INTERNET POLICY REVIEW
Journal on internet regulation

Volume 9 | **Issue** 4

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Reddit quarantined: can changing platform affordances reduce hateful material online?

Simon Copland *Australian National University* simon.copland@anu.edu.au

DOI: https://doi.org/10.14763/2020.4.1516

Published: 21 October 2020
Received: 20 May 2020 **Accepted:** 6 August 2020

Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Copland, S. (2020). Reddit quarantined: can changing platform affordances reduce hateful material online?. *Internet Policy Review*, *9*(4). DOI: 10.14763/2020.4.1516

Keywords: Reddit, Affordances, Misogyny, Manosphere, Digital platforms, Online platforms

Abstract: This paper studies the efficacy of the Reddit's quarantine, increasingly implemented in the platform as a means of restricting and reducing misogynistic and other hateful material. Using the case studies of r/TheRedPill and r/Braincels, the paper argues the quarantine successfully cordoned off affected subreddits and associated hateful material from the rest of the platform. It did not, however, reduce the levels of hateful material within the affected spaces. Instead many users reacted by leaving Reddit for less regulated spaces, with Reddit making this hateful material someone else's problem. The paper argues therefore that the success of the quarantine as a policy response is mixed.

This paper is part of **Trust in the system**, a special issue of *Internet Policy Review* guest-edited by Péter Mezei and Andreea Verteş-Olteanu.

Content moderation is an integral part of the political economy of large social media platforms (Gillespie, 2018). While social media companies position themselves as *platforms* which offer unlimited potential of free expression (Gillespie, 2010), these same sites have always engaged in some form of content moderation (Marantz, 2019). In recent years, in response to increasing pressure from the public, lawmakers and advertisers, many large social media companies have given up much of their free speech rhetoric and have become more active in regulating abusive, misogynistic, racist and homophobic language on their platforms. This has occurred in particular through banning and restricting users and channels (Marantz, 2019). In 2018 for example, a number of large social media companies banned the high-profile conspiracy theorist Alex Jones and his platform InfoWars from their platforms (Hern, 2018), while in 2019 the web infrastructure company Cloudflare deplatformed the controversial site 8chan (Prince, 2019). In 2020 a number of platforms even began regulating material from President Donald Trump, with Twitter placing fact-checks and warnings on some of his tweets and the platform Twitch temporarily suspending his account (Copland and Davis, 2020).

As one of the largest digital platforms in the world, Reddit has not been immune from this pressure. Built upon a reputation of being a bastion of free speech (Ohanian, 2013), Reddit has historically resisted censoring its users, despite the prominence of racist, misogynistic, homophobic and explicitly violent material on the platform (for examples, see Massanari, 2015, 2017; Salter, 2018; Farrell, Fernandez, Novotny, and Harith, 2019). In 2011 for example the former general manager of Reddit, Erik Martin, addressed growing controversies over hateful content, stating:

> We're a free speech site with very few exceptions (mostly personal info) and having to stomach occasional troll reddit (sic) like /r/picsofdeadkids or morally questionable reddits like /r/jailbait are part of the price of free speech on a site like this. (cited in Robertson, 2015b)

However, under increasing pressure from its users, advertisers, lawmakers and the general public in recent years Reddit has slowly begun to shift this approach (Copland and Davis, 2020). Since 2012 Reddit has slowly changed its content policies, including banning: any suggestive or sexual content featuring minors (u/reddit, 2012); the sharing of nude photos without the subject's consent (u/kn0thing, 2015 [1]); attacks and harassment of individuals (u/5days et al., 2015); the incitement of violence (u/landoflobsters, 2017); and attacks and harassment of broad social groups (u/landoflobsters, 2019). Reddit's argument, that it was a free speech site that would not intervene, has slowly come undone.

Reddit has frequently found itself in an uncomfortable position, straddling a fine line between maintaining a culture of free speech, while at the same time not allowing the flourishing of violence and hateful material (Copland, 2018). Reddit has faced dueling pressures from a user base, which finds Reddit's *anything goes* approach as core to the platform's appeal, versus other users who increasingly wish to constrain much of the more extreme material on the site (Massanari, 2015).

In doing so, in addition to the content policies identified above, Reddit has also implemented a unique approach to dealing with hateful content on the platform, that is the quarantine function.

Content moderation from large digital platforms, in particular in regards to hateful material, comprises an array of different techniques (Gillespie, 2018). As noted already, many platforms have become more active in deleting hateful content and banning or suspending the profiles of far-right speakers. In addition to this, many platforms have begun to engage in fact-checking processes in order to stop the spread of disinformation. Facebook, for example, has created an independent fact-checking unit (Facebook, 2020), while Twitter made headlines by fact-checking a Tweet from President Donald Trump (BBC News, 2020). Platforms also rely on the labour of users themselves in order to moderate content (Matias, 2019). Many platforms, including Reddit, contain communities which self-moderate, with volunteer moderators regulating content based on community-defined rules. Users also frequently play the role of content monitors, reporting content to the platform for moderation to occur (Gillespie, 2018).

Reddit's quarantine function is unique from these approaches as it is designed to limit the accessibility of users to identified subreddits, and in turn to limit the spread of questionable content, while not banning the content outright. A quarantine uses the platform's affordances to discourage the spread of particular material and to encourage positive behaviour change within identified *shocking* subreddits.

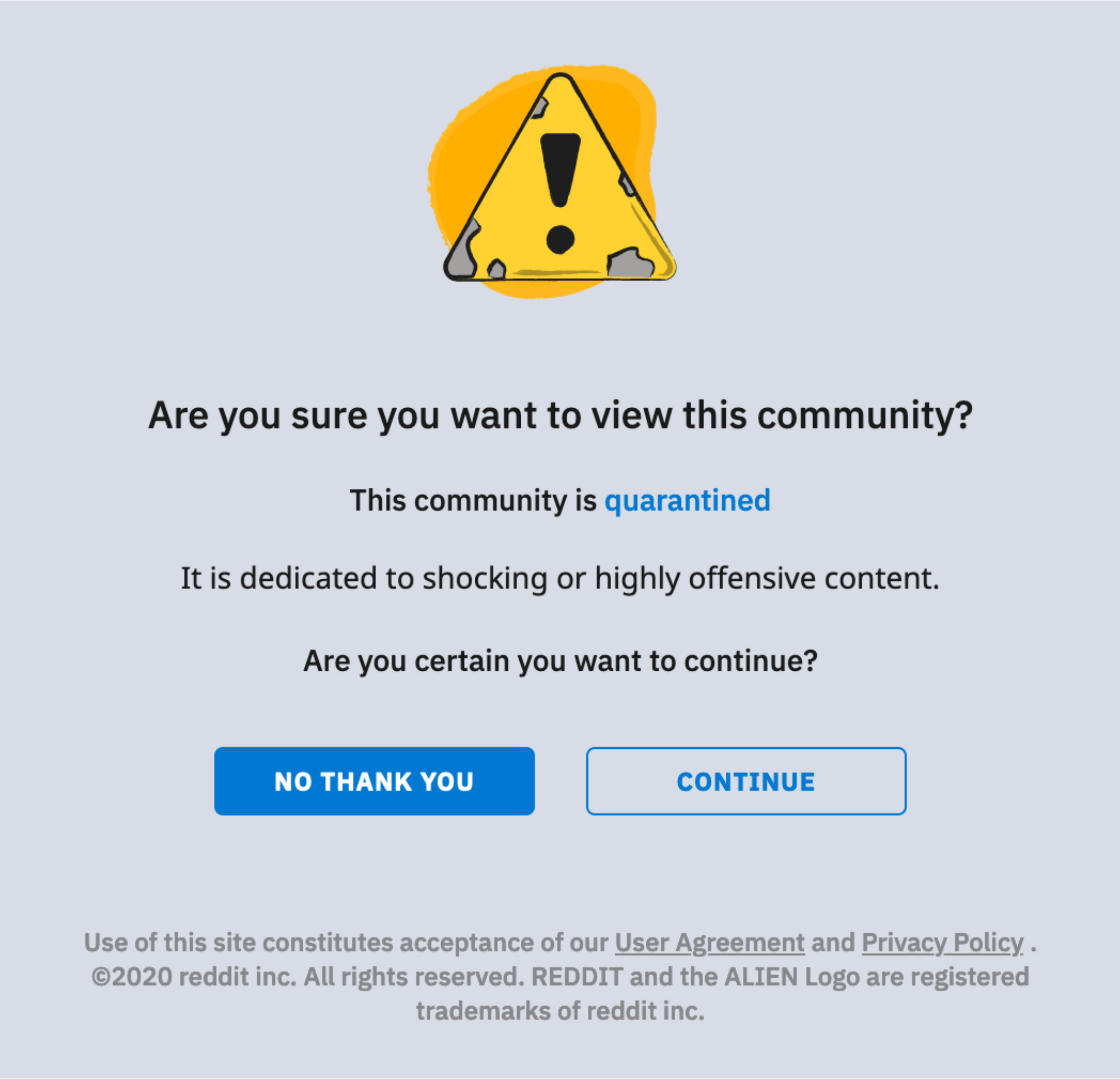
[1] On Reddit users create profile names, which are normally pseudonymous. Users are then identified by a u/(username) – i.e., u/kn0thing. I therefore reference content posted on Reddit by the username, identifying the author, followed by the year of the post or comment.

