Does Content Moderation Lead Users Away from Fringe Movements? Evidence from a Recovery Community

*Giuseppe Russo[1], *Maciej Styzcen[2], Manoel Horta Ribeiro[3], Robert West[1]
[1]EPFL,
[2]Uber,
[3]Princeton University,
giuseppe.russo@epfl.ch, maciejs@uber.com, manoel@cs.princeton.edu, robert.west@eplf.ch

Abstract

Online platforms have sanctioned individuals and communities associated with 'fringe' movements linked to hate speech, violence, and terrorism—but can these sanctions contribute to the abandonment of these movements? Here, we investigate this question through the lens of r/exredpill, a recovery community on Reddit meant to help individuals leave movements within the Manosphere, a conglomerate of fringe Web-based movements focused on men's issues. We conduct an observational study on the impact of sanctioning some of Reddit's largest Manosphere communities on the activity levels and user influx of r/exredpill, the largest associated recovery subreddit. We find that banning a related radical community positively affects participation in r/exredpill in the period following the ban. Yet, *quarantining* the community, a softer moderation intervention, yields no such effects. We show that the effect induced by banning a radical community is stronger than for some of the widely discussed real-world events related to the Manosphere and that moderation actions against the Manosphere do not cause a spike in toxicity or malicious activity in r/exredpill. Overall, our findings suggest that content moderation acts as a deradicalization catalyst.

1 Introduction

Users and communities associated with 'fringe' movements like QAnon, Incel, or Proud Boys have been heavily sanctioned by mainstream social media platforms following their involvement with online harassment and real-world violence (BBC 2017; NBC 2020; CBS 2018). Sanctions applied range from banning community and users permanently from the platform—'hard' content moderation (Horta Ribeiro et al. 2021b)—to reducing the visibility or flagging violations of community guidelines without removing them entirely—'soft' content moderation (Zannettou 2021).

While soft and hard moderation efforts are generally applauded by organizations that combat online violence and extremism (Anti-Defamation League 2020; CCDH 2023), their effectiveness has been a subject of ongoing debate in academia (Zuckerman and Rajendra-Nicolucci 2021). On the one hand, moderation interventions have been shown to reduce the prevalence of hate speech and curtail activity in the targeted communities (Chandrasekharan et al. 2017, 2022). On the other hand, there are concerns about their unintended consequences; banned users often migrate to more radical, less regulated platforms, where their extremist views may intensify (Horta Ribeiro et al. 2021b) and spill over into mainstream platforms (Russo et al. 2023b,a). However, this debate lacks evidence about *recovery*. Do sanctions enacted by platforms lead users to increase their engagement with recovery communities? Does the extra work needed to find fringe content lead users to revisit their beliefs?

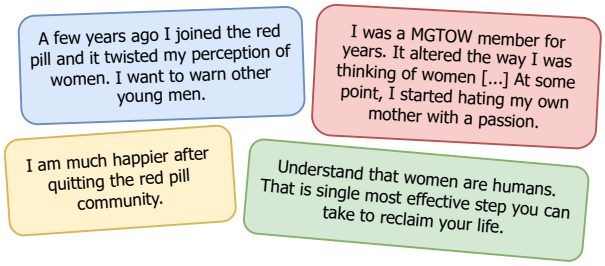

Figure 1: Some of the most upvoted comments and submissions in the `r/exredpill` recovery community (paraphrased due to privacy concerns).

Present work. In this paper, we address these very questions. We ask:

RQ1: Do *soft* moderation interventions (e.g., quarantining) lead to increased participation in recovery communities?

RQ2: Do *hard* moderation interventions (e.g., banning) lead to increased participation in recovery communities?

And finally, to compare the impact of real-world events on participation subsequent participation to the effects of moderation policies, we ask:

RQ3: Do real-world riots and terrorist attacks lead to the participation in recovery communities?

We answer these research questions through the lens of recovery communities, online groups meant to foster user engagement with recovery communities–online spaces where individuals seek support to distance themselves from fringe ideologies. While previous work has focused on aggregate trends of toxicity and participation in communities

*Equal contribution

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

linked with fringe movements (Trujillo and Cresci 2022a; Horta Ribeiro et al. 2021b; Chandrasekharan et al. 2022), we instead study how sanctions impacted participation in recovery communities, a metric that is more tightly linked with what may serve as an initial step toward disengagement from fringe ideologies (e.g., see Fig. 1 for example comments).

We present a case study of r/exredpill, a large recovery that caters to individuals seeking to distance themselves from ideologies associated with the 'The Manosphere'. The Manosphere is a conglomerate of anti-feminist movements (*e.g.*, Incels, Men Going Their Own Way, Men's Rights Activists), all of which had large communities on Reddit, a mainstream social media platform (Horta Ribeiro et al. 2021a; Basu 2020; Farrell et al. 2019). Manosphere-related communities on Reddit were repeatedly subjected to moderation interventions. Most notably, these communities have been *banned*, a 'hard' moderation measure that completely removes the community from Reddit, and *quarantined*, a 'soft' moderation measure that impedes direct access to and promotion of the community.

