and $\mathfrak{D}_T$. All results reported in this paper remained consistent across all three iterations. We purposely choose to construct two independent datasets $\mathfrak{D}_P$ and $\mathfrak{D}_T$ rather than a single dataset $\mathfrak{D}_{P,T}$ which controls for both popularity and toxicity. This is done because we intend to use: (1) $\mathfrak{D}_P$ to specifically examine how toxicity interacts with media pressure and interventions and (2) $\mathfrak{D}_T$ to determine why subreddits with with similar toxicity might differ in their intervention status. Using a single dataset ($\mathfrak{D}_{P,T}$) for this analysis would remove our ability to use toxicity as an independent variable and popularity as a moderating variable, both of which are key to our analysis and subsequent mediation models, due to the absence of any variance in the toxicity and popularity metrics within the dataset.

2.2 Analysis and Results

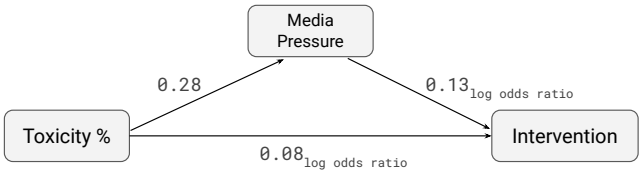
Overview of analyses. We conduct three observational experiments to better understand the influence of toxicity (T) and media pressure (P_{media}) on each other and on administrative interventions for toxic content. Each experiment builds on the previous and eventually provides a test for *H1*.

Analysis 1: What are the characteristics of intervened subreddits? We begin our analysis by simply comparing the distributions and means of toxicity scores (T) and media pressure scores (P_{media}) for active and intervened subreddits in each of our three datasets ($\mathfrak{D}_{3K}$, $\mathfrak{D}_P$, and $\mathfrak{D}_T$).

Differences in distributions of P_{media} and T: In all three datasets, we find that the distribution of P_{media} scores is statistically significantly ($p < 0.05$)different for active and intervened subreddits. Similarly, we see statistically significant differences in T scores for active and intervened subreddits in $\mathfrak{D}_{3K}$ and $\mathfrak{D}_P$ (not in $\mathfrak{D}_T$ which specifically controls for toxicity across the two groups). Looking at the means, we



(a) Mediation effects observed on $\mathfrak{D}_{3K}$. The direct effect ($T \rightarrow I$) is .04 (log odds) and the indirect effect ($T \rightarrow P_{media} \rightarrow I$) is .13 (log odds). Both effects are statistically significant ($p < .05$).



(b) Mediation effects observed on $\mathfrak{D}_P$. The direct effect ($T \rightarrow I$) is .08 (log odds) and the indirect effect ($T \rightarrow P_{media} \rightarrow I$) is .18 (log odds). Both effects are statistically significant ($p < .05$).

Figure 1: A preliminary mediation analysis: Does P_{media} mediate the relationship between T and I? Solid lines indicate statistically significant effects. Values indicate correlation coefficients between variables.

see that on average and across all three datasets, intervened subreddits have over $6\times$ higher P_{media} and $2.5\times$ higher T scores than non-intervened subreddits. Interestingly, we find that even when toxicity scores are controlled ($\mathfrak{D}_T$), the mean P_{media} score of intervened subreddits is nearly $3\times$ higher than their equally toxic non-intervened counterparts. These results suggest that P_{media} may be more predictive (than T) of administrative interventions. However, we note that only 43 of the 120 intervened subreddits had received media attention prior to their intervention. This suggests that P_{media} is not the only influence or predictor of an intervention. A full breakdown of T and P_{media} scores for each group and dataset is provided in Table 1.

Predictive powers of T and P_{media} on administrative interventions: Next, we construct a logistic regression model that uses T and P_{media} to predict administrative interventions. Using subreddits in $\mathfrak{D}_P$, we find that both variables are statistically significant ($p < 0.05$) predictors of administrative interventions with identical odds ratios of 4% — i.e., all else equal, a unit increase in T or P_{media} increases the odds of a subreddit receiving an intervention by 4%. This result suggests that the administrative interventions may be influenced by both T and P_{media} and therefore justifies further investigation into their relationship with each other and with administrative interventions.

Analysis 2: Does media pressure mediate the relationship between toxicity and administrative interventions? Our previous results show that P_{media} and T scores are predictive of administrative interventions on subreddits. Further, we find that T is a statistically significant ($p<0.05$) predictor of P_{media}. Both these findings suggest the possibility of T having its relationship with administrative interventions mediated by P_{media} — i.e., the effects of T on administrative interventions may be explained by T's effects of P_{media}. We explore this with a mediation model using $\mathfrak{D}_P$ and $\mathfrak{D}_{3K}$.

Primer on mediation analysis. Mediation analysis is a standard toolkit for explaining the underlying mechanism of the relationship between two (often) correlated variables — an independent variable (IV) and a dependent variable (DV) (Baron and Kenny 1986). Simply put, a mediation model tests whether the predictive power of IV on the value of DV is reduced when some mediation variable MV is introduced in the regression. If this reduction is statistically significant, we say that MV mediates the relationship between IV and DV — i.e., the relationship between IV and DV may be explained by the effect of IV on MV (and MV to DV). We say that the MV completely mediates the relationship between IV and DV if after the inclusion of MV, the direct effect of IV on DV becomes insignificant — i.e., all of IV's effect on DV is explained by its relationship with MV (and MV with DV).