The quarantine therefore provides a useful moment to examine the efficacy of the use of bans and restrictions on particular content in both limiting the spread of hateful content as well as encouraging changes in behaviour within a platform. Using a case study of two men's rights subreddits quarantined after the revamp of the function in September 2018 (u/landoflobsters, 2018), r/TheRedPill and r/Braincels, this paper examines the efficacy of this approach in dealing with misogynistic content on the platform.

Using data from these two subreddits the paper argues that the quarantine was effective in limiting engagement with these subreddits and in turn limiting the exposure of the broader Reddit community to the misogynistic material within. In particular, the quarantine saw significant drops in activity within the targeted subreddits. However, I argue that despite this, the quarantine had little to no effect in the use of hateful language within the targeted subreddits, with Reddit failing to encourage remaining users to change their behaviour. Instead, the quarantine furthered animosity between users of these subreddits and Reddit overall, in particular resulting in concerted pushes for users to leave Reddit, to instead participate in self-moderated, and in turn less restrictive, platforms. Aligning with research on subreddit bans from Chandrasekharah, Pavalanathan, Srinivasan, Glynn, Eisenstein and Gilbert (2017, p. 18) I argue that the quarantine in turn, potentially made hateful material *someone else's problem*. While the policy reduced access to hateful material on the platform, it did not necessarily do so on the internet overall, instead potentially pushing it to less moderated spaces where it can continue unchecked. As a policy response therefore, this paper argues that the quarantine had mixed results.

The quarantine

On 28 September 2018, Reddit announced a revamp of their quarantine function (u/landoflobsters, 2018). A quarantine implements a range of restrictions on a subreddit. Quarantined subreddits are unable to generate revenue, and their content does not appear on Reddit's front page, nor on other non-subscription based feeds. Quarantined subreddits cannot be found via the search or recommendation function, meaning users need to either search for the subreddit via Google or know the subreddit URL to access it. Quarantined subreddits also do not appear on a Reddit user's subscription feed, requiring subscribers of that subreddit to note that it has been quarantined and then actively seek it out to once again put it in their feed. When a user does attempt to gain access a warning is displayed (see Figure 1), informing users that the subreddit is dedicated to "shocking or highly offensive content” and requiring the user to explicitly opt-in to participate.



Are you sure you want to view this community?

This community is quarantined

It is dedicated to shocking or highly offensive content.

Are you certain you want to continue?

Use of this site constitutes acceptance of our User Agreement and Privacy Policy . ©2020 reddit inc. All rights reserved. REDDIT and the ALIEN Logo are registered trademarks of reddit inc.

FIGURE 1: Quarantine warning

As already noted, the implementation of the quarantine was part of a broader sweep of policy changes from the platform, ones which shifted Reddit away from their unlimited free-speech ethos. These shifts occurred primarily for two reasons. First, Reddit has not been immune to the significant social pressure facing social media companies over their role in public discourse, particularly following the 2016 US presidential election (Marantz, 2019). This has resulted in the co-founders of Reddit Alexis Ohanian and Steve Huffman to publicly walk back much of their early ideas about how Reddit should operate as they reckoned with the influence of their platform (Marantz, 2019). In addition to this, Reddit has begun to rely more heavily on advertising for revenue (Castillo, 2018). This shift brought particular challenges, with advertisers expressing concerns about being associated with the hateful material that appears on the platform. As the forecasting director for eMarketer Monica Peart has argued:

> As a mix of forum and trending news site with a bit of social network, reddit has operated on an ‘open internet’ ethos. While that has yielded organic growth among a hard-to-reach audience, it has also meant a reality where controversial content is the norm. And in a news climate where missteps can tarnish results, that makes some digital advertisers nervous. (Peart, 2019)

The quarantine is an attempt to restrict hateful material through an intervention into the affordances of a digital platform. As Davis and Chouinard (2016, p. 242) argue, an “affordance refers to the range of functions and constraints that an object provides for, and places upon, structurally situated subjects”. In relation to studies of digital platforms, affordances are the functions of the platform that shape how it can, and cannot, be used. Davis and Chouinard (2016) provide categories to allow us to understand how affordances work. These mechanisms define the different ways in which the functions of an artefact define the usage of that artefact. Davis and Chouinard propose that affordances request, demand, allow, encourage, discourage, and refuse:

> Requests and *demands* refer to bids that the artefact places upon the subject. *Encouragement, discouragement,* and *refusal* refer to how the artefact responds to a subject’s desired actions. *Allow* pertains to both bids placed upon on the subject and bids placed upon the artefact. (Davis and Chouinard, 2016, p. 242)

The quarantine function represents an attempt from Reddit to use the architecture of the platform to both discourage and encourage particular behaviours from users. The intent of quarantining subreddits was twofold. Reddit aimed first to limit the spread of content from these spaces, discouraging engagement with identified shocking content. Secondarily however, Reddit also aimed to use the quarantine as a way to encourage positive change within a community, hoping that the process of shutting off communities would encourage users to change their behaviour as a way to re-engage with the broader space. As u/landoflobsters said:

> The purpose of quarantining a community is to prevent its content from being accidentally viewed by those who do not knowingly wish to do so, or viewed without appropriate context. We’ve also learned that quarantining a community

> may have a positive effect on the behavior of its subscribers by publicly signaling that there is a problem. This both forces subscribers to reconsider their behavior and incentivizes moderators to make changes. (u/landoflobsters, 2018)

Case studies

Since announcing the revamp of the quarantine function in September 2018, Reddit administrators became more active in using the function (Copland and Davis, 2020). On the day of the announcement Reddit quarantined two large men's rights subreddits r/TheRedPill and r/Braincels (which has since been banned outright). In other examples, in 2019 Reddit quarantined another men's rights subrerddit r/MGTOW, the subreddit dedicated to the support of Donald Trump, r/The_Donald, and left wing subreddits such as r/communism and r/ChapoTrapHouse. Both r/The_Donald and r/ChapoTrapHouse were banned in mid-2020 (Copland and Davis, 2020).