We study the effect of quarantining (**RQ1**) and banning (**RQ2**) three Manosphere communities (r/MGTOW, r/Braincells, r/redpill) on three key participation-related outcomes on *r/exredpill*: the overall activity within the support community, the influx of new participants, and the migration of users from fringe to recovery communities. Further, we study the impact of three real-world events (Unite the Right Rally, Toronto Van Attack, Capitol Hill Siege) on the same participation-related outcomes (**RQ3**). Our key analyses use two causal inference methods: interrupted time series (ITS) regression and Bayesian structural time series (BSTS) modeling. Last, we conduct extensive robustness checks to ensure that increases in participation in *r/exredpill* were not driven by negative comments or brigading users.

Results. We found little evidence that soft moderation interventions (quarantining) increased participation in online recovery communities(**RQ1**). In contrast, we found that hard moderation interventions (banning) led to substantial increases in activity, newcomer participation, and migration to the recovery community r/exredpill. Following the bans of r/Braincels and r/MGTOW, activity levels increased by 88.4% ($p = 0.003$) and 64.5% ($p = 0.001$), respectively. Newcomer participation rose by 174.3% ($p = 0.004$) and 31.6% ($p = 0.001$), while migration to recovery communities grew by 94.6% ($p = 0.001$) and 22.8% ($p = 22.8$) after these bans (**RQ2**). Surprisingly, while real-world events linked to these communities increased activity in the recovery community by up to 33% and newcomer participation by 16%, their impact was significantly smaller than that of platform-based moderation (**RQ3**).

Implications. Mainstream platforms have attempted to mitigate the influence of fringe communities through visibility reduction and bans. However, these interventions are not without limitations, as fringe communities show remarkable resilience (Horta Ribeiro et al. 2021b; Russo, Ribeiro, and West 2024). Our findings suggest that these moderation actions might also serve as catalysts for deradicalization, potentially guiding users toward recovery communities. Platforms could consider leveraging this effect by strategically nudging fringe community members toward supportive environments, especially when impending moderation actions are anticipated.

2 Background and Related Work

Online Antisocial Communities. Fringe movements hold beliefs considered extreme by society at large (Okholm, Fard, and Thij 2024), and online communities associated with said movements are known to frequently engage in antisocial behavior, propagate conspiracy theories, and promote extremist ideologies (Marwick and Caplan 2018; Russo[1], Stoehr[1], and Ribeiro 2024). Examples of these movements include QAnon (Schulze et al. 2022), the Alt-right (Rieger et al. 2021), and most relevant to the work at hand, the Manosphere, a conglomerate of anti-feminist movements characterized by their hostility towards women (Farrell et al. 2019). Manosphere-related communities prospered on Reddit in the 2010s, in 'subreddits' [discussion forums; see Horta Ribeiro et al. (2021a) for details]. Here, we focus on three Manosphere communities active on Reddit, each associated with a different movement within the Manosphere: r/Braincels, r/MGTOW, and r/TheRedPill. In the paragraphs below, we briefly describe the communities and their associated movements.

r/Braincels was a subreddit associated with the Involuntary Celibate movement. The online community became popular circa 2017 after another subreddit (r/Incels) was banned by Reddit for breaching community guidelines (Baele, Brace, and Ging 2024). The Incel movement abides by "The Black Pill," the idea that some men (Incels) are unable to have romantic and sexual relationships because of their physical appearance (Incel Wiki 2024). Incel communities are notorious for creating hateful and misogynistic content (Horta Ribeiro et al. 2021a). Further, Incels believe that systemic changes are needed to address men's dating issues and have supported and perpetrated acts of violence to achieve them (O'Donnell and Shor 2022).

r/MGTOW was a subreddit associated with the Men Going Their Own Way movement. It was created in 2011, and was among the most popular Manosphere-related subreddits (Horta Ribeiro et al. 2021a). The Men Going Their Own Way movement preaches that society is rigged against men (Lin 2017) and that the only solution is the abandonment of women and, sometimes, of western society in general (Lin 2017; Jones, Trott, and Wright 2020). Members of the movement openly disdain women, and MGTOW-adjacent online communities propagate and normalize misogynistic beliefs via online harassment (Jones, Trott, and Wright 2020)

r/TheRedPill is a subreddit associated with various movements within the Manosphere founded by a New Hampshire state legislator in 2012 (Vox 2017). The subreddit's name alludes to a famous scene from the movie "The Matrix" that, within the Manosphere, refers to the (internally widespread) belief that men, and not women, are disadvantaged in modern (feminist) society (Ging 2019). Much of r/TheRedPill content describes pseudo-scientific 'sexual strategy in a cul-

ture increasingly lacking a positive identity for men,' alluding to Manosphere movements like Pick Up Artists, Men's Rights Activists, and Men Going Their Own Way (Thorburn 2023b; Horta Ribeiro et al. 2021a). Perhaps unsurprisingly, r/TheRedPill was notorious as a hub for misogynistic content on Reddit (The Guardian 2015).

Impact of community-level sanctions. Incidents of online harassment, hate speech, and real-world violence led Reddit to sanction communities associated with fringe movements. Typically, Reddit has applied one of two sanctions: quarantining (soft moderation) and banning (hard moderation). Quarantined subreddits do not appear on user's feeds, are not included in search or recommendations, require users to be logged in to Reddit to view the community, and display a warning that requires users to explicitly opt-in to view the content (Reddit 2024). Banned subreddits are deleted from Reddit, and all their posts and comments become inaccessible. Notably, users participating in a banned subreddit keep their accounts. Prior research shows that quarantines reduce new user recruitment, though the effect is often modest (Chandrasekharan et al. 2017; Trujillo and Cresci 2022b), but do not significantly reduce existing users' toxicity (Chandrasekharan et al. 2022). Bans significantly reduce activity but can push users to other fringe platforms (Horta Ribeiro et al. 2021b), leading to spillover effects back onto mainstream platforms (Russo et al. 2023b; Schmitz, Muric, and Burghardt 2022; Russo Latona et al. 2024).