Testing P_{media} as a mediation variable. We now consider a mediation model which uses T as the independent variable, an indicator variable (I) to represent administrative interventions ($I_s = 1$ if the subreddit s received an administrative intervention and $I_s = 0$ otherwise) as the dependent variable, and P_{media} as the mediating variable. We conduct our mediation analysis on the $\mathfrak{D}_{3K}$ and $\mathfrak{D}_P$ datasets. Note that the $\mathfrak{D}_T$ dataset cannot be used since it explicitly controls for toxicity (the independent variable in our model) which would forcibly remove any effects from $T \rightarrow I$. Our models, the direct $T \rightarrow I$ effects, and indirect $T \rightarrow P_{media} \rightarrow I$ are illustrated in Figure 1a (for dataset $\mathfrak{D}_{3K}$) and Figure 1b (for dataset $\mathfrak{D}_P$). In both cases, we see that the mediation occurring through P_{media} is statistically significant ($p < 0.05$) and that the indirect effect from $T \rightarrow P_{media} \rightarrow I$ is substantially higher than the direct effect from $T \rightarrow I$. In the $\mathfrak{D}_{3K}$ dataset, a unit increase in T will result in a 4% increase in the odds of an intervention solely due to T and a 14% increase in the odds of an intervention because of the effect of T on P_{media}. Similarly, in the $\mathfrak{D}_P$ dataset, a unit increase in T will result in a 8% increase in the odds of an intervention solely due to T and a 19% increase in the odds of an intervention because of the effect of T on P_{media}. Thus, we can conclude that P_{media} has a partial mediating effect on $T \rightarrow I$.

Analysis 3: Further exploring the relationship between toxicity and administrative interventions? We now seek to build a complete model to explain the relationships between T, P_{media}, and I. Our initial analysis which shows that $P_{media} > 0$ only in 43 communities and the existence of only a partial mediation by P_{media} suggests the possibility for additional influences between $T \rightarrow I$. We explore this possibility by incorporating several third variables into our model: (as moderators) subreddit popularity, subreddit topic, and subreddit profitability; and (as mediators) internal pressure and external pressure. We use this model to analyze the $\mathfrak{D}_P$ dataset since it is the most complete and allows effects of toxicity.

Including moderating variables. *Subreddit popularity* is a measure of the average number of active contributors to a subreddit per month. In our model, we specifically investigate how subreddit popularity may influence the relationship between $T \rightarrow P_{media}$ and $T \rightarrow I$. The inclusion of popularity allows us to investigate whether the $T \rightarrow P_{media}$ or $T \rightarrow I$ effects are significant and stronger for subreddits of different popularity levels. Next, to check if the presence of specific topics elicited more negative attention, we included *subreddit topic* as a moderator. We used TF-IDF to create keyword vectors for each subreddit and applied k-means clustering over these vectors to identify groups of similar subreddits. $k = 8$ was selected after manual verification. We manually label each of the eight clusters with one of the following topics: sports, politics, forums, memes, gore, porn, games and health. Each subreddit inherits the topic of the cluster it is clustered in. In our analysis of subreddit topics, we were specifically interested in studying the effects of subreddit topic on $T \rightarrow P_{media}$.

Finally, we introduce a *subreddit profitability* variable as a moderator. In addition to advertising revenue, Reddit is supported by Redditors' purchase of Reddit coins [9]. These coins allow Redditors to reward high-quality posts and comments with awards and reactions. We estimate the amount of non-advertising revenue generated by a subreddit by tracking the average number of awards donated to posts and comments on the subreddit each month. This estimate is used as a proxy for subreddit profitability. In our analysis, we are specifically interested in understanding how subreddit profitability moderates the relationships between $T \rightarrow I$ and $T \rightarrow P_{media} \rightarrow I$.

Introducing mediating variables. Our complete model also seeks to understand if the influence of pressure on administrators originating from within the Reddit community (internal non-media pressure or P_{int}) and pressure on administrators originating from other platforms (external non-media pressure or P_{ext}) may mediate $T \rightarrow I$. In order to measure P_{int} for a subreddit, we obtain all pre-intervention and post-creation comments made on Reddit between 01/2015 and 04/2020 which mention the specific subreddit in a negative sentiment. Then we set $P_{int} = \frac{\text{negative comment mentions of } s}{\text{total comment mentions of } s+L}$. We quantify external pressure for a subreddit by gathering all pre-intervention and post-creation tweets made on Twitter between 01/2015 and 04/2020 which mention the specific subreddit in a negative sentiment. Same as before, we set $P_{ext} = \frac{\text{negative Twitter mentions of } s}{\text{total Twitter mentions of } s+L}$. We select Twitter as our proxy for external pressure due to its ubiquity, size, and prominence in the activist community.

