Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit

ESHWAR CHANDRASEKHARAN*, University of Illinois at Urbana-Champaign, USA
SHAGUN JHAVER*, Rutgers University, USA
AMY BRUCKMAN, Georgia Institute of Technology, USA
ERIC GILBERT, University of Michigan, USA

Should social media platforms override an individual community's self-policing when it repeatedly break rules? What actions can they consider? In light of this hotly debated issue, platforms have begun experimenting with softer alternatives to outright bans. We examine one such intervention called quarantining, that impedes direct access to and promotion of controversial communities. Specifically, we present two case studies of what happened when Reddit quarantined the influential communities r/TheRedPill (TRP) and r/The_Donald (TD). Working with over 85M Reddit posts, we apply causal inference methods to examine the quarantine's effects on TRP and TD. We find that the quarantine made it more difficult to recruit new members: new user influx to TRP and TD decreased by 79.5% and 58%, respectively. Despite quarantining, existing users' misogyny and racism levels remained unaffected. We conclude by reflecting on the effectiveness of this design friction in limiting the influence of toxic communities and discuss broader implications for content moderation.

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

Additional Key Words and Phrases: content moderation; causal inference; design friction; graduated sanctions

> **CONTENT WARNING**: To describe the content on TRP and TD, we include examples of misogynistic and racist content that may offend our readers. We include them because these samples help explain Reddit's response and ground our research.

---

*Both authors contributed equally to this research.

Authors' addresses: Eshwar Chandrasekharan, eshwar@illinois.edu, University of Illinois at Urbana-Champaign, 4226 Siebel Center for Comp Sci, Urbana, Illinois, USA, 61801-2302; Shagun Jhaver, shagun.jhaver@rutgers.edu, Rutgers University, 4 Huntington St, New Brunswick, New Jersey, USA, 08901; Amy Bruckman, asb@cc.gatech.edu, Georgia Institute of Technology, 85 5th St NW, Atlanta, Georgia, USA, 30308; Eric Gilbert, eegg@umich.edu, University of Michigan, 105 S State St, Ann Arbor, Michigan, USA, 48109.

111

1 INTRODUCTION

Over the past few years, many scholars and activists have persistently called for social media platforms to remove content that promotes coded racism, misogyny, and conspiracy theories [90]. At the same time, there is a growing recognition that platforms need to better facilitate critical discussions on sensitive social issues like race, masculinity and immigration. In light of the hotly contested arguments about when and how they should intervene [8, 30], platforms have begun to consider the use of softer alternatives to outright bans. In this paper, we examine the influence of one such approach, called *quarantining*, deployed by the popular social media platform Reddit.

Reddit hosts a variety of online communities—thousands of subreddits address topics ranging from books and TV shows to science and politics. Many of these communities are well-moderated, host valuable content, and facilitate constructive discussions, e.g., r/ChangeMyView [91] and r/science [50]. Others lack redeeming social value, like the long-since banned r/FatPeopleHate (fat shaming) [12] and r/CreepShots (sexualized images of women without their knowledge). However, other subreddits do not neatly fit into either of these categories. Such subreddits contain content that Reddit considers unsuitable for the wider public but still suitable for members to access as they see fit. To handle such cases, Reddit deploys the community-wide intervention of *quarantining*. When a subreddit is quarantined, visitors trying to access it will see a warning screen, and they must make a deliberate choice to view its content. Quarantined subreddits also no longer appear in non-subscription-based feeds, recommendations, or search results (see Figure 1).

Reddit was explicitly founded on the principles of free speech [88]. Therefore, quarantining as a moderation strategy may help the platform fulfill its mission, which is to "allow just about anything so long as it does not prevent others from enjoying Reddit for what it is: the best place online to have truly authentic conversations" [88]. Quarantining also serves as a warning to the community to monitor itself if it wants to avoid being banned. Softer alternatives to punitive total bans like quarantining are an increasingly deployed option in the toolkit of platform interventions. Considering this, we present in this paper a case-study that examines the effects of quarantining two influential Reddit communities—r/TheRedPill and r/The_Donald.

1.1 The r/TheRedPill and r/The_Donald Quarantine

Reddit is a content aggregation platform organized into over 1.6 million[1] user-created communities known as subreddits. r/TheRedPill (TRP) is one such subreddit, created on October 25, 2012. TRP is part of Men's Rights Activists (MRAs), a set of interconnected groups that promote anti-feminist discourse online [68]. With over 250,000 subscribers[2] at the time of its quarantine, TRP is currently one of the most influential MRA communities. Another influential subreddit that was quarantined was r/The_Donald (TD), a community where users frequently posted offensive content and shared racist and far-right talking points [26, 54, 57]. TD was created on June 27, 2015 and had over 750,000 subscribers[3] at the time of its quarantine.

The following is a typical comment on TRP:

> "I believe you are mistaking the message. It's not about the angry 2nd stage TRPers screeching about all women being whores. It's about ignoring these girls' feigned innocence because it's a fucking lie. Better to believe they are whores and not treat them like a fuckin princess. That way they get to earn their place by your side (alongside the other six in your harem)."

---

[1]http://redditmetrics.com/history
[2]https://redditmetrics.com/r/TheRedPill
[3]https://redditmetrics.com/r/The_Donald

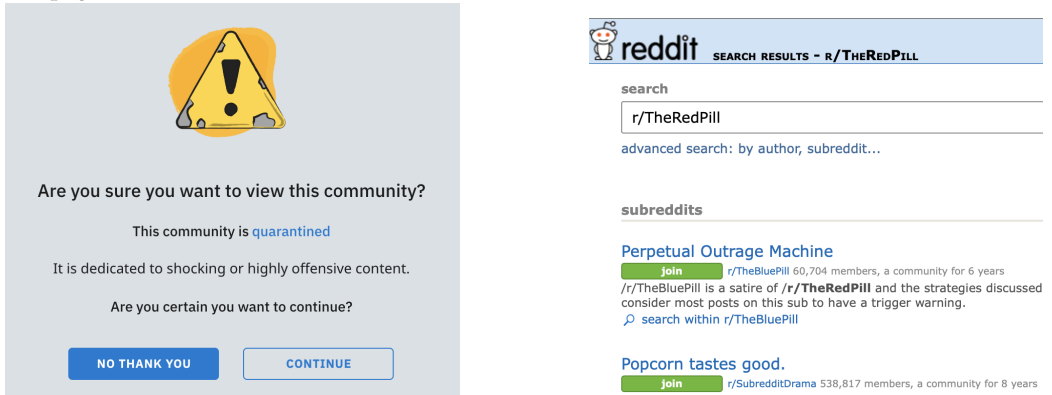(a) Splash page on the interface warns visitors to TRP.   (b) TRP is not shown in search results on Reddit.

Fig. 1. *Design friction* (i.e., intentionally adding points of difficulty to users' interactions with technology) introduced by the quarantine in the form of changes to Reddit's interface, making it harder to find and access TRP content. All users see a new interstitial in Reddit's interface (a), warning them that TRP is quarantined due to its *highly offensive content*. TRP also stopped appearing in Reddit search results during the quarantine (b), and users need to deliberately use TRP's direct URL to access the community. In addition to TRP posts not being indexed by search or the Reddit API, no TRP posts are shown in r/all, a prominent page aggregating the most popular posts from all parts of Reddit. In other words, the quarantine isolated TRP from the rest of Reddit by not surfacing TRP posts to "unaffected" users by chance (though users can still access TRP if they so choose).

TD hosts a wide variety of offensive content but is especially known for its racist and Islamophobic posts. A sample TD comment follows:

> *"No beards here. Just happy to see liberals triggered. I smile when they cry about gun control and how much they hate Trump. Pass laws just to get them foaming at the mouth. Enjoy niggerfaggot"*

On September 28, 2018, Reddit announced the quarantine function and on the same day, it *quarantined* TRP; similarly, Reddit decided to quarantine TD on June 26, 2019.[4] That is, Reddit introduced *design friction* into its interface when users tried to access either subreddit. The design changes to TRP are shown in Figure 1. When Reddit quarantines a subreddit, visitors to it are shown a splash page that requires them to explicitly opt-in to viewing its content. Additionally, the quarantined subreddit and its posts stop appearing in Reddit's indexing and search results. TD was subsequently banned by Reddit on June 29, 2020. Our study was conducted prior to this shutdown, when the subreddit was still under quarantine and active for participation.

In addition to being by far two of the largest subreddits to be quarantined, these communities are also important to study because of the out-sized influence they exerted on Reddit and beyond. r/The_Donald was the largest community for discussing the candidacy and later the presidency of Donald Trump. Digital outreach, particularly on social media like Reddit, was a particularly important part of the Trump election strategy. On the other hand, online groups like r/TheRedPill play an important role in radicalizing disenfranchised men and producing ideological echo chambers that promote violent rhetoric. Recently, media sources have highlighted the connections between

---

[4]Note: The authors of this paper are not Reddit employees, and this research was not financially supported or endorsed by Reddit.

extreme MRAs and mass killings in the U.S. and Canada [39, 56, 76]. Therefore, it is vital to study how community-wide moderation interventions affect the activities of these communities.

1.2 Research Questions

In its official announcements, Reddit states that it quarantines subreddits that "average redditors may find...highly offensive or upsetting" or that promote hoaxes or falsifiable conspiracy theories, like Holocaust denial [78]. The goal of quarantining is to prevent users from accidentally viewing potentially offensive or upsetting content posted on quarantined subreddits [93]. Reddit also expects that quarantining a community may signal the problematic nature of its content to its members, encouraging them to self-monitor their behavior and so escape being banned (i.e., shut down completely). We designed our research questions to study how well quarantining achieves these goals.

Quarantining fundamentally differs from banning, which has proven effective in reducing hate-based behavior [12]. Banning shuts a space down permanently, forcing its members to leave; prior analyses of community bans have primarily focused on understanding the migration patterns to other online spaces [12]. On the other hand, the quarantined space, though isolated from the rest of the site, is still available for member participation. Quarantining, therefore, censurs a community without censoring it [20]. This critical difference sets our work apart from prior work (e.g., [9]). We chose to focus on the immediate effects of quarantine on the "treated" subreddits—TRP and TD—given that these forums were still accessible to users.