This paper studies the quarantining of two manosphere subreddits, r/TheRedPill and r/Braincels, both of which were quarantined on the same day as the announcement of the revamp of the quarantine function in September 2018. r/TheRedPill and r/Braincels are associated with a broad online community called the manosphere, which, primarily focused on discussions around sex and relationships, is broadly anti-feminist in its ideology (Marwick and Caplan, 2018). The manosphere is based in an ideology, described by Nicholas and Agius (2018) as masculinism, which argues that men have become subjugated by feminism, with this subjugation in particular impacting men's access to sexual and romantic relationships.

r/TheRedPill and r/Braincels engage in these questions in two different ways. Members of r/TheRedPill focus on self-help and learning *game*; a range of techniques used to identify and pick up women confidently. Members of r/Braincels, part of a broad sub-community called *incels*, believe that due to a range of factors, primarily associated with their looks, that women refuse to engage in romantic or sexual relationships with them. These men are, by and large, resigned to their singledom (although some hold the more extreme view that they are owed sex by women), and come to manosphere forums to complain and bond about their lack of relationships with women, as well as to complain about women in general. Both subreddits have been identified in literature to contain high levels of misogynistic and other hateful language (e.g., Farrell et al., 2019; Van Valkenburgh, 2019).

It is this misogyny in particular which resulted in the administrators of Reddit quarantining the subreddits. When implementing the quarantine Reddit administrators sent the moderators of each subreddit messages informing them of why they have been quarantined and the avenues they had to appeal the decision. One of the volunteer moderators of r/TheRedPill, u/redpillschool, published the message sent to them. The message from Reddit Administrators was headlined, “This community is quarantined: It is dedicated to shocking or highly offensive content. For information on positive masculinity, please see the resources available at Stony Brook University's Center for the Study of Men and Masculinities,” with the following explanation provided:

> In this case the quarantine was applied for the high degree of misogyny present in this subreddit. To be removed from quarantine, you may present an appeal here. The appeal should include a detailed accounting of changes to community moderation practices. (Appropriate changes may vary from community to community and could include techniques such as adding more moderators, replacing certain moderators, creating new rules, employing more aggressive auto-moderation tools, adjusting community styling, etc.) The appeal should also offer evidence of sustained, consistent enforcement of these changes over a period of at least one month, demonstrating meaningful community transformation. (u/redpillschool, 2018)

Alongside the broader goal of restricting access to these subreddits, Reddit therefore articulated explicitly that it wanted to see a reduction of misogyny within these spaces, presenting this as the way in which both subreddits could exit their quarantine status.

Data analysis

Using these two case studies, the purpose of this paper is to examine the internal and external effects of this policy. First, I study whether the quarantine was an effective way to achieve Reddit’s stated goals of discouraging engagement with hateful material and encouraging shifts towards more positive content within subreddits that foster this material. Beyond this, the article asks, can shifting the affordances of a platform for particular users help reduce misogynistic and hateful material on the internet external to Reddit?

As part of a broader PhD research project I have used the website pushshift.io to download all submissions and comments from both r/TheRedPill and r/Braincels from the start of November 2017 to the end of December 2018. The implementation of the quarantine occurred during my data collection period, providing an opportunity to study the quarantine with an existent large data set. I have examined this data set in three different ways in order to understand the efficacy of the quarantine as a policy approach.

User activity level

I start by examining how the quarantine impacted the level of engagement within affected subreddits. This allows me to see whether Reddit achieved their first goal, which was to limit the exposure of the broader Reddit community from the hateful content within these spaces, and in turn to limit the spread of this material to other parts of Reddit. To analyse this, I created comment frequency histograms of the two affected subreddits (Figure 2 and Figure 3). These histograms examine the combined number of original submissions and comments on these submissions in each subreddit per day, across the collection period for the entire data set from November 2017 to December 2018.

We can see distinct trends in both histograms. Figure 2 is a histogram of the number of submissions and comments per day for r/TheRedPill. r/TheRedPill is established at the beginning of the data collection, with activity levels already high at this point. At its peak across the data collection r/TheRedPill has just over 1,500 submissions and comments on a single day. Submission and comment numbers on r/TheRedPill fluctuates widely on a day to day basis, with frequent highs and troughs. Activity level however drops significantly after the implementation of the quarantine. Despite continued peaks and troughs, these levels overall are approximately one half of what occurred before the quarantine.