Pathways to administrative interventions. The results of our complete mediation analysis are illustrated in Figure 2. First, we see that P_{int} *and* P_{media} *completely mediate the relationship between* T *and* I. The inclusion of P_{int} and P_{media} as mediators between T and I cause the direct effect $T \rightarrow I$ to become insignificant. This allows us to conclude that any effect that T has on I is only because of its effect on P_{int} and P_{media}. Analyzing the pathways to influence I, we see that

[9] https://reddithelp.com/hc/en-us/articles/360043034252

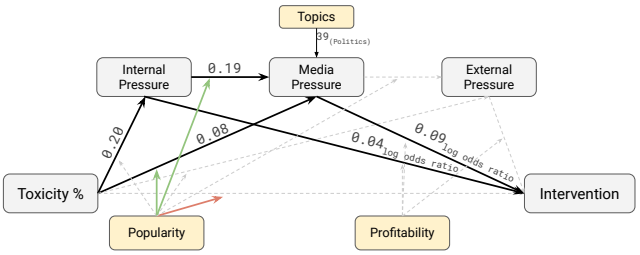


Figure 2: A complete mediation analysis. Solid lines indicate statistically significant effects and dashed lines indicate insignificant effects. Values indicate the correlation coefficients between variables. Variables in yellow boxes were included as moderators. Green and red arrows indicate a statistically significant amplifying and dampening moderation effect, respectively.

all the indirect effects through P_{int} and P_{media} are statistically significant ($p < 0.05$). Of these paths, the indirect effect from $T \rightarrow P_{int} \rightarrow P_{media} \rightarrow I$ is found to be the strongest with a unit increase in T resulting in a 2.3% increase in the odds of an intervention through this path. Smaller effects are observed on the $T \rightarrow P_{media} \rightarrow I$ and $T \rightarrow P_{int} \rightarrow I$ paths where a unit increase in T increases the odds of intervention by 1.2% and 1.7%, respectively. Our model also shows that *subreddit popularity moderates relationships with* P_{media} *and the effect of* T *on* I. Specifically, we find that subreddit popularity is a statistically significant ($p < 0.05$) amplifier in the $T \rightarrow P_{media}$ and $P_{int} \rightarrow P_{media}$ relationships — i.e., the influence of T and P_{int} on P_{media} is higher for more popular subreddits than less popular ones when toxicity or P_{int} are controlled for. This intuitively makes sense — after all, media outlets' interest in covering a subreddit is likely related to the popularity of the subreddit. We also find that the effect of $T \rightarrow I$ reduces as popularity increases and this effect, although small, becomes statistically significant for subreddits with popularity in the 84th percentile and higher. This finding suggests a marginal hesitation to apply administrative interventions to more popular subreddits when toxicity is controlled for.

We note that *Subreddit topic, subreddit profitability, and external pressure yielded no statistically significant influences in our model.* This suggests that subreddit topic generally does not influence media pressure when toxicity is controlled. However, we found that the specific topic of 'politics' was a significant moderator between toxicity and media pressure ($p < 0.05$), suggesting that political subreddits are more likely to get negative media pressure as a result of high toxicity. Finally, we found that subreddit profitability never influences administrative interventions (through the direct or indirect path), external pressure is not influenced by media pressure or toxicity, and external pressure does not influence administrator intervention decisions.

Takeaways. Our results confirm our original hypothesis that in communities with toxic content (T), Reddit's administrative interventions for violating the content policy related to toxicity occur (I) because of media pressure (P_{media}).

However, our analysis shows that the mediating effect of media pressure ($T \rightarrow P_{media} \rightarrow I$) does not completely explain the relationship between T and I. We find that incorporating the effects of internal pressure (P_{int}) in our model yields two additional statistically significant pathways: $T \rightarrow P_{int} \rightarrow I$ and $T \rightarrow P_{int} \rightarrow P_{media} \rightarrow I$ whose addition completely explains any effect from T to I. Taken all together, this suggests a reactionary moderation strategy in which any administrative interventions handed out for toxic content are driven by internal pressure from Redditors and media pressure from negative media attention.

3 What Are the Consequences of a Media-Driven Intervention Strategy?

Overview. Thus far, our analysis has demonstrated that administrative interventions for toxic content are largely driven by internal and media pressure. We now seek to understand whether such reactionary administrative intervention strategies are effective at curbing problematic activities (that are associated with the target subreddit) across the platform. Our hypothesis is that: *(H2) Prior media attention on communities which receive interventions for toxic content: (1) increases the prevalence of problematic activity on the platform and (2) reduces the effectiveness of the issued interventions.* This hypothesis was formulated based on prior social science literature (Phillips 2018; Marwick and Lewis 2017) and historical accounts of media attention resulting in increased traffic to problematic subreddits (Centivany 2016). We test this hypothesis using an interrupted time series analysis to check whether community-specific increases in user growth rates and platform-wide increases in problematic discourse (that is associated with the intervened subreddit) occur as a consequence of the media pressure they receive and the administrative intervention they are handed out. If part (1) of this hypothesis is valid, it suggests that media attention on a problematic community increases the prevalence of the problematic discourse within the community and across the platform. If part (2) of this hypothesis is valid, it suggests that media-driven interventions are less effective at curbing the spread of problematic discourse across the platform than their non-(media)impacted counterparts.

3.1 Methods

Tracking growth rates within an intervened subreddit. For each of the 120 intervened subreddits in our dataset, we compute the daily 'growth' of the community. This growth for a given day is computed by counting the number of unique users that made their first contributions to the community during that day. Put another way, this measures the number of new contributors to a community each day. This metric is used to identify the impact that media coverage has on intervened communities. Note that this metric cannot be used to identify the impact of administrative interventions since the community itself becomes inactive after the intervention. By measuring the growth of a subreddit every day, then performing interrupted time series analysis with the interruption as the day media attention is given to the subreddit, we can observe any rate of change in growth immediately after the interruption. We hypothesize that media attention results in an increase in the growth rate of a community — i.e., more users begin directly participating in the problematic discourse as a result of the media attention.