Recovery communities. Recovery communities on Reddit support individuals dealing with issues like addiction (Gauthier, Costello, and Wallace 2022; Balsamo et al. 2023; D'Agostino et al. 2017), mental health problems (De Choudhury and De 2014), eating disorders (Fettach and Benhiba 2019), and even political extremism and conspiracy beliefs (Harris 2023; Engel, Phadke, and Mitra 2023a). These communities provide networks to aid in recovery or deradicalization. This study focuses on r/exredpill, a key recovery community for members of the Manosphere, including subreddits like r/Braincels, r/MGTOW, and r/TheRedPill. While previous research highlights r/exredpill's potential for de-radicalization (Thorburn 2023a,b; Gheorghe and Yuzva Clement 2023), it has primarily relied on qualitative analysis, with no quantitative studies conducted yet.

Research on the recovery from extremist beliefs often relies on interview-based studies and theoretical models of disengagement. For instance, Xiao, Cheshire, and Bruckman (2021) conducted interviews with current and former chemtrail conspiracy believers, revealing that accidental exposure to counter-narratives (Engel, Phadke, and Mitra 2023b), persuasion by trusted peers, and a desire for social acceptance were key factors in abandoning such beliefs. Similarly, studies on QAnon communities (Jigsaw 2021; Phadke, Samory, and Mitra 2021) have used qualitative methods to highlight the role of disillusionment with failed predictions and unmet promises in prompting recovery.

Theoretical models further frame these processes. Cognitive Dissonance Theory (Harmon-Jones and Mills 2019) suggests that exposure to conflicting information or the need to express opposing views publicly can induce internal tension, leading to attitude shifts. Role Exit Theory (Ebaugh 1988) describes disengagement as a staged process involving 1) doubt, 2) exploration of alternatives, 3) a decisive turning point, and 4) the establishment of a new identity. Aho (1988) identifies various pathways toward radicalization and deradicalization, but most importantly to the work at hand, characterizes voluntary radicalization and deradicalization as a consequence of a change in the push and pull factors.

Present and prior work. Our hypotheses draw on deradicalization theories, such as role exit (Ebaugh 1988) and Aho's Defection Model (Aho 1988). Both theories frame deradicalization as a multi-step process, with a key step linked to external events that can exacerbate and accelerate preexisting doubts, ultimately leading to the abandonment of the community. We hypothesize that moderation policies, like quarantines and bans, may act as "turning points" by disrupting engagement, fostering cognitive dissonance, and prompting belief reassessment (Harmon-Jones and Harmon-Jones 2012; Wacquant 1990). Quarantines might expose users to alternative perspectives, enabling gradual disengagement (Xiao, Cheshire, and Bruckman 2021), whereas bans may trigger abrupt disruptions but risk reinforcing oppositional identities (Liguori 2021; Bérubé et al. 2019).

3 Materials and Methods

3.1 Data Collection

We used the Reddit Archive (the-eye.eu) to retrieve all comments and posts from three prominent fringe subreddits: r/Braincels, r/TheRedPill, and r/MGTOW. We considered all comments and posts from their creation until their eventual banning from Reddit. In the case of r/TheRedPill, which was only quarantined, we collected comments until 180 days after the quarantine event. Consistent with prior work, we considered only comments from users who contributed with more than five comments or posts to any of these subreddits (Kumar et al. 2018; Samory and Mitra 2018). Following Russo, Ribeiro, and West (2024), if a user exceeded the threshold across multiple subreddits, we categorized them under the subreddit with the highest activity to avoid duplicate classifications across communities. Our dataset comprises 9.8 million comments and 574,057 submissions from these communities, with further details presented in Table 1.

The subreddits studied were subject to various moderation actions, including quarantines and bans. Drawing on media reports (Pedroja 2021; Binder 2021) and documentation from the r/reclassified subreddit (which documented Reddit sanctions), we identified four key sanctions applied to r/Braincels, r/TheRedPill, and r/MGTOW:

1. On September 27th, 2018, r/Braincels and r/TheRedPill were quarantined after Reddit updated its policies on content moderation;
2. On October 1st, 2019, r/Braincels was banned;
3. On January 31st, 2020, r/MGTOW was quarantined;
4. On August 3rd, 2021, r/MGTOW was banned.

	Comments	Submissions	Users	Quarantine	Banning
Fringe subreddits					
r/Braincels	2,826,336	216,806	50,379	✓	✓
r/MGTOW	5,449,655	290,503	122,526	✓	✓
r/TheRedPill	3,206,546	118,396	122,684	✓	✗
Recovery subreddits					
r/exredpill	176,035	8,221	12,720	—	—

Table 1: For the subreddits considered in this paper, we depict the number of comments, submissions, and users obtained (columns 1–3) via the data collection of the entire posting (comments+submissions) history. Also, for each subreddit, we include information about the kind of moderation the community received (i.e., quarantine or banning; columns 4–5).