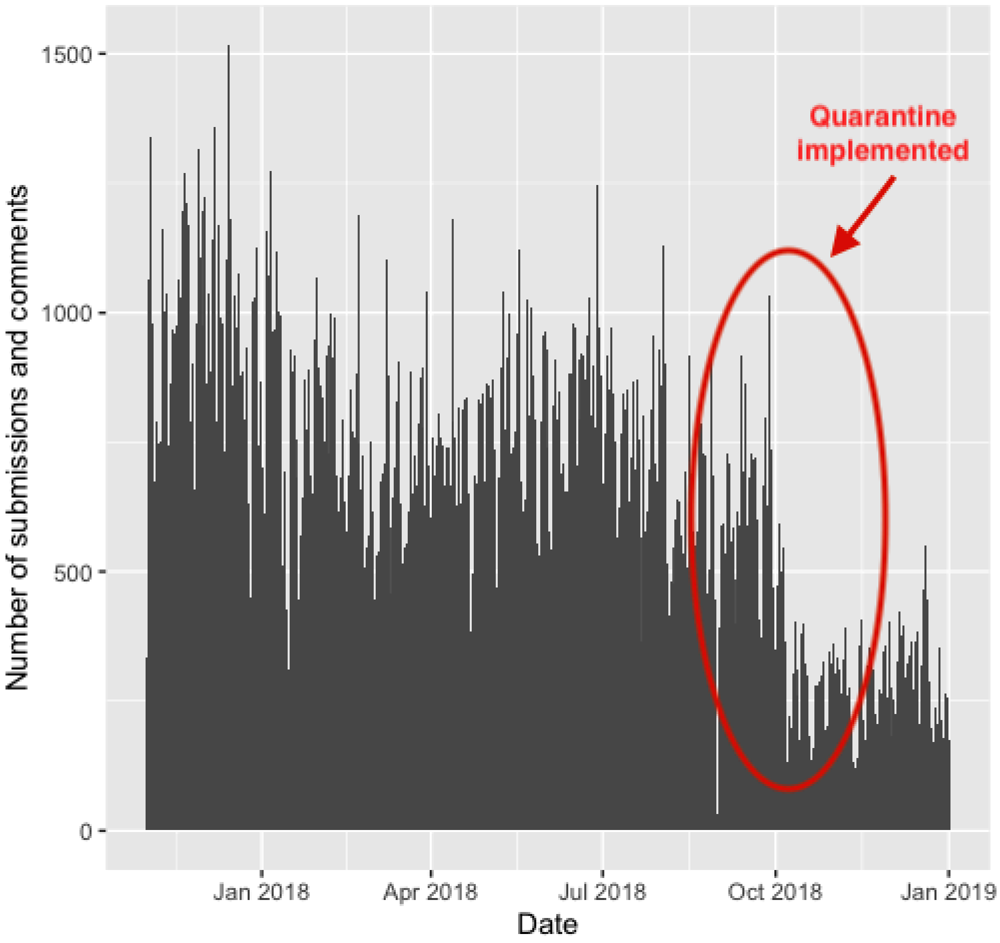


FIGURE 2: The Red Pill submissions and comments frequency histogram

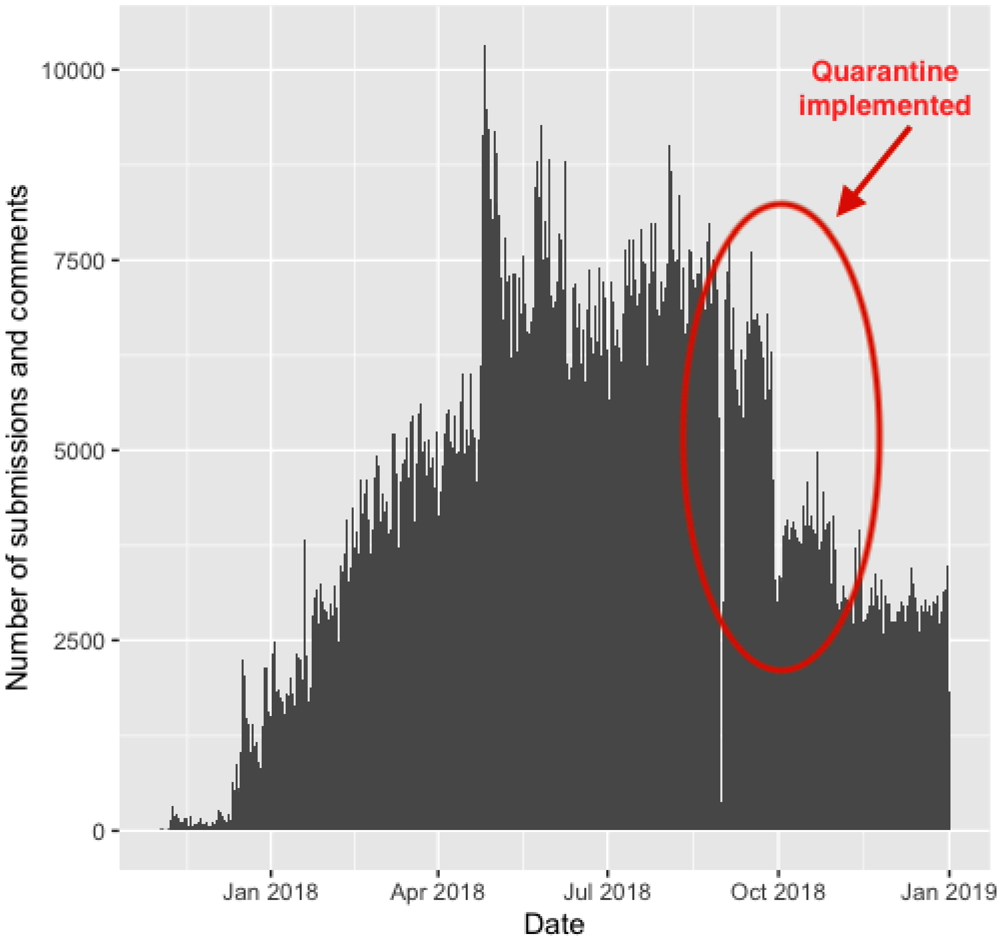


FIGURE 3: The Braincels submissions and comments frequency histogram

Similar patterns occur in r/Braincels. r/Braincels was created in November 2017, at the very beginning of my data collection. We can see a gradual increase in activity in r/Braincels from this period, until a peak in May and June of 2018. On the busiest day of the subreddit there were over 10,000 combined submissions and comments. Activity on r/Braincels stays close to this level until the implementation of the quarantine, after which levels drop significantly. At this moment r/Braincels loses about a half of its activity, dropping to under 5,000 combined submissions and comments per day. Activity levels rise briefly following this significant drop, but then continue to decline toward the end of 2018.

Based on this analysis we can see that, at least in the immediate term, the quarantine achieved Reddit's first goal – i.e., to reduce engagement with particular subreddits and in turn exposure to these subreddits from the broader Reddit community. We can see that this occurred in both affected subreddits, both of which received a drop in activity of about 50%. This analysis therefore shows that the quarantine had an initial, immediate, success. However, while it shows that these subreddits were cordoned off effectively from the rest of Reddit, this does not necessarily prove that it reduced hateful content on Reddit overall. It is entirely possible that users of the affected subreddits simply shifted their hateful content to other parts of the platform. It is unfortunately outside of the scope of this paper to examine these potential broader shifts.

Misogynistic language analysis

While Reddit effectively cordoned off quarantine communities, significantly reducing engagement within them, the platform also had a second goal, which was to encourage users who stayed, to change their behaviour. Therefore, I will now shift to examining the discourse within the affected subreddits to see whether this occurred. What I am looking for here is whether there was a reduction in misogynistic and other hateful language within the two affected subreddits following the quarantine. Any reduction would suggest that the implementation of the quarantine resulted in participants reflecting on the content of the subreddit and making positive changes.

I have conducted an analysis of the levels of misogynistic material in both affected subreddits in the two months prior to and directly after the implementation of the quarantine. I have used a linguistic inquiry approach utilising the software LIWC (Linguistic Inquiry and Word Count) and a dictionary of misogynistic terms created by Farrell et al. (2019). A linguistic approach searches through a data set and calculates the prominence of particular types of terms and language within that data set. These terms and phrases are predetermined through the creation of a *dictionary*. In their analysis of misogyny across the manosphere on Reddit, Farrell et al. (2019) created a dictionary of misogyny comprising 1,168 terms and phrases. Farrell et al. developed this dictionary by both merging other examples of dictionaries of hateful material, alongside identifying misogynistic terms that are specific to manosphere communities. This dictionary is broken up into nine subcategories: *belittling*; *flipping the narrative*; *homophobia*; *hostility*; *patriarchy*; *physical violence*; *racism*; *sexual violence*; and *stoicism* (details of each sub-category can be found in Appendix 1). This approach has some limitations, in particular, that the context of the use of words is not taken into account. Farrell et al.'s dictionary for example includes the word *queer* in the homophobia subcategory. While still used as a slur against LGBTIQ people, queer is also a term that has been reclaimed by many members of this community. This approach does not distinguish between these two uses. Despite this however, a linguistic inquiry approach can provide an insight into the type of language used in a community, and in this case whether there is a high prevalence of terms closely associated with misogynistic and other hateful material.

I have applied a linguistic approach to study the prevalence of misogynistic language with each subreddit in the two months prior to and after the implementation of the quarantine, totalling four months of data analysis. This is done to examine whether there were shifts in the use of hateful language across this period, and in turn whether Reddit was successful in encouraging users of the subreddit to change their behaviour and language. The results, presented in Tables 1 and 2, present the total number of submissions and comments that contain terms from the misogynistic dictionary, as well as the prevalence of each subcategory as a percentage of the total number of submissions and comments.

TABLE 1: Misogynistic language analysis for r/TheRedPill

R/THEREDPILL	PRE QUARANTINE	POST QUARANTINE
TOTAL NUMBER OF SUBMISSIONS AND COMMENTS	38,344	18,409
TOTAL MISOGYNISTIC SUBMISSIONS AND COMMENTS	9,504 (24.78%)	4,710 (25.58%)
BELITTLING	1,836 (4.78%)	991 (5.38%)
FLIPPING THE NARRATIVE	1,187 (3.09%)	723 (3.92%)
HOMOPHOBIA	159 (0.41%)	64 (0.34%)
HOSTILITY	4,347 (11.33%)	1,950 (10.59%)
PATRIARCHY	86 (0.22%)	51 (0.27%)
PHYSICAL VIOLENCE	2,393 (6.24%)	1,283 (6.96%)
RACISM	671 (1.74%)	317 (1.72%)
SEXUAL VIOLENCE	420 (1.09%)	323 (1.75%)
STOICISM	1,295 (3.37%)	582 (3.16%)