Identifying and tracking problematic discourse of an intervened community. For our analysis, we seek to measure if the "problematic discourse" of an intervened community begins to spread across the platform as a consequence of media attention and administrative interventions on the community. This requires us to identify and track this problematic discourse. We do this by using the vocabulary unique to the intervened subreddit as a proxy for the problematic discourse occurring on it. By tracking the prevalence of this unique vocabulary on other subreddits, we effectively measure the adoption of the problematic vocabulary across the platform. We derive the unique vocabulary associated with a community using the *Sparse Additive Generative Model (SAGE)* (Eisenstein, Ahmed, and Xing 2011). SAGE extracts keywords that are unique to our intervened subreddit relative to a set of reference subreddits (the default subreddits in our case). Using this process, we extract the 500 most unique keywords for each intervened community and manually confirming their relevance and specificity to the intervened community. This vocabulary consists of keywords for which the likelihood of being in the intervened community is greater than the likelihood of occurring in the reference subreddits by at least 2.3 standard deviations. For example, (*fakecel, truecel, femoid...*) were the extracted from *r/Incels* and (*eyethespy, thankq, ibor...*) were extracted from *r/greatawakening*. We then count the frequency of occurrence of these keywords across the remainder of the platform (i.e., excluding the intervened subreddit itself) for each day. We note that this approach has also been used in prior work identifying in-group vocabulary and measuring the spread of ideologies (Chandrasekharan et al. 2017). By identifying vocabulary specific to the intervened community compared to the default subreddits, we can use this vocabulary as a proxy for the discourse happening in this subreddit. Next, we measure the prevalence of the vocabulary on all of the Reddit for all of the days in our analysis. Using our interrupted time series analysis, we aim to determine whether an intervention caused significant increase in the time series. A significant increase in the time series would suggest more engagement in the problematic discourse found on the community on Reddit.

Identifying the effects of media coverage and administrative interventions using interrupted time series analysis. Interrupted time series analysis test for any significant changes in the rate of a given variable after an event of interest occurs. It models the time series prior to the event and forecasts the time series after the event. If there is a statistically significant difference in the forecasted and actual time series after the event, the event is said to have an effect on the variable being tracked. We run three interrupted time series analyses for each of our 120 intervened communities: (1) for communities experiencing pre-intervention media attention, using community growth rates as the variable and media at-

tention as the event, (2) for communities experiencing pre-intervention media attention, using the growth in prevalence of intervened subreddit vocabulary on other subreddits as the variable and media attention as the event, and (3) for all intervened subreddits, using the prevalence of intervened subreddit vocabulary on other subreddits as the variable and the administrative intervention as the event. All together, these analyses will identify the impact of media attention on problematic discourse and activity.

Validating results with controlled analysis. In addition to the interrupted time series, we also perform a comparative control-treatment analysis to dispel alternative hypotheses. We compare the changes in user growth and vocabulary growth between treatment and control subreddits. To this end, we construct two sets of control groups while using subreddits which experienced any media attention as the treatment group. For the first control group, we identify (for each treatment subreddit), from the set of subreddits which had no media attention, subreddits which most closely matched the pre-media attention topic (using TF-IDF vectors as described in Section 2.2 followed by measuring cosine similarity between the candidate control and treatment subreddit) and daily user growth rate of the treatment subreddits. Comparing the post-media attention user growth rates of this group with our treatment allows us to identify the impact of media attention on the community growth rate when topic is controlled. For the second control group, we identify (for each treatment subreddit), from the set of subreddits which had no media attention, subreddits which had the most similar pre-media attention vocabulary and user growth rates compared to the treatment subreddit. Comparing the post-media attention vocabulary growth rates of this control group with our treatment group allows us to identify the impact of media attention on the spread of community vocabulary when user growth rates are controlled. We identify subreddits with similar user and vocabulary growth rates for the above construction of the control groups by representing them as time series vectors (each entry corresponds to the growth rate for a particular day) and then computing similarities between the vectors of candidate control subreddits and the treatment subreddits using Dynamic Time Warping (DTW) (Müller 2007). Following the construction of the above control and treatment groups, we compute the percentage change experienced (in user and vocabulary growth) by each treatment subreddit and its corresponding control subreddit. Aggregating percentage changes experienced by treatment subreddits and control subreddits separately, we use a t-test to determine whether the difference between them is significant. A significant difference with the treatment aggregate experiencing higher average percentage increase would suggest the effect of our treatment on user growth or vocabulary growth is significant.

Comparing the effectiveness of media-driven interventions with interventions not impacted by media coverage. Finally, we split our dataset of 120 intervened subreddits into those which generated media pressure (treatment) and those that did not (control). We then compare the effects of administrative interventions on these two groups with a focus on the percentage increase in occurrences of their vocabulary on other subreddits. A statistically significant difference in this variable between the two groups would suggest the possibility that media-driven interventions are less effective at curbing the spread of problematic discourse and ideologies.