Finally, we collected all data of r/exredpill, a recovery community offering peer support to those disengaging from the ideologies promoted in r/Braincels, r/TheRedPill, and r/MGTOW. We gathered in total 8,221 submissions and 176,035 comments from 12,720 users made within a 120-day window before and after each identified moderation event. We defined membership in r/exredpill based on users who posted at least five times within the subreddit.

To accurately measure the effects of moderation events on the subsequent particpation in recovery community, we operationalize participation via three outcome variables:

1. **Activity Volume:** The daily number of comments and submissions in r/exredpill.
2. **Number of New Users:** The number of users posting in r/exredpill for the first time on a given day.
3. **Migrating Users:** The number of users posting in r/exredpill for the first time after previously contributing to one of the radical manosphere communities.

3.2 Estimating the Causal Effect

To estimate the causal effects of soft and hard moderation interventions (**RQ1** and **RQ2**), as well as of real-world events (**RQ3**), we use two causal inference methods: interrupted time series (ITS) regression and Bayesian structural time series (BSTS) modeling.

Interrupted time series analysis (ITS) is a widely used technique for detecting changes in trends, onset, and decay of effects from interventions by examining a series of observations before and after a defined intervention point (Bernal, Cummins, and Gasparrini 2017). The applicability of ITS depends on certain data assumptions. For example, when non-linear trends or specific distributions are present, more advanced regression techniques may be required (Wagner et al. 2002). ITS models must address issues like autocorrelation and seasonality, which can skew effect size estimates if not properly accounted for. In our study, we utilize the ITS regression model fo to illustrate changes in linear trends of key variables around the intervention points, as described in the linear model below:

$$Y_t = \beta_0 + \beta_1 T + \beta_2 D + \beta_3 P + \epsilon, \quad (1)$$

where, Y_t represents the outcome variable of the time series, T is a continuous variable indicating time in days from the start of the observational period, with β_1 capturing the pre-intervention trend. D is a binary variable indicating the presence (1) or absence (0) of the intervention, with β_2 representing the immediate effect of the intervention. P is a continuous variable indicating the number of days since the intervention, with β_3 capturing any post-intervention trend changes. Finally, ϵ represents the model's error term.

The ITS model was fit using Ordinary Least Squares (OLS), chosen for its simplicity and suitability for visualization rather than inferential purposes. While count outcome variables may often be skewed, we prioritized OLS to emphasize absolute changes, aligning with our goal of identifying and visualizing trends. Additionally, this approach served as a robustness check complementary to the Bayesian Structural Time Series (BSTS) model.

Bayesian Structural Time Series (BSTS) Modeling is a Bayesian statistical approach that offers several advantages over traditional ITS analysis. BSTS allows for the decomposition of a time series into components, combined with a dynamic regression framework that uses Monte Carlo Markov Chain (MCMC) simulations to generate counterfactual data and confidence intervals (Brodersen et al. 2015).

The method estimates a synthetic control via a state-space time-series model that uses information from 1) the time-series behavior of the outcomes of interest and 2) a set of multiple control time series similar to the target series. The synthetic control is made on the pre-treatment portion of potential controls, but its value lies in the post-treatment period. As long as the control series received no intervention, it is reasonable to assume the relationship between the treatment and the control series that existed before the intervention will continue afterward. Thus, a plausible estimate of the effect of the intervention can be computed.

To identify such control time series, we select subreddits that share demographic and political characteristics with r/Braincels, r/TheRedPill, and r/MGTOW. These control subreddits predominantly feature young male users with right-leaning views. To ensure comparability, we assessed subreddits across three social dimensions—partisanship, age, and gender—using cosine similarity to match them with the treatment subreddits (using social dimensions provided by (Waller and Anderson 2021). The final control

group includes 48 subreddits such as r/Conservative, r/cigars, r/GunPorn, and r/mancave.

BSTS is a more robust method for estimating intervention effects, particularly in the presence of autocorrelation and seasonality in the data. For our analysis, we utilize the BSTS implementation provided by the CausalImpact R package with MCMC 1000 iterations to ensure robust inference of the intervention effects.

4 Results

We examine the effect of quarantining, banning, and real-world events on participation-related outcomes associated with r/exredpill. We consider two quarantining events (the quarantining of r/Braincels and r/TheRedPill in 2018, and of r/MGTOW in 2020), two banning events (the banning of r/Braincels and r/MGTOW), and three real-world events (Unite the Right Rally, Toronto Van Attack, Capitol Hill Siege). The ITS analysis and the BSTS modeling results are summarized in Tables 4 to 6 (at the end of this document). We show the time series of the key outcomes and the estimated regression lines in Figures 2 and 3.

4.1 RQ1: Effect of quarantining

Quarantining and Activity Volume. Using Interrupted Time Series (ITS) analysis, we found no significant immediate change after the event (β_2 in the model), but we did observe varying impacts on activity trends following these events. The trend in activity in r/exredpill increased following the quarantining of r/Braincels and r/TheRedPill ($\beta_3^{BI+TRP} = 0.16; p = 0.002$), suggesting these events may have sparked discussion or attracted new users. Yet, we find no significant increase in activity following the quarantining of r/MGTOW. Since ITS lacks a control group, platform-wide trends could have influenced the results. To account for this, we applied BSTS modeling. This subsequent analysis casts doubt on the validity of the ITS results, as we find non-significant increases following the quarantining of r/Braincels and r/TheRedPill (15.4%; $p = 0.345$) and of r/exredpill (38.5%; $p = 0.09$). In this context, we conclude that quarantining fringe subreddits did not significantly impact activity volume in r/exredpill.