TABLE 2: Misogynistic language analysis for r/Braincels

R/BRAINCELS	PRE QUARANTINE	POST QUARANTINE
TOTAL NUMBER OF SUBMISSIONS AND COMMENTS	405,955	210,548

R/BRAINCELS	PRE QUARANTINE	POST QUARANTINE
TOTAL MISOGYNISTIC SUBMISSIONS AND COMMENTS	111,845 (27.55%)	58,907 (27.97%)
BELITTLING	13,659 (3.36%)	7,085 (3.36%)
FLIPPING THE NARRATIVE	7,345 (1.80%)	3,565 (1.69%)
HOMOPHOBIA	2,627 (0.64%)	1,473 (0.69%)
HOSTILITY	30,423 (7.49%)	15,714 (7.46%)
PATRIARCHY	581 (0.14%)	430 (0.20%)
PHYSICAL VIOLENCE	10,101 (2.48%)	5,685 (2.70%)
RACISM	10,122 (2.49%)	4,601 (2.18%)
SEXUAL VIOLENCE	2,273 (0.55%)	1,420 (0.67%)
STOICISM	58,983 (14.52%)	32,283 (15.33%)

Prior to the quarantine, misogynistic language appeared in 24.74% and 27.55% of submissions and comments in r/TheRedPill and r/Braincels respectively. In the two months prior to the quarantine r/TheRedPill had a total of 9,504 submissions or comments containing misogynistic language, while r/Braincels had a total of 111,845. These vast differences largely represent the different volume of content within each subreddit. The most common form of misogyny within r/TheRedPill was hostility, being prevalent in 11.33% of submissions and comments. For r/Braincels, stoicism appeared in 14.52% of submissions and comments. This is unsurprising as the stoicism category contains a number of terms directly associated with incel communities, including *incel*, *volcel*, *ricecel*, *truecel*, *rope* and *cope*.

Alongside the overall reduction in activity in both subreddits, there is a clear reduction in the raw number of submissions and comments containing misogynistic language in the immediate aftermath of the quarantine. r/TheRedPill sees a drop in the number of misogynistic submissions and comments from 9,504 to 4,710, and r/Braincels sees a similar drop from 111,854 to 58,907. However, this does not result in any shift in the density of this language within those who remained in the subreddit. Both subreddits see a slight increase in the percentage of submissions and comments containing misogynistic language with this increasing from 24.78% to 25.58% in r/TheRedPill and 27.55% to 27.97% in r/Braincels. Both of these shifts are too small however to suggest that we can draw any significant conclusions from these changes. Similar to the period prior to the implementation of the quarantine, hostility and stoicism were the most common forms of misogyny in r/TheRedPill and r/Braincels respectively.

Therefore, similar to engagement overall there is a sharp reduction–approximately a halving–in the raw number of submissions and comments containing misogynistic language within the subreddits. However, those who remained engaged with the subreddits did not change their language. The percentage of the misogynistic material stayed the same in both subreddits, suggesting that existent users continued using the same language as before, only at a smaller scale. Importantly however, there is also no significant increase in misogynistic language in either subreddit. While Reddit successfully reduced engagement with both subreddits from the broader Reddit community, they did not encourage changes in the use of language within either of these subreddits–either negatively or positively.

Responses to the quarantine

My first two areas of analysis studied the outcomes of the particular goals of the quarantine identified by Reddit itself. However, the impact of the quarantine did not just occur in these two areas. In particular, following its implementation, both affected subreddits saw significant levels of discussion about the quarantine, in particular resulting in debates about whether and how to stay engaged with Reddit. Therefore, I have finally conducted an analysis of how users have responded directly to the implementation of the quarantine through studying discussion of the quarantine. This reaction indicates how users responded to the implementation of the quarantine, and in turn can provide insights both into the failure of Reddit's second goal, as well as the potential long-term impacts of the quarantine policy.

I have pulled all submissions or comments that mention either the term *quarantine* or *quarantined* from both subreddits, ranging from the date of the quarantine (28/09/2018) until the end of my data collection (31/12/2018). The vast majority of these submissions and comments occurred in the first week after the quarantine was implemented. I have cleaned this data, ensuring all posts and comments are related to the policy. This has resulted in a total of ten submissions and 904 comments related to the quarantine in r/TheRedPill and 22 submissions and 728 comments related to the quarantine in r/Braincels. I have conducted a qualitative reading of these submissions and comments, reading all submissions alongside the top twenty comments by score in order to identify key themes. This provides a useful approach to understanding how users responded to the quarantine, and in turn its potential efficacy in encouraging positive behaviour changes within members of these subreddits.

r/TheRedPill

The quarantine was met with immediate anger from members and moderators of r/TheRedPill, who argued that the quarantine represented an attack on their community and on men as a whole.

This manifested in two ways. First, subreddit moderators openly challenged and discredited the logic behind the quarantine and petitioned Reddit to either provide more evidence behind why r/TheRedPill deserved to be quarantined, or to overturn the implementation of the quarantine altogether. The most highly engaged of these posts was from one of the most senior members of the subreddit, u/redpillschool. In a post titled "300,000 Subscribers. The Reddit Administration Tacitly Endorses Male Abuse and Denies its Victims", u/redpillschool engaged in a long rebuttal of the reasons behind the implementation of the quarantine. u/redpillschool in particular targeted the reference to Stonybrook University in the initial letter from Reddit moderators, arguing that, through denying the existence of male victims of domestic violence, the head of Stonybrooks' Centre for Men and Masculinity, Michael Kimmel, tacitly endorses abuse toward men. He then turned their attention to the co-founder of Reddit Steve Huffman, who goes by the username u/spez.

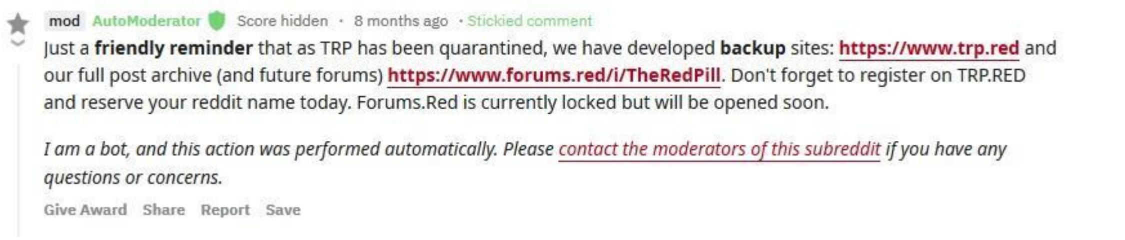
u/redpillschool argued:

> I challenge the admin, /u/spez, or anybody on their team to engage me in courteous, reasoned discussion to defend their assertions and endorsements this week, or to apologize and lift the quarantine. I promise to keep it 100% civil in an all-text debate of ideas with anybody on the reddit administration. I get the feeling nobody is going to be taking me up on this offer. This would be the action of somebody with integrity. I'm sure Michael Kimmel could make an argument that integrity is part of positive masculinity. /u/spez**, you don't want to be toxic do you?**(u/redpillschool, 2018b)

The quarantine thus resulted in a significant split between moderators and users of r/TheRedPill and Reddit as a platform. This led to the second strategy used by r/TheRedPill moderators. Through participating in attacks on men through the quarantine, r/TheRedPill moderators argued that Reddit could no longer be trusted as a space in which the community could gather. There are a number of comments in particular in the data set, in which users said that they believed that the quarantine was simply the starting point, and that Reddit would soon ban the subreddit altogether.

In response to these fears, the moderators of r/TheRedPill directed users to a self-moderated platform outside of Reddit [2]. This is an independently established and self-moderated forum with similar site infrastructure to Reddit, but with a sole focus on Red Pill ideas. The forum is moderated by the same moderators as r/TheRedPill. Users of r/TheRedPill were actively encouraged to make profiles with the same username they had on r/TheRedPill, ensuring continuity of discussion.