Topic	Communities	User growth (post-media)	Voc. growth (post-media)	Voc. growth (post-int.)
Manosphere	Incels	**+463%**	**+202%**	**+201%**
	Braincels	**+1443%**	**+191%**	**+185%**
	shortcels	-105%	**+1206%**	**+1131%**
	TheRedPill	**+38%**	**+221%**	**+213%**
	MGTOW	**+209%**	**+245%**	**+435%**
	JustBeWhite	**-**	**-**	**+219%**
	CringeAnarchy	**+110%**	**+102%**	**+115%**
QAnon	greatawakening	**+3491%**	**-31%**	+41%
	uncensorednews	**+232%**	**-44%**	**-2%**
	TheNewRight	**+51%**	+14%	+31%
	The_Donald	**+530%**	**+117%**	+29%
	new_right	**+419%**	-88%	+29%
	Mr_Trump	**+29%**	**-82%**	**+291%**
Extremist groups	The_Donald	**+512%**	**+117%**	+29%
	DebateAltRight	-	-	+114%
	WhiteRights	**+39%**	-29%	**+41%**
	Physical_Removal	**+353%**	-31%	-42%
	RightwingLGBT	**+555%**	+24%	+131%
	european	-	-	**+331%**
	The_Europe	+32%	-41%	+35%
	new_right	**+419%**	-88%	+29%
	ChapoTrapHouse	**+551%**	**+178%**	**+9%**
	whitebeauty	-42%	+96%	**+126%**
	Average	+412% (28*)	+131% (20*)	+331% (53*)

Table 2: (Partial) Results for impact of media attention and interventions on user growth and vocabulary adoption rate. Bold values denote a statistically significant difference in the forecasted and actual time series ($p < 0.05$). Subreddits are grouped by their manually assigned category. Values in brackets in the Average row denote the number of statistically significant changes.

3.2 Analysis and Results

Overview of analyses. To test our hypothesis ($H2$), we conduct three different analyses with each testing the impact of media pressure and interventions on subreddit growth and spread of problematic discourse.

Analysis 1: For toxic communities, what is the impact of negative media attention on subreddit growth? We now focus on the subset of our intervened subreddits which received negative media attention prior to their administrative intervention for toxic content — 43 in total. We conduct an interrupted time series analysis to test whether there was anomalous community growth (quantified by the rate of new creators joining the community) after the first time they received media attention. Across all 43 intervened subreddits with prior media attention, we see that 28 had statistically significant increases in user growth after the first time they received media attention. The average growth observed was 412%. Despite these alarming increases, the impact of

media attention appears disparate for different communities — e.g., *r/greatawakening* grew 3491% while *r/Mr_Trump* only grew 29% (both are statistically significant from our interrupted time series analysis). Grouping the subreddits by their topics, we see patterns emerge — groups that received the most negative media attention (subreddits in the manosphere, qanon, and extremist ideology categories) also had the largest growth rates from negative media attention. A subset of these results, grouped by 'subreddit topic' are reported in Table 2 in the *User growth (post-media)* column. Our findings provide evidence that negative media attention increases the growth rate for toxic communities. These results were further validated by our controlled analysis which showed that treatment subreddits (i.e., those receiving media attention) experienced statistically significant and higher user growth rates than control subreddits (i.e., those receiving no media attention but having similar pre-media user growth rates and subreddit topics).

Analysis 2: For toxic communities, what is the impact of negative media attention on the spread of problematic community vocabulary? We focus on the 43 intervened subreddits which received negative media attention prior to their interventions. We use an interrupted time series analysis to test whether the vocabulary of problematic subreddits is more commonly adopted across the platform after the first time they received media attention. The interrupted time series analysis returns statistically significant results if, given prior data, the growth of usage of the vocabulary on other subreddits is anomalous after the first media attention. We find that 20 of our 43 subreddits recorded statistically significant changes in the adoption of their vocabulary across the platform. The average increase across all 43 subreddits was 131%. Once again, we find that the effects are disparate across communities – e.g., *r/shortcels* experienced an increase of 1206% while *r/Mr_Trump* experienced a decrease of 82%. Specifically analyzing subreddits by their category, we find that only subreddits in the 'manosphere' experienced a consistent and significant increase in their vocabulary adoption rates after the first time they received media attention. A subset of these results are reported in the *Voc. growth (post-media)* column in Table 2. Our findings show that media attention results in increased adoption of an toxic community's vocabulary across the platform. These results were validated in our controlled analysis which showed that the vocabulary used in treatment subreddits were statistically different and spread more than those used in control subreddits (i.e., those receiving no media attention but having similar pre-media user and vocabulary growth rates).

Analysis 3: What is the impact of administrative interventions on the spread of problematic community vocabulary? We use an interrupted time series analysis to test whether the growth in usage vocabulary of a problematic subreddit across the platform varies depending on whether the subreddit received media attention or not. On average, across all 120 intervened subreddits we find that 53 subreddits had a statistically significant change in vocabulary adoption across the platform. Of these, 26 had received media attention prior to the intervention and 26 had not. The average increase observed in the subreddits that received media attention was 331% and 352% for those that did not. We note that the difference between the two groups was not found to be statistically significant. Breaking down our results by subreddit topic, we find that subreddits in the manosphere were once again found to have their vocabulary consistently adopted across Reddit even after the intervention. This breakdown is illustrated in the *Voc. growth (post-int.)* column in Table 2. Our findings show that subreddits which receive interventions see their vocabulary being adopted across the platform after the intervention, regardless of whether they received prior media attention or not.