Quarantining and Number of New Users. Next, we examine whether quarantining influences the influx of new users to r/exredpill. In both the ITS and the BSTS analyses, we find no statistically significant effect of the moderation intervention on the number of newcomers, suggesting that quarantine did not help popularize r/exredpill.

Quarantining and Migrating Users. Finally, we explore whether quarantining fringe communities spurred migration to r/exredpill from r/Braincels, r/TheRedPill, and r/MGTOW. In the ITS analysis, we find no significant effect on the number of users that migrated from r/Braincels and r/TheRedPill to r/exredpill immediately after the quarantine ($\beta_2 = -0.01; p = 0.318$). Differently, we observe that quarantining led to a significant increase in the trend of migrants that previously participated in r/Braincels or r/TheRedPill ($\beta_3^{BI+TRP} = 0.044; p = 0.02$), suggesting that quarantining these subreddits led to an uptick in user migration to the recovery community in the period following up the quarantine. These results are confirmed by the BST analysis, which shows a 24.6% ($p = 0.045$) increase in migration from r/Braincels and r/TheRedPill to r/exredpill. In contrast, we found no statistically significant changes in the number of migrants from r/MGTOW to r/exredpill in the aftermath of the r/MGTOW quarantine (see Table 6). Altogether, these different results support the notion that quarantining increases the number of migrating users. However, we argue they do not provide substantial evidence that soft moderation interventions increase participation in recovery communities. Given that we are considering three metrics and two different events, this is likely a spurious finding. For instance, a conservative Bonferroni correction would set the significance threshold at 0.0083 (0.05÷6), rendering the effects observed here not statistically significant.

4.2 RQ2: Effect of banning

Banning and Activity Volume. Upon the banning of both r/Braincels and r/MGTOW, we observe a statistically significant increase in the activity volume in both the ITS analysis ($\beta_2^{BI} = 24.06; p < 0.001$, $\beta_2^{MT} = 86.49; p < 0.001$) and in the BSTS model (88.4% and 64.5% increase; $p = 0.032$ and $p = 0.001$). After the banning of r/Braincels, the activity volume of r/exredpill remains similar to the activity level at the time of the banning ($t = 0$). The coefficient β_3 that captures the trend growth after the banning ($t > 0$) is positive (indicating a slightly increasing trend) but not statistically significant. After the banning of r/MGTOW, differently from the r/Braincels ban, we observe a statistically significant decreasing trend ($\beta_3 = -0.61; p = 0.001$) in activity volume. However, even if the trend decreases, the activity volume remains consistently higher than the pre-ban activity volume. These results indicate that the banning of Manosphere communities increased activity in r/exredpill.

Banning and Newcomers. Considering the influx of newcomers, we observe that the banning events influenced the number of users joining r/exredpill. The number of new users who posted in the recovery community rose significantly following the bans of r/Braincels and r/MGTOW ($\beta_2^{BI} = 5.30; p = 0.001, \beta_2^{MT} = 11.74; p < 0.000$). The number of newcomers in the days following the ban exhibits an increasing statistically significant linear trend for r/Braincels ($\beta_2^{BI} = 0.04; p = 0.04$) and a not statistically significant decreasing trend for r/MGTOW. The BSTS analysis further supports the ITS, showing an increase in the number of newcomers following the r/Braincel and r/MGTOW ban of 174.3% ($p = 0.004$) and 31.7% ($p = 0.001$), respectively.

Banning and Migrating Users. Finally, we examine the migration of users from the banned fringe communities to r/exredpill. Following the bans of r/Braincels and r/MGTOW, we observed an uptick in the number of users from these communities posting in the recovery subreddit. The ITS analysis highlights a positive and significant increase in migration to r/exredpill after the r/Braincels and ($\beta_2^{BI} = 1.24; p = 0.03$) and r/MGTOW banning ($\beta_2^{MT} = 21.10; p < 0.001$). Similarly to what we observed