It is not clear whether the quarantine would "work." Quarantining, which is essentially labeling a subreddit as "taboo," could lead to a Streisand effect,[5] i.e., the subreddit could become even more popular due to the quarantine and attract more new users than before.

Reddit's quarantine of TRP and TD provides a rare, natural experiment to examine the long-term effects of isolating radicalized groups. To examine the effects caused by the quarantine on these two communities, we ask the following research questions:

- **RQ1**: How were the participation levels within TRP and TD affected by the quarantine?
- **RQ2**: To what extent was the influx of new users to TRP and TD affected by the quarantine?
- **RQ3**: How was the use of misogynistic and racist language within TRP and TD, respectively, affected by the quarantine?

RQ1 and RQ3 aim to identify the effects of the quarantine on community members participating in TRP and TD (i.e., "Did Reddit worsen or improve the amount and quality of discourse in the quarantined group?"). RQ2 investigates the effects of quarantining TRP and TD on their ability to attract new users (i.e., "Did the Reddit quarantine prevent TRP and TD from assimilating new users?").

1.3 Summary of Methods, Findings and Implications

Methods. We answer our research questions by examining observational data from Reddit through a *temporal analysis* of TRP and TD's subreddit timelines. These included all comments and submissions made in the quarantined subreddits six months before to six months after they were quarantined. Working with over 85M Reddit posts and comments from Reddit, we chose metrics that include posting volume and frequency of words used from hate speech lexicons. We then used *causal inference methods* to examine variations in levels of interaction and racism/misogyny, accounting

---

[5]The term "Streisand Effect" emerged from an incident where an attempt by the actress Barbra Streisand to restrict online views of her Malibu Mansion on a public website had the paradoxical effect of stimulating public interest and leading to many more views than if she had done nothing [43].
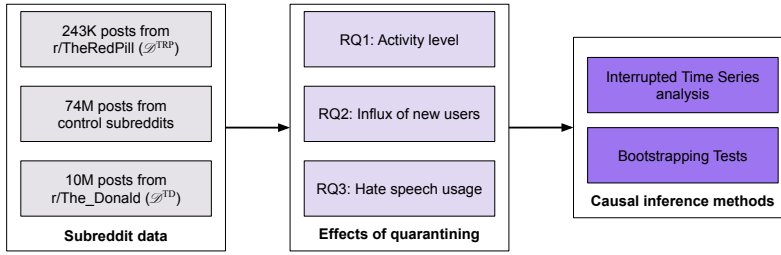
Fig. 2. Overview of the research pipeline employed in this study. We collected user timelines from the r/TheRedPill (TRP) and r/The_Donald (TD) subreddits and 100 control subreddits corresponding to each of them for our analyses. Our research questions explored the effects of Reddit's quarantine on posting activity level, influx of new users, and hate speech usage within TRP and TD. We used causal inference methods to answer these research questions.

for ongoing temporal trends within TRP and TD in addition to Reddit-wide trends. Figure 2 provides an overview of the research pipeline employed in this paper.

Findings. We determined that the *quarantine severely disrupted the influx of new users* to both subreddits—the rate of new users joining TRP and TD dropped drastically, by over 79.5% and 58%, respectively. We also found that already assimilated users within both subreddits exhibited changes in posting volume: activity levels of TRP users decreased substantially (by 52.4% and statistically significant w.r.t. corresponding control subreddits), but activity levels of TD users increased slightly (by 3.8%, but not statistically significant w.r.t. corresponding control subreddits). Despite these changes in posting activity, we found scant changes with respect to levels of misogynistic (TRP) or racist (TD) language.

Implications. For HCI, our work demonstrates the efficacy of introducing simple design friction to counteract antisocial behavior in online communities. For online moderation, we provide a computational framework that internet platforms can adopt to evaluate the effectiveness of moderation interventions. In her groundbreaking analysis of how to successfully manage common-pool resources, Ostrom argues for using graduated sanctions as a design principle [72]. Our work explores the idea of graduated sanctions in the context of community-wide moderation interventions. Given the philosophical debates around whether platforms should be allowed to ban certain types of speech [8], quarantining provides a *softer* alternative to banning. Our findings show one way that platforms can use low-cost design solutions like quarantining to compromise between containing antisocial activities and preserving freedom of speech.

2 BACKGROUND AND RELATED WORK

We next review related work to provide background information about the Manosphere and Red Pill communities, The_Donald community, online moderation, and platform interventions.

2.1 The Manosphere

The *manosphere* is an umbrella term for online communities that hold misogynistic ideas, perpetuate the belief that men are under attack, and propagate anti-feminist discourse [34, 40, 60]. Over the past few years, the manosphere has gained considerable media attention because of its linkages to the Isla Vista and Oregon mass shootings [14, 23, 27]. Campaigns of online harassment, death threats and rape threats against female gamers and journalists have also brought public attention to the manosphere [34, 62]. Prior work has linked these communities to the increasingly influential

Alt-Right, a political ideology that embraces misogynistic and white supremacist ideas [52, 57, 67]. The manosphere encompasses different communities that focus on a perceived crisis in masculinity, such as involuntary celibates (or incels), men going their own way (MGTOW), pick up artists (PUAs), traditional Christian conservatives (tradCons), and gamer/geek culture communities [83, 95].

2.1.1   The Red Pill communities. Another salient part of the manosphere are the *Red Pill* communities, which are built on a philosophy that rejects modern feminism and progressivism. The Red Pill concept is derived from the 1999 film The Matrix, in which the main character, Neo, is offered a choice to take *the blue pill*, which means living a life of delusion, or *the red pill*, which means becoming aware of the harsh realities of life. In the context of the men's rights movement, the Red Pill philosophy is aimed at awakening men to the purported misandry and brainwashing engendered by feminism [34]. This compelling cultural motif has brought together divergent factions of the *MRA*, enabling a rapid propagation of the Red Pill philosophy across multiple platforms.

The rhetoric of the Red Pill groups is dominated by evolutionary psychology, especially genetic determinism. Proponents use these theories to explain male and female behaviors of sexual selection, often superficially interpreting them to claim that women are inherently hypergamous and irrational and need to be dominated [34]. Red Pill communities prescribe a new form of masculinity that aims to attain the traditional hegemonic aims of sexual conquest and social dominance [67]. They have also 'geekified' concepts from evolutionary biology to cultivate a uniquely misogynist, racist, and heterosexist lexicon, including terms like "beta," "negging," "looksmaxx," "cuck," and "zero night stand" [34].

The technological affordances of social media platforms—such as speed, anonymity, disruption of geography, and social disembodiment—have allowed MRA and Red Pill groups to assert male hegemony in new ways [34]. For example, these platforms have facilitated the creation of echo chambers where members unite around emotionally charged claims of victimhood [67]. Massanari has observed that the algorithmic politics of platforms like Reddit prioritize the interests of young, white, cis-gendered, and heterosexual males and therefore provide fertile ground for anti-feminist activism [62]. The reproducibility and searchability of online content have further contributed to the viral spread of the Red Pill philosophy [49]. Since TRP is central to the ecosystem of MRA communities [67], disrupting its recruitment pipeline would likely have large downstream effects in the manosphere. In view of this, our work explores whether platforms like Reddit can play an active role in disrupting the spread of misogyny through design frictions like quarantining.

2.2   The Donald Subreddit

The_Donald subreddit (TD) was created on June 27, 2015, around the time that Donald Trump, then primarily known as a business and television personality, announced his presidential run. Users on this subreddit regularly discussed Trump-related news, with the vast majority supporting his candidacy and later his presidency. Ranked as one of the most popular online forums for supporting President Trump [4], TD also heavily influenced the news content on other social media sites like Twitter [100].

Soon after its creation, this subreddit began causing considerable controversy, both within the Reddit community and in the news media, for its offensive content [26]. Several Reddit users pointed out the "hateful" posts this subreddit hosted and argued that such posts undermined open political deliberation on the platform [26]. TD slowly evolved from being a political community focused on discussing the Trump campaign to becoming a hub for far-right talking points, coded racism, and trolling [54]. Many posters on TD were alt-right users who supported President Trump while frequently participating in racist, sexist, homophobic, and anti-Muslim subreddits [57]. TD

members often adopted a "troll slang" [61] and promoted conspiracy theories, like Pizzagate [70] and the Seth Rich murder conspiracy [33].

On June 26, 2019, Reddit quarantined TD, presumably in light of the problems detailed above. In this work, we examine the after-effects of this quarantine, specifically in terms of change in activity levels and influx of new users. Prior research has found that large proportions of users from the influential anti-black, racist subreddit called r/CoonTown migrated to TD after it was banned by Reddit in 2015 [12]. This migration more than doubled the posting activity of r/CoonTown users within r/The_Donald [12]. Given the influence of subreddits like r/CoonTown on anti-black, racist speech within TD, our study focuses on measuring this toxic impact.

2.3 Online Moderation and Platform Interventions

Grimmelmann defines *content moderation* as "the governance mechanisms that structure partici-pation in a community to facilitate cooperation and prevent abuse" [37]. Content moderation is a fundamental service that social media platforms provide to shape what content is allowed to remain online and what is removed, the accessibility of different content types, and the actions that accompany content removals [31]. It is vital that platforms maintain robust content moderation mechanisms to maintain their reputations and keep existing content usable [45].

As platforms grow in size, it becomes increasingly challenging for them to efficiently moderate high volumes of user-generated content efficiently [35, 45, 58]. They have therefore established a variety of complex, intricate systems to implement content moderation [11, 47, 48, 64, 92]. In recent years, researchers in HCI as well as other disciplines have started to examine the complexities and challenges involved in these systems [22, 25, 31, 37, 41, 55, 80, 85, 99]. Platforms regulate content at different levels of granularity—from moderating individual users or sanctioning an entire community to making site-wide changes—as we review below.