This push occurred primarily through a recurring comment by the *AutoModerator* (see Figure 5). Subreddit moderators have the capacity to configure AutoModerators to automatically remove particular posts or comments, or to reply to posts or comments with predetermined messages. Starting eleven days after the implementation of the quarantine, moderators of r/TheRedPill configured the AutoModerator to reply as the first, pinned, comment on every post informing users that the subreddit has been quarantined and directing users to the external Red Pill site (see Figure 4).



Just a **friendly reminder** that as TRP has been quarantined, we have developed **backup** sites: **https://www.trp.red** and our full post archive (and future forums) **https://www.forums.red/i/TheRedPill**. Don't forget to register on TRP.RED and reserve your reddit name today. Forums.Red is currently locked but will be opened soon.

I am a bot, and this action was performed automatically. Please contact the moderators of this subreddit if you have any questions or concerns.

FIGURE 4: the red pill AutoModerator message

This push did not just occur through the messages of the AutoModerator. To show this Figure 5 presents a URL analysis of submissions and comments from the quarantine data set. The use of a URL in a submission or comment is done primarily to direct users to outside sources, either as links to outside news stories, videos or other content, or in this instance, as a way to direct users to outside forums. I first removed all the comments from the AutoModerator, as these comments would skew the results of any analysis from other users. This left a total of 22 submissions and 243 comments. Using the R package Quanteda, which is a programming tool for the quantitative analysis of textual material, I isolated URLs referenced in submissions and comments. I then developed a frequency plot to identify the web-

[2] The Red Pill Network

sites to where users were most commonly directing others. One limitation of this analysis is that it does not include hyperlinks. This URL analysis however can provide a base analysis of where users were being directed to in the aftermath of the quarantine.

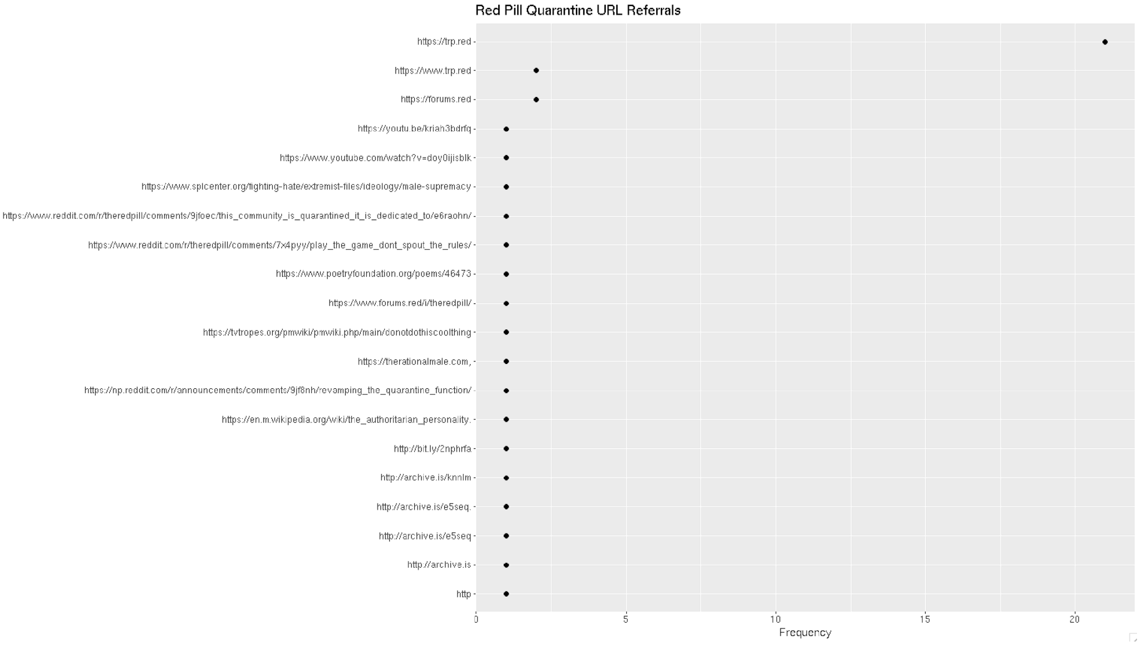


FIGURE 5: Red pill quarantine url referrals

Alongside the message from AutoModerator, what we can see in this analysis is a push to the outside forum, alongside its sub-pages. Out of the 22 submissions and 243 comments, users were directed to this forum a total of 25 times. No other website was referred to more than once. While not a significant movement, it indicates an effort from users to direct others to the outside forum. Qualitative analysis of this data indicates this was driven in particular by moderators and other highly-regarded users, with the push coming from the top down.

This does not mean that users necessarily flooded to this outside platform, and unfortunately I am unable to conduct analysis to see how many actually followed the moderators directions and left Reddit for the external forum. This is because I do not have the relevant user data for this outside forum, and in turn cannot trace user migration from Reddit to this platform. However, research on previous bans of hateful subreddits have shown the bans have resulted in a large number of users of affected subreddits moving to different, often less regulated, platforms (Chandrasekharan et al., 2015; Newell et al., 2016). These shifts occurred without the concerted efforts to do so as we saw in r/TheRedPill following the quarantine. While the specific data is not available in this case, it is reasonable to think that the quarantine resulted in some, or many, users of the subreddit moving to this external platform.

r/Braincels

While the responses from r/TheRedPill was coherent and organised, primarily emanating from moderators and other senior figures, the response from r/Braincels was blurrier. The first responses to the quarantine for example do not come from the moderators, but instead from regular users. Users quickly started asking questions and suggesting potential ways forward. Some for example, asked whether users should go elsewhere, others suggested it may end up being good for the subreddit as it would mean only dedicated incels would stay in the space, and some highlighted fears that the quarantine would eventually result in the subreddit being banned (a fear that was realised when r/Braincels was banned from Reddit in October 2019).

When the moderators engaged they did not suggest any means forward, instead providing basic information. The first engagement from moderators came from u/Cristalline144hz, who on the day the quarantine was implemented posted a range of basic information about what it meant. This was followed by a lively discussion about how r/Braincels should respond. Similar to arguments prominent in r/TheRedPill, some, such as u/Uqtpa argued that the quarantine represented the *misandry* of Reddit, which was unfairly targeting men. As they argued:

> *It's ridiculous. There is far more misandry in subreddits such as againstmensrights, therealmisandry, inceltears, twoxchromosomes, feminism, askfeminists and trollxchromosomes. All those subreddits have displayed way more misandry than this subreddit has displayed misogyny. And there are probably more misandric subreddits, yet this subreddit is the one that gets quarantined and probably deleted soon. Men aren't even allowed to criticize women in an online forum, but women can write/say the most horrible things about men without any consequences. (u/Uqtpa, 2018[3])*

Similar to r/TheRedPill, this theme of misandry and censorship ran throughout much of the responses to the quarantine. Others even pointed out that other manosphere spaces, such as r/MGTOW avoided the quarantine, suggesting a lack of consistency in Reddit's approach. Others however thought the quarantine was a good thing, as it allowed the community to continue on without the influence of

[3] r/Braincels was banned in October 2019. While it is still possible to collect data through Reddit's API and sites such as pushshift.io, submissions and comments are no longer available to view online.

so called “normies”. As u/WeAreLostSoAreYou argued: “why do you guys want to find a new sub? this is great. no more normies. i love it. lets stay quarantined. we have enough people to keep thriving with steady activity” (u/WeAreLostSoAreYou, 2018).