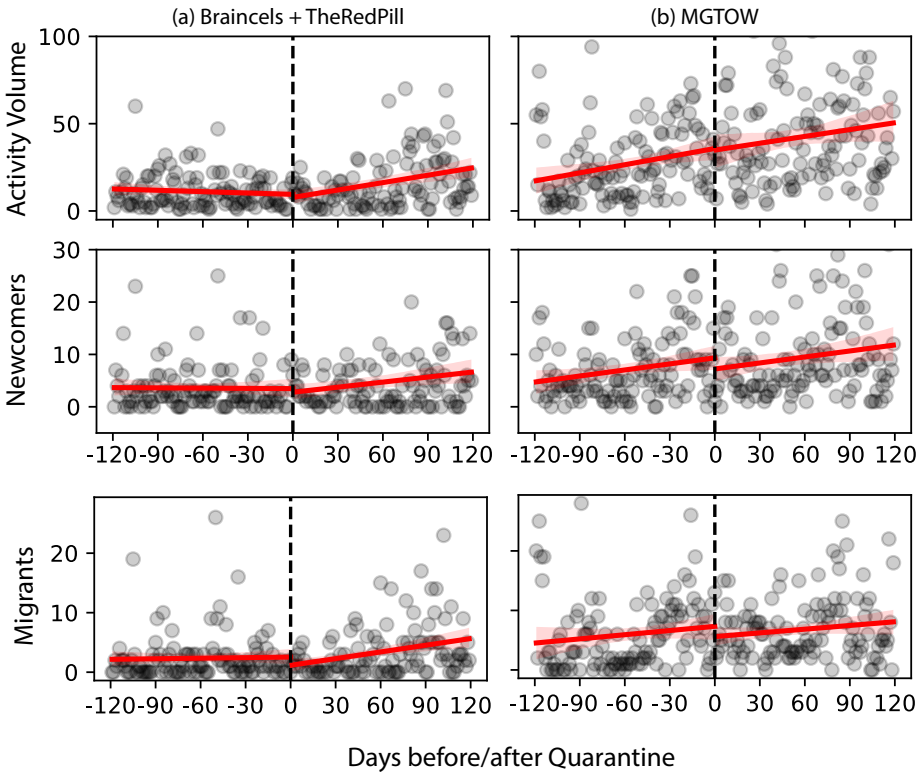

Figure 2: Effects of Quarantine on r/exredpill Activity Volume, Newcomers, and Migrants. We show the results obtained after fitting the ITS model on the activity volume, newcomers, and migrants that joined r/exredpill in the aftermath of the quarantine. In the left column (a), we show the ITS analysis assessing the effect on activity volume and newcomers after the quarantine of r/Braincels and r/TheRedPill. In right column (b), we show the same analysis after the quarantine of r/MGTOW. In the bottom row, we show the ITS analysis for the users who joined r/exredpill after having participated in r/Braincels (a), r/TheRedPill (b), and r/MGTOW (c).

in the case of Activity Volume and Newcomers, we observe one slightly decreasing linear trend after the Braincels ban ($\beta_3^{BI} = -0.01; p = 0.318$), and a increasing linear trend after the ban of MGTOW ($\beta_3^{MT} = 0.04; p = 0.441$). However, neither change in trend is statistically significant. The BSTS analysis estimates a 94,6% ($p = 0.001$) increase in migrating users from r/Braincels corresponding to an absolute increase of 46 users, and a 22.8% increase from r/MGTOW ($p = 0.006$), 21 additional users. These findings suggest that banning fringe communities not only curtails their activity but also prompts a subset of users to seek out recovery and support.

4.3 RQ3: Comparison with Real-World Events

To assess the impact of real-world events connected to the Manosphere on participation in recovery communities, we analyzed three significant events: the Unite the Right rally (UR) on August 11, 2017, the Toronto Van Attack (TA) in April 23, 2018, and the Capitol Hill Siege (CH) in January 6, 2021. These events were chosen for their ideological ties to the Manosphere. A self-identified Incel perpetrated the Toronto Van Attack (Guardian 2019), and both the Unite the Right Rally and the Capitol Hill Siege were associated with far-right groups tightly associated with the Manosphere (Mamié, Horta Ribeiro, and West 2021).

We first examined their effect on activity volume within r/exredpill. The BSTS analysis revealed statistically significant increases in activity following all three events, with activity volume rising by 11.2% ($p = 0.012$), 22.1% ($p = 0.06$), and 33.5% ($p = 0.043$) for the Unite the Right rally, the Toronto Van Attack, and the Capitol Hill Siege, respectively. The Interrupted Time Series (ITS) analysis similarly detected increases in activity immediately following these events ($\beta_2^{UR} = 2.82; p = 0.529$, $\beta_2^{TA} = 4.20; p = 0.018$, $\beta_2^{CH} = 11.65; p = 0.006$). However, no statistically significant changes in trends were observed after the events. Notably, compared to the effects of bans on r/MGTOW (+64.5%) and r/Braincels (+88.4%), the activity increases following real-world events were much smaller. Next, we analyzed the influx of new users into r/exredpill. The BSTS analysis found a statistically significant increase (+19%, $p = 0.012$) in new users following the Capitol Hill Siege. Still, no significant effects on newcomer influx were identified following the Unite the Right rally or the Toronto Van Attack. In contrast, the ITS analysis did not reveal statistically significant changes in newcomers immediately following (β_2) or after (β_3) any of the events. Finally, considering the migration of users from fringe communities to r/exred-