2.3.1 User-level moderation. The bulk of prior research on evaluating content moderation strategies has focused on user-level interventions [3, 53, 84]. Kiesler et al. theorized how design principles of clarifying community norms and conferring rewards and sanctions based on user behavior may encourage users to voluntarily comply with norms [53]. Seering et al. empirically tested the effectiveness of proactive and reactive moderation tools in preventing antisocial behavior among users [84]. Several studies have investigated the efficacy of using automated tools to help users make successful contributions on Wikipedia [3, 66, 96, 101]. More recently, Jhaver et al. evaluated the effectiveness of establishing community guidelines and offering explanations for content removal in improving user attitudes and behaviors [44, 47]. This extensive literature has helped establish strong theoretical foundations for how user-level interventions can be deployed to improve community outcomes.

2.3.2 Community-level interventions. While prior research has scrutinized the efficacy of many user-level interventions, we know comparatively little about community-level interventions. Multi-community platforms can choose to moderate entire communities. For example, Reddit can ban an entire community if it fails to comply with its guidelines[6]. Alternatively, Reddit administrators can choose to quarantine a community, which forms the focus of the current work. These community-level moderation interventions have widespread implications for a large number of users. Prior research most closely related to our work studied the effects of banning two controversial Reddit communities and found that banning helped reduce the use of hate speech on Reddit [12]. Some prior work has explored the phenomenon of quarantining on Reddit: Shen and Rose recently conducted a topic analysis of how Reddit users with different political ideologies react to Reddit's

---

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

quarantine policy [86]. Carlson and Cousineau examined the ethics of quarantining using existing ethical frameworks for content moderation and argued that quarantining allows Reddit to continue getting ad revenue from harmful subreddits while appearing to publicly denounce them [8]. We add to this literature by empirically examining a quarantine's effect on activity and hate speech levels of two prominent communities. We argue that examining the efficacy of quarantining is critical to develop a holistic understanding of how platforms can intervene to disrupt online radicalization even as they allow controversial groups to exist to uphold freedom of speech.

We also contribute to prior literature on graduated sanctions. In her book on community self-regulation, Ostrom writes that "graduated punishments ranging from insignificant fines all the way to banishment, applied in settings in which the sanctioners know a great deal about the personal circumstances of the other appropriators and the potential harm that could be created by excessive sanctions, may be far more effective than a major fine imposed on a first offender" [72]. Applying Ostrom's insights to online settings, Kiesler et al. recommend using graduated sanctions against offending users to increase the legitimacy and effectiveness of sanctions in online communities [53]. Our work expands the idea of graduated sanctions to the context of community-wide moderation interventions.

2.4 Design Friction

Design friction refers to intentionally adding "points of difficulty encountered during users' interaction with technology" [21]. The goal of design friction is to disrupt "mindless" automatic interactions and make users reflect before they take action [21, 65]. Even small barriers prior to an interaction, called *microboundaries*, can help users avoid engaging in behaviors that might not align with their personal values [9, 21].

Platforms generally try to facilitate low-friction interactions, e.g., by allowing users to rapidly share photos, access their friends' profiles, or up-vote posts. However, *some* online friction is highly desirable in certain situations, such as in avoiding the potentially instantaneous damage caused by encountering hate speech [94]. Recognizing this, many platforms have deployed moderation interventions that use design friction. For instance, Instagram attempts to moderate the growth of antisocial groups like pro-eating disorder (pro-ED) communities by prefixing search results for pro-ED hashtags such as #proana with a content warning [29]. Search engines display information about suicide helplines as the top results when people query suicide-related information [16]. Pinterest has started blocking search results for some anti-vaccination queries in an effort to combat the anti-vaxxer disinformation [98]. Reddit itself also uses splash page messages that require users to opt-in to view content on many non-quarantined communities, for example, pornographic (NSFW[7]) subreddits. These interventions represent design frictions that constrain access to antisocial content.

Quarantining on Reddit incorporates two manifestations of design friction. First, it removes the quarantined content from the main /r/all feed and search results, making it accessible only through direct URL link hits. Second, accessing quarantined communities requires an extra click-through acknowledgement, which makes users pause before viewing content in communities they would rather avoid. By examining quarantining, our research aims to engender conversations about the use of design friction to inhibit the growth of offensive speech and radical groups online.

3 DATA

We began by reviewing dozens of Reddit forums that discussed quarantining, and TRP and TD were by far the largest and most discussed quarantined subreddits. We therefore chose to focus on these two communities and our analyses only speak to the issues of quarantining these large

---

[7]Not Safe For Work

Table 1. Data description. $\mathcal{D}$ denotes the collection of submissions and comments made in r/TheRedPill (TRP) and r/The_Donald (TD) subreddits in the period from six months before to six months after the quarantine.

Key	Value
Number of posts in $\mathcal{D}^{\text{TRP}}$	243,758
Number of posts in $\mathcal{D}^{\text{TD}}$	10,832,057
Number of users in $\mathcal{D}^{\text{TRP}}$	24,077
Number of users in $\mathcal{D}^{\text{TD}}$	173,748
Number of control subreddits for TRP	100
Number of control subreddits for TD	100
Number of posts from control subreddits for TRP	19,174,200
Number of posts from control subreddits for TD	55,414,558

Table 2. List of top 10 control subreddits for TRP compiled based on the high proportion of co-posting by users from r/TheRedPill. We also provide the description for each subreddit as excerpted from the subreddit page. We compared the effects observed within the TRP subreddit to the effects observed within these subreddits in order to control for Reddit-wide trends.

Control Subreddits	Self-Description
asktrp	Here we have Red Pill Discussion for personalized questions about specific situations, people, scenarios.
askMRP	The Red Pill Rangers! Using the Red Pill In Their Marriages!
The30DayChallenge	Like all things Red Pill you want to take life by the balls and own it...Be in command of your fitness–that includes your training regimen.
marriedredpill	We are men that subscribe to The Red Pill (TRP) philosophy of sexual strategy, and are dedicated to applying it in marriage or in Long Term Relationships.
RPChristians	We are a subreddit that exists to provide a safe place for Christians to discuss problems, solutions and insights on relationships, sex, and biblical masculinity according to biblical principles.
PurplePillDebate	PurplePillDebate is a neutral community to discuss sex and gender issues, specifically those pertaining to /r/TheBluePill and /r/TheRedPill.
WhereAreAllTheGoodMen	We're just a bunch of clueless NiceGuys with kindness coins that don't seem to work in women's holes, so that the sex we're entitled to falls out.
RedPillWomen	This community was created as a harbor for RP minded women whose goal is to build a lasting and happy relationship with a great man.
DarkEnlightenment	The Dark Enlightenment, or neoreaction, focuses on the fundamentally flawed tenets of modern western culture.
BattleOfTheSexes	A place for no-holds-barred discussion and debate of all things sex and gender.

Reddit communities. Note that Reddit does not provide an official list of all communities that have been quarantined. Therefore, we relied on Reddit search for finding community discussions about quarantining to identify the quarantined communities. We are neither affiliated nor funded by Reddit, and our request to receive a list of all quarantined subreddits to Reddit did not receive any response. Though our focus on just two communities is a limitation[8], we hope that our findings

---

[8]Note that although we present results for just these two communities, our analysis method required us to identify 200 additional control subreddits as well as collect and analyze data from each of them. This is an involved process that included many manual and computationally intensive steps, which made it difficult for us to scale our study to analyze more quarantined subreddits. We chose to conduct a rigorous analysis on two case studies and describe how to scale it to other communities instead of conducting a shallow analysis on a larger number of case studies.

Table 3. List of top 10 control subreddits for TD compiled based on the high proportion of co-posting by users from r/The_Donald. We also provide the description for each subreddit as excerpted from the subreddit page. We compared the effects observed within the TD subreddit to the effects observed within these subreddits in order to control for Reddit-wide trends.

Control Subreddits	Self-Description
The_Europe	Make Europe Great Again - we update on European right wing leaders and upcoming elections to help the movement.
redacted	[REDACTED]. A wise man once said - [removed]
DrainTheSwamp	Make sure the swamp doesn't win!
AskThe_Donald	a Pro Trump, Pro Right-wing partisan subreddit
YouPostOnTheDonald	You post on the The_Donald you BIGOT REEEE. Links and images of The_Donald users getting BTFO with the "cuck comeback"
tucker_carlson	A subreddit dedicated to discussing Fox news host Tucker Carlson
TheNewRight	Home of the New Right. The New Right is a big-tent group within New Right Network, dedicated to mobilizing, countering the Left and energizing the Right!
kotakuinaction2	Promoting ethics in journalism, and opposing censorship and political correctness.
HillaryForPrison	This sub is for everyone who wants to put Crooked Hillary IN FEDERAL PRISON WHERE SHE BELONGS.
metacanada	a place to mock the hive mind of /r/Canada

and computational approach will inspire further research on alternative approaches to complete bans. We were motivated to take this approach by the success of our prior papers [12, 79] that examined the effectiveness of banning two subreddits and informed subsequent studies on the efficacy of moderation interventions [59, 71, 87] and deplatforming [46, 81].

We first collected a dataset of all submissions and comments that were posted on TRP and TD during 2018-2019. We obtained this data using the Google BigQuery service.[9] In the remainder of this paper, we refer to both submissions and comments as "posts." TRP was quarantined on September 28th 2018 and TD was quarantined on June 26th 2019. In all further analysis to observe the temporal effects of the quarantine on TRP and TD, we examined all the posts made on these subreddits from 180 days before to 180 days after the date of quarantine—April 1st 2018 to March 27th 2019 for TRP and December 28th 2018 to December 23th 2019 for TD. We denote these posts using $\mathcal{D}^{\text{TRP}}$ and $\mathcal{D}^{\text{TD}}$, respectively. Table 1 describes our data.