Takeaways. We validated one of our hypotheses ($H2(1)$) that media attention on problematic communities increases the user growth rate in the community itself and increases the adoption of the community's vocabulary across the platform. Our findings were unable to validate our second hypothesis ($H2(2)$) that prior media attention on problematic communities reduced the effectiveness of administrative interventions. All together, our study allows us to conclude that media attention on a problematic community does lead to an increase in problematic activity within and outside the community itself. However, reactionary administrative interventions do not appear to have a significantly different impact on the communities which receive media attention.

4 Related Work

Our research was influenced by and makes contributions to research that can broadly be classified into two categories: platform moderation strategies and their consequences; and the influence of externalities on platform moderation.

Platform moderation strategies and their consequences. The dilemma of how to moderate effectively without resorting to extreme restrictions on discourse is not new to platforms as they increasingly find themselves grappling with challenges arising from being too strict or too lenient. Angwin (Angwin 2009) highlighted how restrictions and moderation on Friendster led to mass user migrations to more lenient platforms such as MySpace. Conversely, overly lenient moderation also presents problems for platforms. For example, the failure to address trolls and misogynistic content led to a loss of users along with the withdrawal of several offers to purchase and invest in Twitter (Ingram 2016). Increasingly, however, we find platforms offering moderation strategies as a commodity: some advertise increased safety and protection for its users (e.g., Reddit and Twitter) while others advertise no restrictions on discourse (e.g., Gab, Parler, and 4chan). Several studies, detailed below, have shown the former to suffer from inconsistency in moderation while the complete lack of moderation in the latter has been found to encourage extremism and toxicity (Zannettou et al. 2018; Hine et al. 2017).

The challenge of consistent and timely moderation. Numerous works have tracked discourse on platforms, specifically to measure the effectiveness of community-level interventions to suppress dangerous discourse. Early research conducted on Reddit (Chandrasekharan et al. 2017) showed the effectiveness of interventions applied to *r/fatpeoplehate*

and *r/coontown*. The study revealed a significant downturn in the amount of incivility amongst community members after the intervention was applied. However, this finding has been contradicted by several recent studies which have shown users and discourse from banned communities migrating to newer communities while maintaining or increasing their incivility (Habib et al. 2019; Ali et al. 2021; Horta Ribeiro et al. 2021). Habib et al. hypothesize that increasing inconsistency by platform administrators may be the reason for this contradiction. Our research which suggests a reactionary moderation strategy supports this hypothesis. Researchers have highlighted that inconsistencies associated with moderation may be attributed to the high cost and inherently poor scalability of human moderation and have proposed machine-learning based tools to assist moderators (Reddit moderators, specifically) identify communities at risk of violating platform rules (Habib et al. 2019; Chandrasekharan et al. 2019). Additionally, primarily relying on human moderators has been shown to have a severe effect of their mental health (Lagorio-Chafkin 2018; Roberts 2014; Wohn 2019). There are also opposing views to the adoption of machine-learning based tools for assistance in moderation due to their disposition to introduce obscurity and opacity in decisions (Gorwa, Binns, and Katzenbach 2020).

The consequences of inconsistent moderation. The effects of moderation inconsistencies have been found to be substantial. In the context of the 2016 US Presidential elections, several researchers (Benkler, Faris, and Roberts 2018; Allcott and Gentzkow 2017) found that discourse on social media platforms played a significant role in amplifying propaganda and fake news. These problems continue to arise today as online platforms provide a home for fringe elements promoting violent or problematic conspiracy theories. Failure to act effectively against such harmful ideologies by way of timely moderator interventions has been shown to result in the development of more extreme ideologies amongst community members. For example, researchers (Mamié, Ribeiro, and West 2021) showed that anti-feminist communities acted as a pathway to more radical alt-right communities. Further, the recent attack on the US Capitol and protests in Charlottesville that resulted in multiple deaths are both known to have been planned in online communities including large platforms such as Twitter and Parler (Prabhu et al. 2021). The importance of timely interventions on toxic content has been further highlighted by Scrivens et al. (Scrivens, Wojciechowski, and Frank 2020) who showed that there existed a gradual increase in the approval of toxic content in response to consistent toxic posting by community members. These results are in line with other studies showing how communities can become more extreme over time (Simi and Futrell 2015; Wojcieszak 2010; Caiani and Kröll 2015; Wright, Trott, and Jones 2020; Ribeiro et al. 2020).

External forces influencing platform moderation. Platform moderation does not operate without influence from external (particularly, economic and regulatory) forces. Numerous research efforts have analyzed the impact of the online advertising ecosystem on platform moderation. Bozarth et al. (Bozarth and Budak 2021) show how many fake news websites are mostly funded by top-tier advertising firms and an effective strategy towards combating fake news would be to have these advertisers blacklist these sites. Braun et al. (Braun, Coakley, and West 2019) showed how the 'Sleeping Giants', an activist group, strategically reported events of misinformation and racism to brands and advertisers (rather than the platforms themselves) in an effort to pressure them to withdraw their advertisements. This direct impact on the revenue streams of online platforms was found to cause changes in the moderation of misinformation and racist content. Along a similar vein, in 2019, YouTube experienced a series of boycotts from advertising agencies and brands in retaliation to the proliferation of toxic content. This event, now known as the 'Adpocalopyse' resulted in a large number of changes in YouTube's content policies, comment moderation, as well as video monetization policies (Kumar 2019; Dunna et al. 2022; Caplan and Gillespie 2020). These studies reflect the impact that pressure from advertisers can have on the moderation policies of online platform. Our work suggests that Reddit may not be an exception.