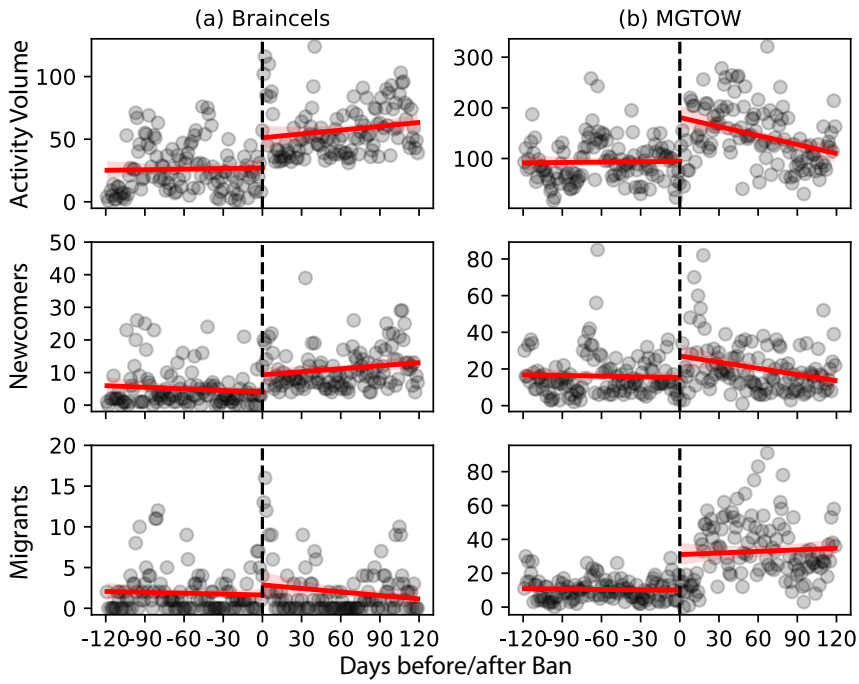

Figure 3: Effects of Banning on r/exredpill Activity Volume, Newcomers, and Migrants. We show the results obtained after fitting the ITS model on the activity volume, newcomers, and migrants that joined r/exredpill in the aftermath of the banning. In the left column (a), we show the ITS analysis assessing the effect on activity volume and newcomers after the ban of r/Braincels. In right column (b), we show the same analysis after the banning of r/MGTOW.

Event	p-value	relative effect	95% CI
Braincels ban	0.105	82%	[-13%, 250%]
MGTOW ban	**0.001**	-55%	[-63%, -44%]
Braincels & TRP quarant.	0.279	-0.57%	[-46%, 155%]
MGTOW quarantine	0.362	9%	[-20%, 56%]

Table 2: Effect of content moderation events against 'Manosphere' communities on the fraction of comments with a deleted body in `r/exredpill`.

pill, neither the ITS nor the BSTS analyses found any statistically significant changes in migration patterns following these events. Altogether, these results provide some evidence that real-world riots and terrorist attacks boost participation to recovery communities. Yet, most important for the work at hand, they highlight the magnitude of the effect sizes observed as a result of hard content moderation interventions.

5 Robustness Checks

A critical aspect of our analysis is ensuring that the observed increase in activity within r/exredpill following the moderation of fringe communities is not driven by negative motivations, such as brigading from users of the sanctioned communities. To address this concern, we conducted a series of robustness checks focusing on the content of the comments posted in r/exredpill after these moderation events.

Deleted content. We examined the fraction of deleted comments and submissions in r/exredpill before and after the moderation events. If the increased participation in r/exredpill was driven by former users brigading against r/exredpill, we would expect to see a rise in deleted content following the interventions. Yet, our analysis found no significant changes in the deleted content fraction in three of the four events studied. The only exception was a decrease observed following the ban of r/MGTOW, which we attribute to an upward trend in deleted comments during the pre-intervention period rather than the intervention itself. Overall, these findings suggest that the increase in activity within r/exredpill did not lead to a rise in inappropriate or rule-violating content. We provide the results of this analysis in Table 2.

Toxicity Analysis. We investigated whether the increased activity in r/exredpill could be attributed to an influx of toxic behavior from users of the moderated communities. To do this, we analyzed the toxicity levels of comments posted before and after each moderation event. We used Google's Perspective API (Jigsaw 2022) to annotate all comments with a toxicity score and applied Interrupted Time Series (ITS)

Event	Outcome variable	p-value	relative effect	95% CI
Unite the Right Rally (UR)	activity volume	**0.012**	**11.2%**	**[6.1%, 16.3%]**
	new users	0.645	0.8%	[-2.3%, 3.9%]
	migrating users	0.732	-1.1%	[-3.5%, 1.3%]
Toronto Van Attack (TA)	activity volume	0.06	22.1%	[15.1%, 29.1%]
	new users	0.559	-0.6%	[-2.7%, 1.5%]
	migrating users	0.687	1.2%	[-1.8%, 4.2%]
Capitol Hill Siege (CH)	activity volume	**0.043**	**33.5%**	**[26.3%, 40.7%]**
	new users	**0.012**	**19.0%**	**[12.3%, 25.7%]**
	migrating users	0.831	-0.3%	[-2.6%, 2.0%]

Table 3: BSTS results of the effects of real-world events on volume activity, newcomers, and migrating users within r/exredpill.

and Bayesian Structural Time Series (BSTS) models to detect significant changes. Our analysis found no statistically significant increases in average toxicity levels of r/exredpill around the time of the moderation events. Specifically, the ITS models showed no immediate change in toxicity at the time of intervention or in the follow-up period. Similarly, the BSTS models did not indicate any significant shifts in toxicity levels. These results held for all the moderation events considered.