3.1 Control Subreddits

The causal inference question in our analysis is whether the quarantining of TRP and TD caused significant changes within them post-quarantine—especially effects on posting activity level, new user influx, and levels of misogyny and racism. In order to establish causality for any observed effects, we selected subreddits that were *highly likely to be quarantined* as control subreddits in our analyses. We use the co-posting behavior of users in our treatment subreddits as a proxy for the likelihood of being quarantined by Reddit. By comparing the effects observed within TRP and TD to the effects observed within their corresponding control subreddits over the same time period following the quarantine, we obtain lower-bounds for the effects caused within TRP and TD; in other words, this comparison accounts for any Reddit-wide effects observed following the quarantine, thereby establishing the causal link. Following the approach used in prior work [12], we selected our control subreddits as follows: we compiled a list of all subreddits where treatment

---

[9]Example table used for data collection: https://bigquery.cloud.google.com/table/fh-bigquery:reddit_comments.2018_03

users posted pre-quarantine, and then selected the top 100 subreddits based on the percentage of TRP (TD) users posting in these subreddits pre-quarantine (compared to their overall population). We used the PRAW Reddit API[10] to verify that the top 100 subreddits we selected were still active and had not been banned or quarantined at the time of our analysis. Tables 2 and 3 show the top 10 control subreddits selected through this process. Manual inspection[11] of these subreddits indicated that many of them contained misogynistic and racist content. Next, for each control subreddit, we retrieved all posts made within the period 180 days before and after the date of quarantine of their corresponding treatment (TRP or TD) subreddit. We also retrieved posts submitted within an additional 30 days period prior to the starting date to identify new users in each subreddit.

4 METHODS

We divided the timelines of all subreddits into time windows of 10 days each. We assigned a time window index w_i to each time window, with w_0 denoting the time window beginning at the date of the subreddit's quarantine (i.e., 09/28/2018 for TRP and 06/26/2019 for TD). In order to answer our research questions, we focused on three important indicators of user behavior within subreddits: (1) posting activity levels, (2) influx of new users, and (3) usage of toxic speech. Next, we detail how each indicator was operationalized.

4.1 RQ1: Posting Activity Levels

We measured posting activity levels within a subreddit using the volume of posts made in each time window w. For this, we iterated through each post in the subreddit timeline, and assigned it to the corresponding time window (based on the timestamp indicating when the post was made). Next, we aggregated all posts assigned to each time window. The total number of posts contained in each time window represented the subreddit's posting activity level for that time window.

4.2 RQ2: Influx of New Users

For each time window w, we defined *new users* as those users who had never posted on the subreddit before w. To distinguish new from preexisting users, we set aside a buffer period of three time windows (30 days) for which we did not compute new users. We collected all unique users who posted in the first three time windows in a set *seenUsers*. Starting from the fourth time window, we collected unique users in each time window, *currentUsers(w)*, and curated new users in that window, *newUsers(w)*, by compiling those users who were not in *seenUsers* (i.e., *newUsers(w)* is the set difference, *currentUsers(w)* − *seenUsers*). We iterated through each subsequent time window, adding new users seen in that window to the set *seenUsers* (i.e., *seenUsers* = *seenUsers* ∪ *newUsers(w)*). Posting on TRP or TD is a strong indicator of participation, compared to users just "lurking" [7] or viewing content; posts are also more reliable and easier to track.

4.3 RQ3: Misogyny and Racism Levels

4.3.1 Misogyny lexicon. Farrell et al. studied the manosphere by investigating the flow of extreme language across seven openly misogynistic subreddits [24]. In order to determine the prevalence of misogyny within these communities, they combined theoretical insights from feminist critiques of language to create nine lexicons capturing specific misogynistic rhetoric [24]. We used this lexicon consisting of 1,300 misogynistic terms to examine TRP misogyny levels.

---

[10]https://praw.readthedocs.io/en/latest/
[11]The first author used the validation strategy detailed in Sec. 4.3.3 to annotate whether a random sample of 100 posts from each subreddit contained misogynistic or racist content.
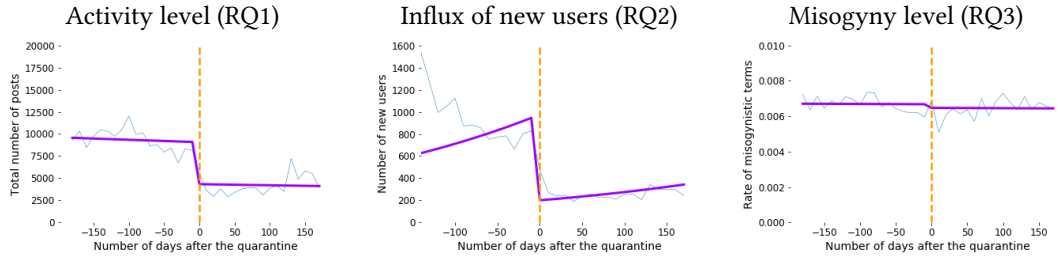
Fig. 3. Variations in posting activity levels, influx of new users, and misogyny levels within TRP. We compute rates for each metric based on all posts from TRP, using time-windows of 10 days, before and after the quarantine. As shown by the purple lines that "fit" the underlying time series, there were dramatic decreases in posting activity level and influx of new users to TRP (both changes were found to be highly significant through ITS analysis and bootstrapping tests), while misogyny levels remained unaffected (we found no significant changes). Note that the graphs appear to show a drop even before quarantining because we accumulate the metrics in each 10-day time window at its start, e.g., net posting activity within the first 10 days after quarantining is shown in the graph at time t=0.

4.3.2 Racism lexicon. In their study of Reddit bans [12], Chandrasekharan et al. created lexicons for detecting racism on the platform by using actual posts made within a racist subreddit banned for harassing behavior—r/CoonTown. The generation of this lexicon was guided by the European Court of Human Rights' definition of hate speech, including "comments which are necessarily directed against a person or particular groups of people," focusing on race, religion, and "aggressive nationalism and ethnocentrism," as well as homophobic speech [97]. Though hate speech and harassment lack specific and all-encompassing definitions, this usage-based approach provides a useful starting point to examine the prevalence of racist speech on different parts of Reddit. We used this lexicon, which consists of 100 racist terms, to examine racism levels within TD.

4.3.3 Instrument validation. To validate the instrument and ensure that the lexicons actually detected misogynistic and racist content, we manually annotated 100 randomly selected TRP posts and 100 randomly selected TD posts containing at least one term from the misogyny and racism lexicons respectively. This robustness check helped us answer the following: "Does the existence of lexicon terms in a post indicate that the overall post is misogynistic or racist in nature?" One of this paper's authors reviewed each post sampled from TRP and labeled whether the post was misogynistic or not using the definitions adopted by Farrell et al [24]. A similar labeling strategy was used to review posts sampled from TD and determine whether they were racist or not. Through this validation step, we found that 87 out of the 100 TRP posts containing at least one misogyny lexicon term were indeed misogynistic in nature. In other words, the misogyny lexicon provides a high precision of 87% when detecting misogynistic posts within TRP. Likewise, the racism lexicon was found to provide a precision of 82% when detecting racist posts within TD. During error analysis, we found that the "false positives" were mostly posts that contained only 1-2 lexicon terms and comments that either replied to or quoted information that other authors posted earlier. On average, we found that posts containing greater numbers of lexicon terms were found to be more misogynistic and racist.

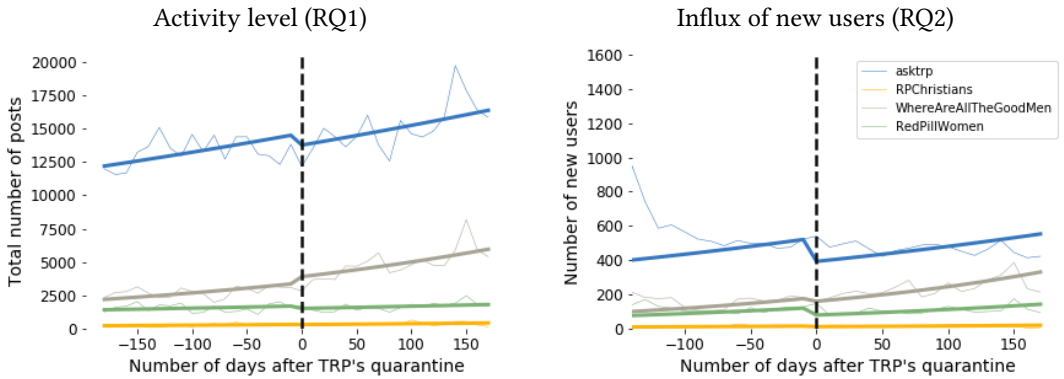## Activity level (RQ1)   Influx of new users (RQ2)

Fig. 4. Variations in posting activity levels and influx of new users in 4 subreddits selected at random from the top 10 control subreddits for TRP. We computed rates for each metric based on all posts from 4 random control subreddits, using time-windows of 10 days, before and after the date of TRP's quarantine. We verified that no control subreddit was banned or quarantined. Through bootstrapping tests, we determined that the effects observed in TRP were significantly *more extreme* than the effects in control subreddits (visually apparent through comparisons with Figure 3).

Activity level (RQ1)   Influx of new users (RQ2)   Racism level (RQ3)
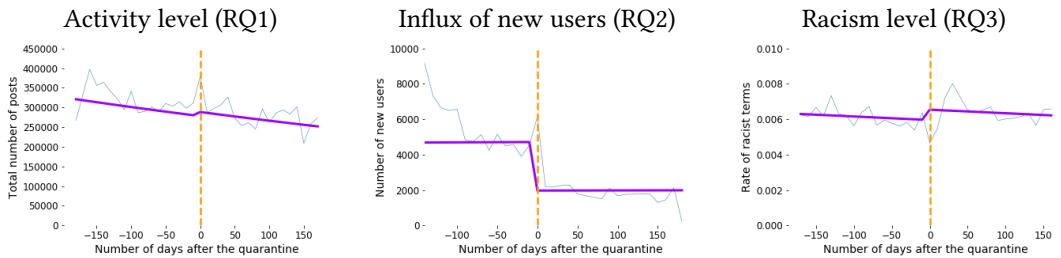
Fig. 5. Variations in posting activity levels, influx of new users, and racism levels within TD. We computed rates for each metric based on all posts from TD, using time-windows of 10 days, before and after the quarantine. As shown by the purple lines that "fit" the underlying time series, there was a dramatic decline in the influx of new users to TD (this change was found to be highly significant using ITS analysis and bootstrapping tests), but there were no drastic changes in posting activity or racism levels.