Here the quarantine is perceived as a positive development, and in fact, for some the implementation of the quarantine showed that the subreddit was on the right track. Some users argued that the quarantine would allow r/Braincels to stop being so “cucked” and in turn to return to its roots. As u/FireAlarm911 argued:

I'd rather remain quarantined than "change the direction of the sub". It was already getting too cucked in here, conforming to their wishes will kill this sub more than quarantine, I'm sure they consider pretty much anything critical of women as “mysoginist". (u/FireAlarm911, 2018)

This suggests a very strong unwillingness to change any behaviour in response to the quarantine. Finally, similar to much of what happens on r/Braincels, some users responded to the quarantine in meme format. u/David_Allen_Cope turned to Braveheart for their inspiration. They posted a comment stating :“They may quarantine our sub but they'll never quarantine our [FREEEEEEEEEDOM!] https://imgur.com/8zTqFvV.” (u/David_Allen_Cope, 2018) The link goes to a cartoon image of the famous scene in the movie Braveheart, in which Mel Gibson rallies his troops before battle. In the cartoon Gibson’s character is replaced by a drawing of the right-wing-meme “Pepe the Frog” - popular within manosphere forums - with the same battle paint on his face. This meme equates members of r/Braincels with William Wallace and the Scots fighting for their freedom in the First War of Scottish Independence.

Conclusion

This paper has studied the efficacy of the quarantine as a means of restricting and reducing hateful and misogynistic material on large digital platforms.

As already articulated, the goal of the quarantine from Reddit’s perspective was twofold: to cordon off affected subreddits from the broader platform, in turn reducing access to hateful material, and to change discourse within affected subreddits in a more positive direction. Based on the analysis presented, the results of the quarantine are mixed. The quarantine resulted in a drop in engagement within both affected subreddits by approximately one half. This suggests that Reddit was successful in reaching its first goal. Based on the linguistic analysis however, those who continued to participate did not reduce their levels of misogyny, highlighting a failure in reaching goal two. This failure could, in part, be explained by how users responded to the quarantine in the first place. In both instances there was a large amount of anger and concern at the implementation of the quarantine, with this anger in particular manifesting in a campaign in r/TheRedPill to encourage users to move to an external platform. Some users in r/Braincels even celebrated the quarantine as a sign that the subreddit was heading in the right direction, arguing they would prefer to stay quarantined so they did not have to change their language. While some users did suggest the quarantine highlighted a need for change, these voices were in the minority, as highlighted by the lack of change in language within both affected subreddits.

What does this mean for the use of technical and editorial means to restrict hateful content on digital platforms? Digital platforms such as Reddit increasingly sit in an uncomfortable position. Large platforms face increasing pressure to ban or heavily moderate hateful material, often finding a need to strike a balance between pleasing users and advertisers, adhering to regulations, and maintaining their role as spaces of free speech (Gillespie, 2018).

In managing these tensions, platforms act as silos. The focus of the platform-wide moderators of Reddit is on managing hate speech on Reddit and Reddit alone. There is little to no concern placed on the broader implications of moderation decisions on content on the web, overall. Acting as a silo therefore, Reddit can easily see the quarantine largely as a success. While the quarantine did not change the density of misogynistic material with either subreddit, it did reduce the levels of this material significantly in raw numbers. Reddit successfully isolated these communities from the rest of the platform, likely reducing the overall level of misogynistic material on the platform (although this platform-wide analysis has not been tested in this paper, meaning this is a speculative prediction).

This does not necessarily mean however that the quarantine reduced the levels of misogyny and other hateful material on the web overall. In particular, the quarantine resulted in a significant campaign from moderators of r/TheRedPill to push users to an external site, one managed entirely by Red Pill moderators. In this site users were, and remain, able to engage in misogynistic and other hateful behaviour with no external moderation at all. This site is restriction-less. While we do not have data as to how many users actually migrated to this platform, in the case of r/TheRedPill, the effect of the quarantine was potentially similar to previous bans of subreddits on the platform. In analysing these bans, Chandrasekharah et al. argued:

“In a sense, Reddit has made these users (from banned subreddits) *someone else’s problem*. To be clear, from a macro perspective, Reddit’s actions likely did not make the internet safer or less hateful. One possible interpretation, given the evidence at hand, is that the ban drove the users from these banned subreddits to darker corners of the internet” (Chandrasekharah et al., 2017, p. 18).

These moves occurred alongside a significant debate within right-wing and extremist organisations, with a range of high-profile individuals, in particular in the United States, arguing the right should create and migrate to their own platforms (Marantz, 2019). A number of alternative platforms have already been created. This includes Gab, a right-wing version of Twitter, and Voat, which is seen as a replacement for Reddit (Marantz, 2019). A significant part of these shifts has been an articulated distrust in large digital media platforms who have been seen as biased and discriminatory to right-wing voices. While the response from r/TheRedPill and r/Braincels differed, this distrust was a strong theme throughout. Users in particular articulated concerns that the quarantine specifically targeted men and was misandrist, or that it was targeting some subreddits unfairly when others went unscathed. The quarantine, therefore either likely created or furthered distrust from users *vis-à-vis* Reddit as a platform.

In turn, while the quarantine likely reduced hateful material on Reddit, it has potentially created a situation in which members of the manosphere create, at least in regards to their engagement around manosphere ideas, a mini filter bubble of content and social connections that connect only with their own pre-defined beliefs and opinions (Pariser, 2012). While recent empirical research has called into question the widespread existence of filter bubbles online (Zuiderveen Borgesius, Trilling, Möller, Bodó, de Vreese and Helberger, 2016; Bruns, 2019), some research has shown they may exist at the extremes of the ideological spectrum (Rietzschel, 2017). In other words users seeking out manosphere material may increasingly only engage with that content on self-moderated forums, while using other platforms for other social and political connections. If this has occurred this presents new challenges, in particular as newer forums have much less content moderation than large digital platforms. In these spaces therefore, this material is able to flourish unchecked from the oversight either of platform moderators or other users. More research is required to examine the impact recent shifts to self-moderated forums from manosphere and far-right communities have on the creation and spread of hateful material on the web.

Overall therefore, the results of the quarantine as a policy prescription are mixed. For Reddit itself the quarantine would likely be seen as a success. This has been shown by its increased use since the revamp in 2018, with administrators in particular targeting high-profile subreddits such as r/The_Donald and r/ChapoTraphouse (before both were banned in 2020). However, this does not necessarily mean that it changed the perspectives of users affected by the policy or that it reduced levels of hateful material on the web overall. This paper provides a first step of analysis of this policy approach. More work is needed to further this understanding of the role of approaches such as quarantines as ways to reduce hateful content on the internet.

References

BBC News. (2020, May 27). Twitter tags Trump tweet with fact-checking warning. *BBC News*. https://www.bbc.com/news/technology-52815552

Bruns, A. (2019). Filter bubble. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1426

Castillo, M. (2018, July 5). Reddit—One of the world's most popular websites—Is trying to cash in through advertising. *CNBC*. https://www.cnbc.com/2018/06/29/how-reddit-plans-to-make-money-through-advertising.html

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (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*. https://doi.org/10.1145/3134666

Copland, S. (2018, October 29). Reddit Quarantined. *Cyborgology*. https://thesocietypages.org/cyborgology/2018/10/29/reddit-quarantined/

Copland, S., & Davis, J. (2020, July 2). Reddit removes millions of pro-Trump posts. But advertisers, not values, rule the day. *The Conversation*. https://theconversation.com/reddit-removes-millions-of-pro-trump-posts-but-advertisers-not-values-rule-the-day-141703

Davis, J., & Chouinard, J. (2016). Theorizing affordances: From request to refuse. *Bulletin of Science, Technology & Society*, *36*(4), 241 – 248. https://doi.org/10.1177/0270467617714944

Facebook. (2020). *Facebook's third-party fact-checking program*. Facebook Journalism Project. https://www.facebook.com/journalismproject/programs/third-party-fact-checking

Farrell, T., Fernandez, M., Novotny, J., & Alani, H. (2019). Exploring misogyny across the manosphere. *WebSci '19 Proceedings of the 10th ACM Conference on Web Science*, 87–96. https://doi.org/10.1145/3292522.3326045

From 1 to 9,000 communities, now taking steps to grow reddit to 90,000 communities (and beyond! (2015). *Reddit*. https://www.reddit.com/r/announcements/comments/2x0g9v/from_1_to_9000_communities_now_taking_steps_to/

Gillespie, T. (2010). *The politics of 'platforms'*. *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Hern, A. (2018, August 6). Facebook, Apple, YouTube and Spotify ban Infowars' Alex Jones. *The Guardian*. https://www.theguardian.com/technology/2018/aug/06/apple-removes-podcasts-infowars-alex-jones

Marantz, A. (2019). *Antisocial: How online extremists broke America*. Picador.