5 Discussion and Conclusions

Limitations and challenges. This work is fundamentally a best-effort study to understand one aspect of the relationship between platform administrators and the media, using observational data. Consequently, each of our contributions has their own limitations. First, our study considers both bans and quarantines as equal interventions. This may not be the case since they each might interact differently with media pressure. Unfortunately, owing to the small number of quarantined subreddits, this hypothesis is not possible to test with any statistical rigor and any conclusions might result in questionable validity. We expect that as Reddit begins quarantining more subreddits, future work will be able to explore the differences in the causes of bans and quarantines with statistical significance. Second, given the use of observational data and our inability to experimentally manipulate media pressure, we are unable to make strong causal claims regarding the relationship between media pressure and administrative interventions. This resulted in our need to frame a weaker hypothesis. We note, however, despite much debate regarding the use of mediation analyses for making causal inferences, the approach has been leveraged for precisely this purpose in many prior studies and one could argue that our models satisfy all the criteria required to make a causal inference (Pearl 2014; Pieters 2017). Next, our study required us to develop proxies for several parameters such as subreddit profitability, topics, and external pressure. It is unclear if our analysis found no impact from these variables due to the inaccuracy of our proxies or the actual absence of effects from them. We note that in the absence of groundtruth, however, one can only make best-effort approximations. Finally, our study is also limited by our decision to treat the media attention and interventions applied on each subreddit as independent events. This might have implications in scenarios where one subreddit receives negative media attention and this results in the closure of multiple related communities (e.g., Reddit banned five communities associ-

ated with encouraging self-harm on the same day). However, the alternate decision (grouping all subreddits receiving an intervention together as a single class) is also fraught with challenges that arise from the assumption that all simultaneous interventions occur due to the same effect.

Takeaways and implications. At a high-level, our study provides evidence of: (1) a reactionary (media- and internal-) pressure-driven administrative strategy being leveraged by Reddit, (2) the harms of giving media attention to toxic communities, and (3) the statistically similar (in)effectiveness of media- and non-media driven administrative interventions. Each of these findings has profound implications for platform administrators and media outlets.

Implications for platforms. As online social platforms increasingly find their communities becoming the originators and propagators of toxic and harmful content, calls to regulate them have started emerging. Particularly relevant is §230 of the US Communications and Decency Act which grants complete immunity to online platforms for publishing or censoring speech on their platforms — i.e., §230 guarantees no judicial consequences for moderation and administration decisions. Changes to this regulation have been proposed by both sides of the American political spectrum and, if enacted, are expected to have severe implications for moderation strategies employed by platforms such as Reddit. For example, any change which results in liability for publishing a users' toxic content will likely render reactionary administrative strategies, such as the one uncovered in our work, untenable. Further, although our findings suggest no significant difference in the effectiveness of interventions driven by media attention and otherwise, they do provide evidence that reactionary interventions do facilitate an increase in problematic behavior across the platform. Both these findings suggest the benefits of investing in and adopting proactive intervention strategies.

Implications for media outlets. Our study simultaneously highlights the importance of and the dilemma faced by the media in platform moderation. On the one hand, in the presence of reactionary platform administration and the absence of regulatory demands, it is imperative that the media hold platforms accountable for their administrative decisions. On the other hand, our findings also show that shining the media spotlight on problematic communities results in the growth and spread of the problematic activity. Thus, it remains unclear how media outlets should proceed — must they continue to hold platforms accountable or should they avoid publicizing problematic communities? Journalists have faced similar dilemmas in the past while negotiating reporting on hate crimes, suicides, and school shootings where they are faced with the consequences of possibly inspiring "copycat" behavior. In each such case, institutions of journalism such as the Society for Professional Journalists, the Poynter Institute, Thomson Reuters, and others have sought input from a variety of stakeholders in order to develop guidelines or "best practices" for these reports. Our research suggests the need for and value of such guidelines for reporting toxic online content and communities.

Acknowledgements

The authors are grateful to the anonymous reviewers for their feedback. This research was supported by the Air Force Office of Scientific Research under award #9550-20-1-0346. The opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the funding bodies.

References

Ali, S.; Saeed, M. H.; Aldreabi, E.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. Understanding the Effect of Deplatforming on Social Networks. In *13th ACM Web Science Conference 2021*.

Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *J. economic perspectives*.

Angwin, J. 2009. *Stealing MySpace: The battle to control the most popular website in America*. Random House.

Baron, R. M.; and Kenny, D. A. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Benkler, Y.; Faris, R.; and Roberts, H. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bozarth, L.; and Budak, C. 2021. Market forces: Quantifying the role of top credible ad servers in the fake news ecosystem. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 83–94.

Braun, J. A.; Coakley, J. D.; and West, E. 2019. Activism, advertising, and far-right media: The case of sleeping giants. *Media and Communication*, 7(4).

Caiani, M.; and Kröll, P. 2015. The transnationalization of the extreme right and the use of the Internet. *International Journal of Comparative and Applied Criminal Justice*.