5 CAUSAL INFERENCE STRATEGY

The causal inference question we posed was whether the quarantining of a radical community like TRP or TD causes a decrease in radicalization within the subreddit (as measured by posting activity level, influx of new users, and misogyny/racism levels). Ideally, from a study design perspective, Reddit would have randomly chosen the subreddit to quarantine from a list of candidate subreddits. However, TRP and TD were not randomly chosen (to our knowledge), so we employed a number of techniques to approximate the results we would have observed if they had been randomly chosen. To determine whether the post-quarantine effects observed in TRP and TD could be explained by random chance, we calculated *p*-values for many key estimates in this paper. For many of these estimates (namely in RQ1, RQ2), we used classical *p*-values based on asymptotic results.
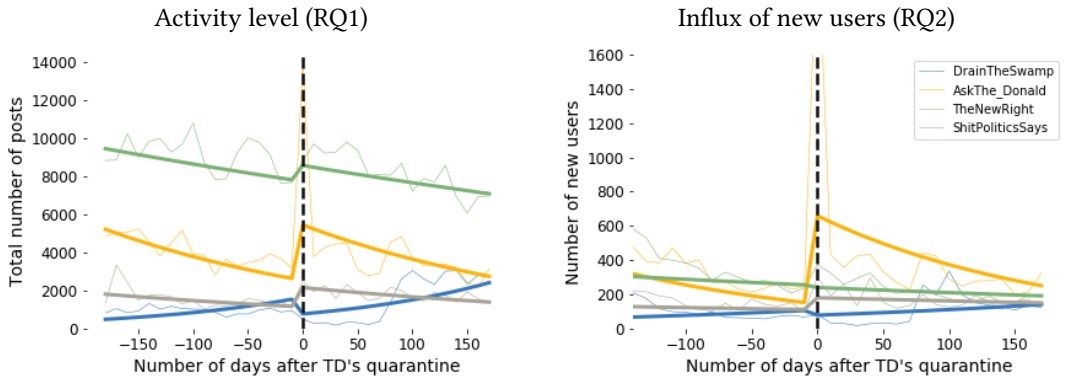
Fig. 6. Variations in posting activity levels and influx of new users in 4 subreddits selected at random from the top 10 control subreddits for TD. We computed rates for each metric based on all posts from 4 random control subreddits, using time-windows of 10 days, before and after the date of TD's quarantine. We verified that no control subreddit was banned or quarantined. Using bootstrapping tests, we determined that the reduction in the influx of new users observed in TD was significantly *more extreme* than the corresponding effects in control subreddits (visually apparent through comparisons with Figure 5).

First, we used a causal inference strategy called *Interrupted Time Series* [5] to measure the causal effects of the quarantine, accounting for ongoing temporal trends within the subreddits. We did this by comparing pre- and post-quarantine differences in posting activity level, influx of new users, and racism/misogyny levels within TRP and TD. Next, we employed bootstrapping tests to account for Reddit-wide trends. We did this by comparing the effects observed in TRP and TD to those observed in their corresponding control subreddits (examples shown in Table 2 and Table 3) that could potentially have been quarantined (see Figures 3 and 4 for effects within TRP, and Figures 5 and 6 for effects within TD).

5.1 Interrupted Time Series: Accounting for Ongoing Temporal Trends within TRP and TD

The temporal variations in posting activity level, influx of new users, and misogyny/racism levels in TRP and TD are shown in Figure 3 and Figure 5 respectively. Based on these figures, we observed substantial drops in posting activity and number of new users in TRP, and a dramatic drop in the number of new users in TD. Note that some drops in activity levels existed even before the quarantine, and these were not caused by the quarantine. There is a possibility that this evidence of a downward trend in activity levels and influx of new users was due to a general trend and not specifically due to the quarantine. To control for this, we performed an Interrupted Time Series (ITS) regression analysis. This let us causally attribute effects in TRP and TD to the quarantine.

In an ITS analysis, the dependent variable (e.g., posting activity level) for a subreddit is tracked over time, and a regression is used to determine if a *treatment* at a particular point in time caused a change in behavior (i.e., interrupted the time series). Taking the example of TRP, the *treatment* is the quarantine, initiated on 09/28/2018 (the date the subreddit was quarantined). As illustrated in Figure 7, an ITS regression models the behavior of the time series before the treatment is applied in order to predict how the series would have looked had the treatment not been applied. This is most often done by regressing the behavior data on time while also including in the regression an indicator variable for the post-treatment time periods (e.g., [12]). If we were to simply compare the
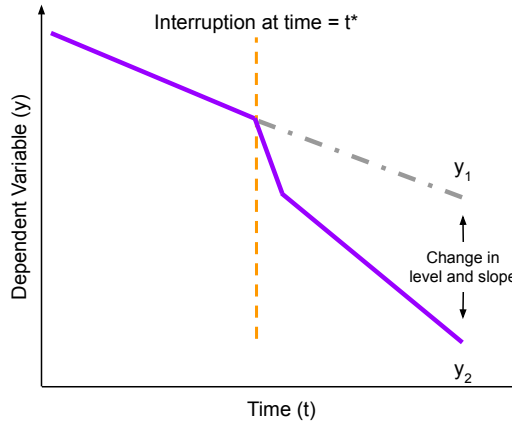
Fig. 7. Illustration of how the Interrupted Time Series (ITS) [5] strategy accurately estimates changes (in slope and level) caused by the interruption (denoted by the line at t^*), by accounting for any ongoing temporal trends within the time series (y). In other words, by computing the difference between the slopes and levels of line y_1 (extrapolated old series) and y_2 (actual new series), ITS analysis estimates the causal effect introduced by the event (interruption) at t^*.

averages of the behavior pre- and post-treatment without regressing the behavior also on time, then one could easily misinterpret a steadily increasing (or decreasing) time series as a treatment effect [5]. In this sense, an ITS regression shows that the behavior actually changed at the treatment time (e.g., 09/28/2018), instead of being simply a general trend, which could appear to be attributed to any randomly chosen time.

5.2 Bootstrapping Tests: Accounting for Reddit-wide Trends

We used significance tests to determine whether the observed effects in TRP could reasonably be ascribed to the randomness introduced in selecting the sample. If not, we have evidence that the effect observed in the sample reflects an effect that is present in the population [38, 51]. In order to account for Reddit-wide trends leading to decreases in activity levels across all subreddits, we developed a p-value using a bootstrapping test [38]. To conduct this test, we estimated the change in the dependent variables (i.e., posting activity levels, influx of new users) with respect to the date of quarantine for all control subreddits.

First, we ran the ITS analysis for all 100 control subreddits and obtained a new value of the coefficient for the ITS, β, for each dependent variable. This step yielded a distribution of 100 coefficients for the ITS coefficient, β. Figures 4 and 6 illustrate the effects observed in some control subreddits. Next, these distributions are compared to the actual β_{TRP} and β_{TD} from our analysis. If the actual β_{TRP} was "extreme" in relation to the distribution, then this provided evidence that the effect estimate we observed was unlikely to have been produced by chance. To formalize the notion of "extreme," we used p-value, which denotes the proportion of the β distribution that exceeds β_{TRP}. For our analysis, because we were interested in drops in posting activity levels and numbers of new users, p was the proportion of the β distribution that was smaller than the actual β_{TRP}. We bootstrapped by running multiple simulations to compute the distribution of βs for control subreddits. In each simulation, we randomly sampled 50 coefficients, and computed *how extreme β_{TRP} was* when compared to the sampled distribution. Because the distribution was

Table 4. Interrupted Time Series regression results for posting activity level, influx of new users and the usage of misogynistic/racist language in TRP and TD. The β coefficient from ITS regression, as well as the one-tail p-values are also included. Results show a highly significant causal effect of the quarantine on the substantial decrease in posting activity levels and influx of new users in TRP. However, there was no significant change in the use of misogynistic language caused by the quarantine. In TD, the influx of new users caused by the quarantine declined appreciably, but there were no drastic changes in posting activity levels or the use of racist language. Overall, our findings indicate that the quarantined disrupted the incoming streams of new users to both TD and TRP, but it did not change behaviors of preexisting members of these communities. Here, p<0.001: ***

Metric	β_{TRP}	% change TRP	β_{TD}	% change TD
Activity level	-0.7418***	-52.37%	0.0367***	3.74%
Influx of new users	-1.58***	-79.51%	-0.8732***	-58.24%
Misogyny level	-0.0002	-0.0	-0.0000	-0.0%

constructed by simulation, we set up the simulations to terminate when the p-value of sampled distributions reached saturation (i.e., no change in p-value following consecutive simulations). We repeated a similar procedure to compute *how extreme β_{TD} was* when compared to the effects in its corresponding control subreddits.

6   RESULTS

We now summarize results from our analyses.

6.1   Changes in Posting Activity Level (RQ1)

Results from the ITS analysis are shown in Table 4. We observed that quarantining significantly reduced—by over 52.37% (β_{TRP} = -0.7418, p-val < 0.0001)—posting activity in TRP (shown in Figure 3). By comparing change in posting activity across control subreddits using bootstrapping tests, we found that a one-sided significance test gave p-val < 0.0001, indicating strong evidence that the TRP activity decline was indeed caused by the quarantine, not by a general Reddit-wide decrease in posting activity. On the other hand, we observed that quarantining significantly increased—by over 3.74% (β_{TD} = 0.0367, p-val < 0.0001)—posting activity in TD (shown in Figure 5). When we compared change in activity level across control subreddits using bootstrapping tests, we found that a one-sided significance test gave p-val > 0.05. This indicates that the increased posting activity in TD was not extreme compared to control subreddits.