Massanari, A. (2015). *Participatory culture, community, and play*. Peter Lang Publishing.

Massanari, A. (2017). Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media and Society*, *19*(3), 329–346. https://doi.org/10.1177/1461444815608807

Matias, J. N. (2019). The civic labour of volunteer moderators online. *Social Media + Society*, *5*(2). https://doi.org/10.1177/2056305119836778

Newell, E., Jurgens, D., Saleem, H. M., Vala, H., Sassine, J., Armstrong, C., & Ruths, D. (2016). User migration in online social networks: A case study on Reddit during a period of community unrest. *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, 279–288. http://cgi.cs.mcgill.ca/~enewel3/pdfs/user-migration-in-online-social-networks.pdf

Nicholas, L., & Agius, C. (2018). *The persistance of global masculinism: Discourse, gender and neo-colonial re-articulations of violence*. Palgrave MacMillan. https://doi.org/10.1007/978-3-319-68360-7

Ohanian, A. (2013). *Without their permission: How the 21st Century will be made, not managed*. Business Plus.

Pariser, E. (2012). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin Books.

Peart, M. (2019). *Reddit to cross $100 million in ad revenues in 2019*. EMarketer Newsroom. https://www.emarketer.com/newsroom/index.php/reddit-to-cross-100-million-in-ad-revenues-in-2019/

Prince, M. (2019, August 5). Terminating service for 8Chan [Blog post]. *Cloudflare*. https://blog.cloudflare.com/terminating-service-for-8chan/

Rietzschel, A. (2017, July 11). Wie es in Facebooks Echokammern aussieht – von links bis rechts [What it looks like in Facebook's echochambers – from left to right]. *Süddeutsche Zeitung*. https://www.sueddeutsche.de/politik/mein-facebook-dein-facebook-wie-es-in-den-echokammern-von-links-bis-rechts-aussieht-1.3576513

Robertson, A. (2015a, July 10). *Reddit bans "Fat People Hate" and other subreddits under new harassment rules*. https://www.theverge.com/2015/6/10/8761763/reddit-harassment-ban-fat-people-hate-subreddit

Robertson, A. (2015b, July 15). Was Reddit always about free speech? Yes, and no. *The Verge*. https://www.theverge.com/2015/7/15/8964995/reddit-free-speech-history

Salter, M. (2018). From geek masculinity to Gamergate: The technological rationality of online abuse. *Crime, Media, Culture*, *14*(2), 247–264. https://doi.org/10.1177/1741659017690893

Solon, O. (2017, November 9). "Incel": Reddit bans misogynist men's group blaming women for their celibacy. *The Guardian*. https://www.theguardian.com/technology/2017/nov/08/reddit-incel-involuntary-celibate-men-ban

u/5days, u/ekjp, & u/kn0thing. (2015). Promote ideas, protect people [Blog post]. *Reddit Blog*. https://redditblog.com/2015/05/14/promote-ideas-protect-people/

u/bnano999. (2019). *R/braincels just got banned* [Post]. Reddit. https://www.reddit.com/r/SubredditDrama/comments/dbfx4c/rbraincels_just_got_banned/

u/David_Allen_Cope. (2018). [User page]. Reddit.

u/FireAlarm911. (2018). [User page]. Reddit.

u/landoflobsters. (2017). *Update on site-wide rules regarding violent content* [Post]. Reddit. https://www.reddit.com/r/modnews/comments/78p7bz/update_on_sitewide_rules_regarding_violent_content/

u/landoflobsters. (2018). *Revamping the quarantine function* [Post]. Reddit. https://www.reddit.com/r/announcements/comments/9jf8nh/revamping_the_quarantine_function/

u/landoflobsters. (2019). *Changes to our policy against bullying and harassment* [Post]. Reddit. https://www.reddit.com/r/announcements/comments/dbf9nj/changes_to_our_policy_against_bullying_and/

u/reddit. (2012). *A necessary change in policy* [Post]. Reddit. https://www.reddit.com/r/blog/comments/pmj7f/a_necessary_change_in_policy/

u/redpillschool. (2018). *300,000 subscribers. The Reddit admin tacitly endorses male abuse and denies its victims* [Post]. Reddit. https://www.reddit.com/r/TheRedPill/comments/9jypns/300000_subscribers_the_reddit_administration/

u/redpillsschool. (2018). *Our appeal of the quarantine to Reddit admin* [Post]. Reddit. https://www.reddit.com/r/TheRedPill/comments/9lcyio/our_appeal_of_the_quarantine_to_the_reddit_admin/

u/Uqtpa. (2018). [User page]. Reddit.

u/WeAreLostSoAreYou. (2018). [User page]. Reddit.

Van Valkenburgh, S. P. (2019). "She thinks of him as a machine": On the entanglements of neoliberal ideology and misogynist cybercrime. *Social Media + Society*, *5*(3), 1–12. https://doi.org/10.1177/2056305119872953

Zuckerberg, D. (2018). *Not all dead white men: Classics and misogyny in the digital age*. Harvard University Press.

Zuiderveen Borgesius, F. J., Trilling, D., Möller, J., Bodó, B., Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review*, *5*(1). https://doi.org/10.14763/2016.1.401

Appendix one: Lexicons of misogyny

This is a description of each of the different subcategories of the misogynistic dictionary, as detailed by Farrell et al. (2019, p. 5).

- The concept of stoicism is based on Zuckerberg’s (2018) analysis of the manosphere. It encapsulates terms and expressions of endurance of pain or hardship because of the lack of intimacy of beauty. Terms like ‘kiss-less’, ‘hug-less’ or ‘involuntary celibate’ are part of this category.
- Patriarchy encapsulates that which has to do with women being

considered less than men, or some men being better than others by virtue of having traditional masculine qualities.

- Flipping the Narrative encapsulates terms and expressions that refer to men being oppressed by women or (indirectly or directly) by other men.
- Sexual Violence encapsulates any word explicitly connected with sexual violence (and nothing else).
- Physical Violence encapsulates any work explicitly connected with physical violence that is not explicitly sexual.
- Hostility includes violent verbs, and slurs that are not immediately racist or homophobic. However, if a verb is ambiguous (such as fucking), but it is made into a slur (such as fucker) it is coded as hostility.
- Belittling encapsulates any word that is disrespectful or degrading of women's experiences.
- Homophobia encapsulates any word related to being homosexual or that mocks being homosexual. This category does not distinguish between terms that have to do explicitly with women and those that have to do with men, for reasons that this was a modifying category to assess general violent attitudes.
- Racism referred to any word that was supposed to represent a specific group of people based on where they are from or their perceived race or ethnicity. Xenophobia and racism are not separated in this category. We included terms like 'fresh off the boat', 'paddy' (pejorative for Irish person) and 'kraut' (pejorative for German person), as well as terms referring specifically to race. The reason for this division is, similarly to the above, to simplify modifying categories about violent attitudes.

Published by

Alexander von Humboldt
INSTITUTE FOR INTERNET
AND SOCIETY

in cooperation with

CREATe

centre internet et société

R&I
IN3
Internet interdisciplinary Institute
Universitat Oberta de Catalunya