Caplan, R.; and Gillespie, T. 2020. Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media+ Society*.

Centivany, A. 2016. Values, ethics and participatory policymaking in online communities. *Proceedings of the Association for Information Science and Technology*, 53(1): 1–10.

Chandrasekharan, E.; Gandhi, C.; Mustelier, M. W.; and Gilbert, E. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction*, (CSCW).

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. (CSCW).

Chen, A. 2012. Unmasking Reddit's Violentacrez, The Biggest Troll on the Web. https://tinyurl.com/mryavhdp.

Dunna, A.; Keith, K.; Zuckerman, E.; Vallina-Rodriguez, N.; O'Connor, B.; and Nithyanand, R. 2022. Paying Attention to the Algorithm Behind the Curtain. *ACM Conference on Computer Supported Collaborative Work (CSCW 2022)*.

Eisenstein, J.; Ahmed, A.; and Xing, E. P. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer.

Gorwa, R.; Binns, R.; and Katzenbach, C. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1): 2053951719897945.

Habib, H.; Musa, M. B.; Zaffar, F.; and Nithyanand, R. 2019. To Act or React: Investigating Proactive Strategies For Online Community Moderation. *arXiv preprint arXiv:1906.11932*.

Habib, H.; Musa, M. B.; Zaffar, F.; and Nithyanand, R. 2022. Are Proactive Interventions for Reddit Communities Feasible? *Proceedings of the International AAAI Conference on Web and Social Media*.

Hine, G.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Horta Ribeiro, M.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; and West, R. 2021. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–24.

Ingram, M. 2016. Disney, Salesforce Dropped Twitter Bids Because of Trolls. https://tinyurl.com/e5te5cjr. Accessed: 2021-10-09.

Kumar, S. 2019. The algorithmic dance: YouTube's Adpocalypse and the gatekeeping of cultural content on digital platforms. *Internet Policy Review*.

Lagorio-Chafkin, C. 2018. *We Are the Nerds: The Birth and Tumultuous Life of Reddit*. Hachette.

Mamié, R.; Ribeiro, M. H.; and West, R. 2021. Are Anti-Feminist Communities Gateways to the Far Right? Evidence from Reddit and YouTube. *arXiv:2102.12837*.

Marwick, A.; and Lewis, R. 2017. Media manipulation and disinformation online. *New York: Data & Society Research Institute*, 7–19.

Mittos, A.; Zannettou, S.; Blackburn, J.; and De Cristofaro, E. 2020. "And We Will Fight for Our Race!" A Measurement Study of Genetic Testing Conversations on Reddit and 4chan. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Müller, M. 2007. Dynamic time warping. *Information retrieval for music and motion*, 69–84.

Obadimu, A.; Khaund, T.; Mead, E.; Marcoux, T.; and Agarwal, N. 2021. Developing a Socio-Computational Approach to Examine Toxicity Propagation and Regulation in COVID-19 Discourse on YouTube. *Information Processing & Management*, 102660.

Pavlopoulos, J.; Thain, N.; Dixon, L.; and Androutsopoulos, I. 2019. Convai at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert. In *Proceedings of the 13th international Workshop on Semantic Evaluation*, 571–576.

Pearl, J. 2014. Interpretation and identification of causal mediation. *Psychological methods*.

Phillips, W. 2018. The oxygen of amplification. *Data & Society*, 22: 1–128.

Pieters, R. 2017. Meaningful Mediation Analysis: Plausible Causal Inference and Informative Communication. *Journal of Consumer Research*.

Prabhu, A.; Guhathakurta, D.; Subramanian, M.; Reddy, M.; Sehgal, S.; Karandikar, T.; Gulati, A.; Arora, U.; Shah, R. R.; Kumaraguru, P.; et al. 2021. Capitol (Pat) riots: A comparative study of Twitter and Parler. *arXiv preprint arXiv:2101.06914*.

Ribeiro, M. H.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; Long, S.; Greenberg, S.; and Zannettou, S. 2020. The Evolution of the Manosphere Across the Web. *arXiv preprint arXiv:2001.07600*.

Roberts, S. T. 2014. *Behind the screen: The hidden digital labor of commercial content moderation*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Scrivens, R.; Wojciechowski, T. W.; and Frank, R. 2020. Examining the developmental pathways of online posting behavior in violent right-wing extremist forums. *Terrorism and Political Violence*.

Simi, P.; and Futrell, R. 2015. *American Swastika: Inside the white power movement's hidden spaces of hate*.

Tufekci, Z. 2012. If Reddit Really Regrets "Not Taking Stronger Action Sooner", What Will It Do in the Future? https://tinyurl.com/5dymrax5.

Wohn, D. Y. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*.

Wojcieszak, M. 2010. 'Don't talk to me': Effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *New Media & Society*.

Wright, S.; Trott, V.; and Jones, C. 2020. 'The pussy ain't worth it, bro': assessing the discourse and structure of MGTOW. *Information, Communication & Society*.

Zannettou, S.; Bradlyn, B.; De Cristofaro, E.; Kwak, H.; Sirivianos, M.; Stringini, G.; and Blackburn, J. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*.

Zannettou, S.; ElSherief, M.; Belding, E.; Nilizadeh, S.; and Stringhini, G. 2020. Measuring and characterizing hate speech on news websites. In *12th ACM Conference on Web Science*.