6.2   Sharp Decreases in the Influx of New Users (RQ2)

Through the ITS analysis (results in Table 4), we observed that the quarantine caused a significant decrease of over 79.51% (β_{TRP} = -1.58, p-val < 0.0001) in the influx of new users to TRP and a decrease of over 58.24% (β_{TD} = -0.8732, p-val < 0.0001) in the influx of new users to TD (shown in Figure 3 and Figure 5, respectively). Comparing the change in rate of new users posting within different subreddits using bootstrapping tests, we found that a one-sided significance test gave p-value < 0.0001 for both TRP and TD. This provides strong evidence that the decreases in influx of new users to TRP and TD were indeed caused by the quarantine, not by a general Reddit-wide decrease in new user influx. These results suggest that the quarantining of TRP and TD did not cause users from other factions of Reddit to post on TRP and TD, indicating an absence of the "Streisand effect." Thus, in the cases of r/TheRedPill and r/The_Donald, Reddit's intervention of quarantining drastically reduced the number of new, unaffected users posting on TRP and TD.

6.3 No Significant Effects on Hate Speech Usage (RQ3)

Through the ITS analysis, we found that the quarantine did not cause a significant change in the usage of misogynistic words in TRP (β_{TRP} = −0.0002, p-val > 0.5) or racist words in TD (β_{TD} = −0.0000, p-val > 0.5). When using all 1,300 words in the misogyny lexicon, we observed a slight downtick in the use of misogynistic terms immediately after the quarantine (shown in Figure 3). Upon manual inspection, we determined that most posts following the quarantine focused on the event itself (i.e., Reddit's decision to quarantine TRP), criticising the admins for their actions, and planning their next steps in case Reddit decided to ban TRP (i.e., shut TRP down permanently, similar to bans imposed on hateful subreddits in the past [12]). This caused the slight decline in the use of misogynistic words immediately after the quarantine. However, this decrease in misogynistic speech was not significant, and it was only temporary. Over time, TRP users resumed their misogynistic speech, leading to levels of misogyny that were comparable to the pre-quarantine time period. In other words, the level of misogynistic language in the TRP community was not affected by the quarantine. Likewise, the levels of racism prevalent on TD was not affected by the quarantine. Again, we observed a slight uptick in the use of racist words in the TD subreddit immediately after the quarantine, but this increase was not significant, and it was only temporary. Since we found no significant trends in our ITS analyses for hate speech usage in TRP and TD, we did not conduct the follow-up bootstrapping analyses.

6.4 Additional Checks

We verified that quarantining TRP and TD did not create conditions within control subreddits that skewed the baseline coefficient distribution in the opposite direction. For example, the *extremeness* (i.e., p-val) of the decrease in the rate of new users joining TRP, as computed through the bootstrapping tests, was not caused by increases in new user influx within all of the control subreddits. In order to check for this, we ran the ITS analysis for posting volume and new user influx within all control subreddits (100 for TRP and 100 for TD). We found that only 9 out of the TD's 100 control subreddits had a significant increase in posting volume (wrt the date of TD's quarantine), while 15 out of TRP's 100 control subreddits had a significant increase in posting volume. Additionally, only 2 out of the TD's 100 control subreddits had a significant increase in new user influx while none of TRP's 100 control subreddits had a significant increase in new user influx. Therefore, the *extremeness* of the observed decreases in posting volume within TRP and new user influx to TRP and TD were not inflated due to a simultaneous increase in these variables within all of the control subreddits.

7 DISCUSSION

We now summarize what our findings reveal about the effectiveness of quarantining as a moderation strategy to curb the flourishing of hateful groups online. We also highlight the theoretical and design implications of this research, discuss its limitations, and identify directions for future work.

7.1 The Effectiveness of TRP's and TD's Quarantine

7.1.1 Posting activity in TRP dropped drastically, but remained misogynistic. Posting activity in TD slightly increased, and remained racist. Through our ITS analysis, we found that the quarantine significantly reduced posting activity in TRP. Despite this reduction, the relative use of misogynistic terms in TRP posts showed no significant change. Additionally, quarantine slightly increased the posting activity in TD and the relative use of racism terms in its posts showed no significant change. This lack of change in the use of toxic terms suggests that users who are already entrenched within radical groups resist changing their behavior. As mentioned earlier, Reddit noted that one of the

primary goals of quarantining is to compel users to rethink their behavior and reduce offensive posts. For example, when quarantining TRP, Reddit administrators messaged its moderators that to exit their quarantine status, the community would need to demonstrate a sustained reduction of misogyny. However, our findings suggest that quarantining did not serve this goal. A lack of reduction in posting activity in TD post quarantine also indicates that quarantining a community does not consistently deter its activity levels. Thus, if the platform's goal is to curb activity in an offensive community, quarantining by itself is not a reliable solution.

7.1.2 Influx of new users to TRP and TD declined. Before we began our analyses, it was not clear how quarantining would affect the stream of new users into quarantined sites. Labeling a subreddit as "highly offensive" could have resulted in a Streisand effect [43], i.e., making it even more popular and therefore attracting more new users than before. However, we found that this was not the case: despite widespread discussions about the quarantine in multiple Reddit communities, the numbers of new users to TRP and TD significantly dropped due to the quarantine. This demonstrates how design friction—indicating to new users the offensiveness of a group and introducing barriers to the access of offensive spaces—can substantially disrupt the growth of such groups.

In recent years, internet platforms have often been accused of serving as a venue to radicalize disenfranchised groups [2, 69]. However, the effectiveness of quarantining as a strategy to impede the integration of new individuals into TRP and TD communities offers important empirical evidence of how platform interventions can successfully address these societal problems. Platforms like Reddit have thousands of active user-created communities, which makes monitoring the activity of every community challenging. Nevertheless, our findings suggest that platforms can provide societal value by devoting resources to detecting, labeling and isolating antisocial groups.

7.1.3 Cross-subreddit effects of quarantining. Through the ITS analysis within the control subreddits, we found there were no drastic increases in posting volume and new user influx within these subreddits that were caused by the quarantine. In fact, there was a decrease (on average) of 40.78% in posting volume and a decrease of 41.47% in rate of new users posting within TD's control subreddits. Similarly, there was a decrease (on average) of 20.16% in posting volume and a decrease of 31.69% in rate of new users posting within TD's control subreddits. Moreover, we found that 26 out of the TD's 100 control subreddits had a significant decrease in posting volume (wrt to the date of TD's quarantine), while 60 out TRP's 100 control subreddits had a significant decrease in posting volume. More importantly, 63 out of the TD's 100 control subreddits had a significant decrease in new user influx while 89 of TRP's 100 control subreddits had a significant decrease in new user influx. These results provide early evidence that quarantining likely depressed the activity levels and new user influx in many other communities frequented by existing users of TRP and TD. At the very least, quarantining TRP and TD did not *spread the infection* to other parts of Reddit.

7.1.4 Calls to migrate elsewhere. In additional analyses, we explored whether the reduction in activity on TRP and TD could be explained by community managers asking users to migrate on a different site. Moderators usually communicate such official announcements through posts that are stickied at the top of the subreddit. We manually reviewed all posts submitted on TRP and TD between the time these subreddits were quarantined and the end date for our analyses (i.e., six months after the quarantine). We could not find any official (stickied) posts asking users to migrate from TRP within the time period of our study. Indeed, as of Sept 1, 2020, the TRP subreddit still remained accessible. However, we found one post on TD, submitted by a moderator on Nov 19, 2019, that asked users to bookmark an external site https://thedonald.win, presumably to prepare users to migrate there in case of a ban. To evaluate whether our results are affected by this official announcement on TD, we ran our ITS analysis only up until the time window before this post was

submitted. Our results show that even for this restricted period, quarantining significantly reduced the new user influx (β=-.927, p<=.001). This suggests that even before the official calls for migration to a different site, quarantining had disrupted the activity on TD.

7.2 Design and Theoretical Implications

7.2.1 Effectiveness of design friction as a moderation strategy. For HCI, our work demonstrates the efficacy of introducing design friction [21, 65] to counteract antisocial behavior in online communities. Quarantining resulted in the following design changes: (1) TRP and TD were relegated to a "sort of bad-actor closet [20]" through removal from aggregated content for non-community users; and (2) all visitors to them were shown a splash page saying that the community was quarantined and highly offensive, and asked if they would still like to view the community. This meant that participants had to deliberately choose to continue, in order to participate in the quarantined subreddit. While our methods do not allow us to disentangle the effects of these two changes, overall our findings show the potential for design friction to serve as a new moderation strategy in the toolbox of social platforms. Given the philosophical debates around freedom of speech online [8], and simultaneously, a platform's responsibility to protect its users and brand from harm [94], quarantining provides a *softer*, perhaps more ethical, alternative to banning. Further, since community-wide moderation interventions like quarantining are necessarily human-made, they bypass the criticisms that platforms are increasingly facing about the pitfalls of automating their moderation [32, 36].

7.2.2 Framework to evaluate the effectiveness of interventions. For online moderation, we provide a computational framework that internet platforms can use to evaluate the effectiveness of moderation interventions. We would like to stress that our methods and causal framework are independent of the target community, and our code is publicly available.[12] This makes it easy to extend our approach to examine the effects of such interventions on other communities in the future.

The empirical work in this paper suggests that when narrowly applied to specific, large misogynistic and racist groups, quarantining can work to reduce the influx of new users and contain the behavior. Lately, governance scholars have used Reddit quarantining as a lever to discuss the ethics of content moderation by private companies [8, 20]. For example, Carlson and Cousineau have explored existing ethical frameworks for content moderation such as the Santa Clara Principles [82] and Rawls' [77] veil of ignorance to develop a decision-making model that provides platforms guideposts for addressing complex content regulation questions. We argue that the efficacy of such strategies, as demonstrated in this paper for quarantining, should also inform these ongoing conversations [8, 20] around their possible future use. Further, our framework can be easily adapted to evaluate other moderation interventions in the future.

7.2.3 Platform interventions as graduated sanctions. In this work, we studied quarantining which is a community-level intervention. However, social platforms predominantly tend to use user-level interventions such as post removals (automated and manual), temporary banning, and suspensions. This raises the question: when should platforms rely on user-level interventions, and when should they resort to community-level interventions? In the distributed moderation design of Reddit, the local volunteer moderators of a community are largely responsible for implementing user-level interventions and the small group of paid Reddit administrators carry out community-level interventions [13]. This arrangement naturally leads to a hierarchy of graduated sanctions in which communities are punished increasingly severely—from removals of individual user postings to community bans. In their influential work "Regulating behavior in online communities," Kiesler et al.

---

[12]We will make our code available after peer review.

theorized that "graduated sanctions increase the legitimacy and thus the effectiveness of sanctions" [53, 72]. In view of this, it is possible that community-level interventions will be more acceptable if they follow a stream of user-level interventions or local governance failures. Graduated sanctions can also help keep platforms accountable, ensuring that the reason for each escalation of sanctions is clear to users. Indeed, in line with our findings, Reddit noted that the TD community did not reform their behavior, and consequently banned TD on June 29, 2020.

It is also important to note that quarantining is just one of the many levers of change available to platforms. Though actions like quarantining and banning are often widely covered by the news media [74, 89] and reflect well on the platform's moderation, they need to work alongside efficient user-level interventions and coordination with local community moderators to ensure overall platform health.

7.3 Limitations and Future Work

7.3.1 Focus on effects within the quarantined communities. In this study, we examined whether the quarantined subreddit could become more popular due to the quarantine and attract more new users than it had previously. Since the communities still exist, we chose to focus our research questions on understanding the changes within them post-quarantine instead of its members' migration to other communities. Our results clearly demonstrate the effectiveness of quarantining in delimiting the influx of new users to such problematic forums. Even though these quarantining events were widely discussed on Reddit, we did not observe a Streisand effect of significantly increased user influx. However, it is possible that quarantining may just shift problematic behavior and lead to increased activity elsewhere. We provide evidence that quarantining did not lead to drastic increases in activity on similar other active subreddits (Sec. 7.1). Yet, a more rigorous analysis of members' migration to other websites is beyond the scope of this paper and is covered in a different paper [79]. In summary, our cross-platform analyses show that quarantining did not significantly increase the posting activity of these communities on other websites. Going forward, comparative analyses of offensive groups that continue to operate while quarantined on mainstream platforms versus groups that self-host on external websites would substantially inform the use of community-level sanctions.

7.3.2 Focus on two specific communities. We focused on two specific quarantined communities, TRP and TD subreddits. To really understand whether quarantining "works," we need to study more cases of quarantined subreddits. Nevertheless, this research examines important aspects of the quarantining of most prominent subreddits, and is therefore a key first step towards answering that larger question. Amidst the growing amounts of research on online moderation and new ways to counter antisocial behavior [11, 15, 48], we provide a powerful tool for causally evaluating moderation interventions. Our open-source code would also help community managers to evaluate the effects of platform sanctions on their own communities.

A wide range of subreddits get quarantined. Quarantined subreddits include communities that focus on racist tropes, pro-eating disorders, cannibalism, and holocaust denial. We observed that while a few subreddits have their quarantine status removed and return to normal operations after some time, the majority of subreddits are subsequently banned. It would be interesting to analyze whether the subreddits that come out of quarantine maintain improved discourse over long periods of time or whether they return to incendiary behavior.

7.3.3 Lexicons for measuring hate speech. We applied several robustness checks to ensure that our findings are causal in nature, i.e., they reflect the actual consequences of quarantining and do not simply mirror ongoing temporal or Reddit-wide trends. Although we used a comprehensive set of lexicons to analyze trends of misogyny and racism, our analyses are limited by the selection of

terms in these lexicons. It is likely that these communities have devised new misogynistic and racist terms that are not captured in the lexicons we use. How language evolves within these groups and how it shapes anti-feminist and racist discourse offers an interesting direction for future work.

7.3.4 New v/s inactive users. We reserved a buffer of 30 days before the start date of our analyses to identify new users. However, some users who we identify as *new users* may not be new but just recently inactive. We computed that the average gap between two posts for any user has a median of 8 days. Thus, we expect that our buffer period of 30 days would ensure that most *new users* are truly new. Even when that is not the case, the new engagement by previously inactive users speaks to the same issues as our analyses for new users, i.e., the extent to which quarantining affects the attraction of non-active members to the community.

7.3.5 Facilitating discussion on race and masculinity. We note that we did not make the decision to quarantine TRP—Reddit administrators did. While our findings suggest their decision was wise, it is important to recognize that there are many topics about men and masculinity (e.g., high male suicide rates, society's neglect of male mental health) that need discussion in our society today. Many scholars have emphasized the importance of recognizing multiple masculinities—black, white, working-class, middle-class—and the relations between them for an intersectional analysis of gender relations, reducing men's perpetration of violence, and increasing support for gender justice [17–19, 75]. A crucial future direction is identifying how to support constructive discussions about masculinity and protect them from incursions of hate speech. Similarly, the quarantine of TD helped reduce the influx of new users to a toxic community, and we consider this a desirable outcome. Still, we need to find ways to support important discussions about race and conservative principles that do not devolve into conspiracies, faux outrage over free speech, and racism.

7.3.6 Potential for quarantining to increase insularity. Prior research has shown that mediated communication often facilitates the formation of exclusionary communities that in more extreme cases may become echo chambers [73] where users intensify shared views, foster intolerance, and are sheltered from ideological opposition [1, 6, 28, 42]. By alienating sanctioned communities from the rest of the platform, quarantining may increase the insularity of those communities, reduce the wider public pressure to limit their regressive social attitudes, and encourage moves toward extremism. Isolated communities are also more likely to "favor exit over voice [10]," i.e., they would rather exit the platform than express constructive criticism to help improve platform operations [63]. Future research should therefore empirically evaluate the dynamics of how insularity within a community changes in response to quarantining.

7.3.7 Supporting community-wide moderation interventions. Finally, platforms like Reddit have difficult decisions to make when they choose to quarantine (or ban) a particular group. While a European approach makes hate speech illegal, Americans tend to hesitate because choosing where to draw the line between offensive and not offensive speech can lead to a slippery slope. Better data analysis and visualization tools could help administrators make more informed decisions about where and how to draw that line.

It is also important to recognize that platforms do not take community-level moderation decisions unilaterally. These decisions are often a response to a long, drawn-out set of complaints by end-users and moderators, media rebukes, and pressure from the public, lawmakers and advertisers [8]. Further, administrations' decisions to regulate controversial communities have previously resulted in widespread collective protests by moderators [63]—such incidents may influence how admins wield community-level actions in the future. Delineating the participatory aspect [10] of such policymaking processes is an exciting direction for future research.

8 CONCLUSION

In this paper, we examined the effectiveness of quarantining, a community-wide moderation intervention, in changing end-user behavior in two influential communities, the TRP and TD subreddits. This low-cost intervention dramatically diminished the influx of newcomers to each community but it did not modulate the offensive discourse. We conclude that quarantining can be a surprisingly effective compromise between censoring offensive speech and preserving freedom of expression in regulating content that exists on the margins of acceptability. Going forward, additional research and thoughtful design work can help to develop more moderation strategies that can achieve such compromises.

REFERENCES

[1] Kimberley Allison and Kay Bussey. 2020. Communal quirks and circlejerks: A taxonomy of processes contributing to insularity in online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 12–23.

[2] Cristina Archetti. 2013. Terrorism, Communication, and the Media. In *Understanding Terrorism in the Age of Global Media*. Springer, 32–59.

[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] BBC. 2019. Reddit places pro-Donald-Trump forum in quarantine. https://www.bbc.com/news/technology-48783866

[5] James Lopez Bernal, Steven Cummins, and Antonio Gasparrini. 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International journal of epidemiology* 46, 1 (2017), 348–355.

[6] LA Brainard. 2009. Cyber-communities. *International encyclopedia of civil society* (2009), 587–600.

[7] Cody Buntain and Jennifer Golbeck. 2014. Identifying social roles in Reddit using network structure. In *Proceedings of the 23rd international conference on world wide web*. ACM, 615–620.

[8] Caitlin Ring Carlson, Luc Cousineau, and Caitlin Ring Carlson. 2020. Are You Sure You Want to View This Community ? Exploring the Ethics of Reddit's Quarantine Practice. *Journal of Media Ethics* 00, 00 (2020), 1–12. https://doi.org/10.1080/23736992.2020.1819285

[9] Marta E. Cecchinato, Anna L. Cox, and Jon Bird. 2015. Working 9-5? Professional Differences in Email and Boundary Management Practices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3989–3998. https://doi.org/10.1145/2702123.2702537

[10] Alissa Centivany and Bobby Glushko. 2016. " Popcorn Tastes Good" Participatory Policymaking and Reddit's. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1126–1137.

[11] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

[12] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 31.

[13] 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.

[14] Soraya Chemaly. 2015. Mass killings in the US: Masculinity, masculinity, masculinity. *The Huffington Post* (2015). https://www.huffpost.com/entry/mass-killings-in-the-us-w_b_8234322

[15] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Ninth International AAAI Conference on Web and Social Media*.

[16] Qijin Cheng and Elad Yom-Tov. 2019. Do search engine helpline notices aid in preventing suicide? Analysis of archival data. *Journal of medical internet research* 21, 3 (2019), e12235.

[17] Patricia Hill Collins. 2004. *Black sexual politics: African Americans, gender, and the new racism*. Routledge.

[18] Robert William Connell. 2005. *Masculinities*. Polity.

[19] Will H Courtenay. 2000. Constructions of masculinity and their influence on men's well-being: a theory of gender and health. *Social science & medicine* 50, 10 (2000), 1385–1401.

[20] Luc S Cousineau. 2020. DISPLACED DISCUSSION: THE IMPLICATIONS OF REDDIT QUARANTINE AND THE MOVEMENT OF THEREDPILL TO SELF-HOSTING. *AoIR Selected Papers of Internet Research* (2020).

[21] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. (2016), 1389–1397. https://doi.org/10.1145/2851581.2892410

[22] 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

[23] Caitlin Dewey. 2014. Inside the 'manosphere' that inspired Santa Barbara shooter Elliot Rodger. *The Washington Post* (2014). https://www.washingtonpost.com/news/the-intersect/wp/2014/05/27/inside-the-manosphere-that-inspired-santa-barbara-shooter-elliot-rodger/

[24] Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring Misogyny across the Manosphere in Reddit. (2019).

[25] Casey Fiesler, Jialun "Aaron" Jiang, Joshua McCann, Kyle Frye, and Jed R Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Twelfth International AAAI Conference on Web and Social Media*.

[26] Claudia Flores-Saviaga, Brian C Keegan, and Saiph Savage. 2018. Mobilizing the Trump train: Understanding collective action in a political trolling community. *ICWSM* (2018).

[27] Ollie Garkey. 2016. Elliot Rodger, Gunman in California Mass Shooting, Was Influenced by the Men's Rights Movement. *Daily Kos* (2016). https://www.dailykos.com/stories/2014/05/24/1301671/-Elliot-Roger-Gunman-in-California-Mass-Shooting-was-influenced-by-the-Men-s-Rights-Movement

[28] R Kelly Garrett. 2009. Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of computer-mediated communication* 14, 2 (2009), 265–285.

[29] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (2018), 4492–4511. https://doi.org/10.1177/1461444818776611

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

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

[32] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data and Society* 7, 2 (2020). https://doi.org/10.1177/2053951720943234

[33] David Gilmour. 2017. 4chan and Reddit users set out to prove Seth Rich murder conspiracy. https://www.dailydot.com/layer8/reddit-the-donald-4chan-seth-rich-murder

[34] Debbie Ging. 2017. Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and Masculinities* (2017), 1097184X17706401.

[35] April Glaser. 2018. Want a Terrible Job? Facebook and Google May Be Hiring. https://slate.com/technology/2018/01/facebook-and-google-are-building-an-army-of-content-moderators-for-2018.html

[36] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society* 7, 1 (2020). https://doi.org/10.1177/2053951719897945

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

[38] Tim Hesterberg, David S Moore, Shaun Monaghan, Ashley Clipson, and Rachel Epstein. 2005. Bootstrap methods and permutation tests. *Introduction to the Practice of Statistics* 5 (2005), 1–70.

[39] Laura Hudson. 2018. The internet is enabling a community of men who want to kill women. They need to be stopped. https://www.theverge.com/2018/4/25/17279294/toronto-massacre-minassian-incels-internet-misogyny

[40] Zhane Hunte. 2019. "Female Nature, Cucks, and Simps": Understanding Men Going Their Own Way as part of the Manosphere.

[41] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized social signals: Computationally-derived social signals from account histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[42] Kathleen Hall Jamieson and Joseph N Cappella. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment.* Oxford University Press.

[43] Sue Curry Jansen and Brian Martin. 2015. The Streisand effect and censorship backfire. (2015).

[44] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "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, Article 192 (2019), 33 pages. https://doi.org/10.1145/3359294

[45] 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

[46] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proc. ACM Hum.-Comput. Interact.*

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

CSCW, Article 150 (2019), 27 pages. https://doi.org/10.1145/3359252

[48] 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

[49] Deborah G Johnson. 1997. Ethics online. *Commun. ACM* 40, 1 (1997), 60–65.

[50] Ridley Jones, Lucas Colusso, Katharina Reinecke, and Gary Hsieh. 2019. R/Science: Challenges and Opportunities in Online Science Communication. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.

[51] Matthew Kay, Gregory L Nelson, and Eric B Hekler. 2016. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4521–4532.

[52] Annie Kelly. 2017. The alt-right: Reactionary rehabilitation for white masculinity. *Soundings* 66, 66 (2017), 68–78.

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

[54] Jason Koebler. 2016. https://www.vice.com/en_us/article/53d5xb/what-is-rthedonald-donald-trump-subreddit

[55] 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

[56] Helene Odden Langås. 2019. *"Have ya seen some of the women we gotta rape?": Exploring discourses on rape in an antifeminist Internet forum.* Master's thesis.

[57] Matthew N Lyons. 2017. Ctrl-alt-delete: The origins and ideology of the alternative right. *Somerville, MA: Political Research Associates, January* 20 (2017).

[58] Alexis Madrigal. 2018. Inside Facebook's Fast-Growing Content-Moderation Effort. https://www.theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632/

[59] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. " You Know What to Do" Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[60] Alice E Marwick and Robyn Caplan. 2018. Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Studies* 18, 4 (2018), 543–559.

[61] Joan Massachs, Corrado Monti, Gianmarco De Francisci Morales, and Francesco Bonchi. 2020. Roots of Trumpism: Homophily and Social Feedbackin Donald Trump Support on Reddit. In *12th ACM Conference on Web Science* (Southampton, United Kingdom) *(WebSci '20)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/3394231.3397894

[62] Adrienne Massanari. 2017. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19, 3 (2017), 329–346.

[63] J Nathan Matias. 2016. Going dark: Social factors in collective action against platform operators in the Reddit blackout. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 1138–1151.

[64] J Nathan Matias and Merry Mou. 2018. CivilServant: Community-led experiments in platform governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 9.

[65] Thomas Mejtoft, Sarah Hale, and Ulrik Söderström. 2019. Design Friction. (2019), 41–44. https://doi.org/10.1145/3335082.3335106

[66] 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* (San Antonio, Texas, USA) *(CSCW '13)*. ACM, New York, NY, USA, 839–848. https://doi.org/10.1145/2441776.2441871

[67] Joseph Mountford. 2018. Topic modeling the red pill. *Social Sciences* 7, 3 (2018), 42.

[68] Angela Nagle. 2015. *An investigation into contemporary online anti-feminist movements.* Ph.D. Dissertation. Dublin City University.

[69] Peter R Neumann. 2013. Options and strategies for countering online radicalization in the United States. *Studies in Conflict & Terrorism* 36, 6 (2013), 431–459.

[70] Abby Ohlheiser. 2016. Fearing yet another witch hunt, Reddit bans 'Pizzagate'. https://www.washingtonpost.com/news/the-intersect/wp/2016/11/23/fearing-yet-another-witch-hunt-reddit-bans-pizzagate/

[71] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[72] Elinor Ostrom. 1990. *Governing the commons: the evolution of institutions for collective action.* Cambridge University Press, Cambridge.

[73] Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think.* Penguin.

[74] Robert Peck. 2017. The Hate Fueled Rise of r/The_Donald-and Its Epic Takedown. https://www.wired.com/story/the-hate-fueled-rise-of-rthe-donald-and-its-epic-takedown/

[75] Tal Peretz. 2016. Why Study Men and Masculinities? A Theorized Research Review. *Graduate Journal of Social Science* 12, 3 (2016).

[76] Casey Quinlan. 2019. It's no coincidence the synagogue shooter posted about the 'red pill movement'. https://thinkprogress.org/no-coincidence-california-synagogue-shooter-john-earnest-posted-about-the-red-pill-movement-aa46b31c01ad/

[77] John Rawls. 1971. *A Theory of Justice.* Oxford University Press.

[78] Help Reddit. 2020. Quarantined Subreddits. https://www.reddithelp.com/hc/en-us/articles/360043069012

[79] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Robert West. 2021. Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. In *CSCW*.

[80] Sarah T Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press.

[81] Richard Rogers. 2020. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication* (2020), 0267323120922066.

[82] santaclaraprinciples.org. 2021. Santa Clara principles on transparency and accountability in content moderation. https://santaclaraprinciples.org/

[83] Rachel Schmitz and Emily Kazyak. 2016. Masculinities in cyberspace: An analysis of portrayals of manhood in men's rights activist websites. *Social Sciences* 5, 2 (2016), 18.

[84] 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* (Portland, Oregon, USA) *(CSCW '17).* ACM, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277

[85] 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.

[86] Qinlan Shen and Carolyn Rose. 2019. The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy. In *Proceedings of the Third Workshop on Abusive Language Online.* 58–69.

[87] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[88] Bijan Stephen. 2018. Reddit updates its quarantine policy with an appeals process. https://www.theverge.com/2018/9/28/17914240/reddit-update-quarantine-policy-appeals-process

[89] Emily Stewart. 2019. Reddit restricts its biggest pro-Trump board over violent threats. https://www.vox.com/recode/2019/6/26/18760288/reddit-the-donald-trump-message-board-quarantine-ban

[90] 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.

[91] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web.* 613–624.

[92] TL Taylor. 2018. Regulating the networked broadcasting frontier. In *Watch me play: Twitch and the rise of game live streaming.* Princeton University Press, Chapter 5.

[93] u/landoflobsters. 2018. Revamping the Quarantine Function. https://www.reddit.com/r/announcements/comments/9jf8nh/revamping_the_quarantine_function

[94] Stefanie Ullmann and Marcus Tomalin. 2020. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology* 22, 1 (2020), 69–80.

[95] Christopher Vito, Amanda Admire, and Elizabeth Hughes. 2018. Masculinity, aggrieved entitlement, and violence: considering the Isla Vista mass shooting. *NORMA* 13, 2 (2018), 86–102.

[96] 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.

[97] Anne Weber. 2009. *Manual on hate speech.* Council Of Europe.

[98] Mark Wilson. 2019. Pinterest is escalating its fight against anti-vaxxers as measles surge in the U.S. https://www.fastcompany.com/90396075/pinterest-is-escalating-its-fight-against-anti-vaxxers-as-measles-surge-in-the-u-s

[99] Donghee Yvette Wohn. 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.* ACM, 160.

[100] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference*. 405–417.

[101] 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