LLM Moderation Analysis. While toxicity offers a useful indicator for assessing whether increased participation was driven by retaliation from moderated communities, it remains a controversial measure for content moderation (Friedl et al. 2023; Gargee et al. 2023). Inspired by previous works that used Large Language Models (LLM) to annotate data (Latona et al. 2024; Davidson et al. 2024), we used GPT-4-turbo, a large language model, to further evaluate the nature of comments posted by newcomers and migrants in r/exredpill. We applied two strategies. First, we provided GPT-4-turbo with the community guidelines of r/exredpill and a comment posted on r/exredpill from migrating users, asking if it violated any rules. Second, we described the ideologies of the three fringe communities (r/Braincels, r/MGTOW, and r/TheRedPill) and asked if the comment aligned with those ideologies. GPT-4-turbo labeled 97% of the comments as compliant with the community guidelines, and 99% did not align with the values of the fringe communities. This analysis reinforces that the post-moderation activity in r/exredpill reflects recovery rather than a continuation of fringe ideologies. We provide the prompts used in the Appendix.

Observation Window. To ensure the robustness of our findings, we tested various observation window lengths, including 60, 90, 150, and 180 days. Our sensitivity analysis showed that the results remained consistent across these different windows. This consistency suggests that our findings are not influenced by the choice of observation window. We selected the 120-day window as it provides a practical balance between capturing relevant trends and maintaining temporal proximity to the events studied.

Placebo Testing. To evaluate the reliability of our results, we conducted placebo tests using Bayesian Structural Time Series (BSTS) modeling. Specifically, we introduced "placebo intervention" dates at -120 days and +120 days relative to each studied intervention. We report these results in Tables 7 and 8 These placebo tests were applied to all quarantine and banning events where our main analysis detected changes. Additionally, we repeated this process using alternative observation windows of 60, 90, 150, and 180 days. Across all scenarios, the results consistently showed no effects for the placebo intervention dates.

Manual Inspection. To further understand the content of the posts published by either newcomers or migrants, we go beyond the automatic content analysis and perform a human judgment analysis. We selected 200 random comments made by users who are either newcomers or migrants from the r/exredpill recovery community from one of the r/Braincels, r/MGTOW, and r/TheRedPill communities after the moderation action had been taken. Two human annotators, both authors of this paper, labeled these sentences by marking whether they contain pro-Manosphere content or attacks against the r/exredpill community. In addition, we report interannotator rates with Cohen's κ. In 94% of the comments, no pro-Manosphere or community attack was identified. Instead, many comments reflected a change in views from those previously held by users in the fringe communities.

6 Discussion

Our study investigates the effects of moderation policies on fringe communities and their potential to influence participation in recovery communities. We shed light on how these interventions shape user behavior and recovery processes by examining soft (quarantines) and hard moderation (bans) interventions alongside real-world events associated with fringe ideologies.

Our key findings are threefold. First, we find that, contrary to our initial hypothesis (**RQ1**), quarantines of fringe communities had no substantial impact on recovery community participation. Activity volume and newcomer influx remained essentially unchanged following quarantines, suggesting that visibility reduction alone may not be sufficient to drive users toward recovery. Second, we find banning fringe communities led to a marked increase in participa-

tion across all considered outcomes. These results suggest that hard moderation can act as a turning point, encouraging former members of fringe communities to seek support and begin the process of deradicalization (**RQ2**). Third, our analysis suggests real-world events boosted participation in the recovery community, but the effects observed were smaller than those observed following bans (**RQ3**).

The robustness checks conducted across toxicity analysis, LLM moderation, manual inspection, control subreddit comparison, and deleted content analysis consistently support our conclusion: (hard) moderation interventions targeting fringe communities on Reddit led to increased participation in recovery communities like r/exredpill. This increased participation does not appear to be driven by negative or toxic behavior butreflective of genuine engagement with recovery processes.

Relation to existing social science theories. We discuss our findings in light of two prominent social science theories: the Role Exit Theory (Ebaugh 1988) and Aho (1988)'s Defection Model. Ebaugh (1988) theorizes the presence of turning points, events that lead to someone exiting a role. Our results indicate that bans, but not quarantines, may be understood as "turning points." Also, in light of the social exit theory, we argue that recovery communities may help users redefine their identity (or, in the lingo of the theory, creating the "ex-role"). Aho (1988) theorizes deradicalization as a consequence of a change in the push and pull factors. For example, relationships with other people in a radical group could "pull" individuals toward the hate group, whereas relationships with minorities targeted by the group could "push" them away from it. In that context, this study analyzes the force of banning and quarantining as "push factors," finding that banning seems enough to 'flip the scale' for many individuals, whereas quarantining is not.

But why are community-wide bans impactful, whereas quarantines are not? We hypothesize that bans may sever social ties within the community (Horta Ribeiro et al. 2021b), increasing the likelihood of users encountering alternative beliefs and counter-narratives within recovery communities. On the other hand, quarantines may merely isolate the community (Chandrasekharan et al. 2022), limiting opportunities for self-reflection or exposure to counter-narratives.

Implications. The results of this study highlight the potential for platform-based moderation to facilitate positive outcomes beyond merely reducing harmful activity (Chandrasekharan et al. 2017, 2022). Specifically, banning fringe communities appears to have the unintended but beneficial effect of driving a small but meaningful proportion of users to recovery communities, potentially initiating their journey toward deradicalization. Notably, we find that 4.7% of all users with at least five posts in one of the moderated communities posted in the recovery community. Additionally, 78.3% of the users who posted in the recovery community during this period continued to engage actively, averaging 12.7 posts over the following 120 days. This highlights the benefits of interventions that provide easier access to recovery communities or exposure to counter-narratives tailored to specific cohorts of users. In light of this, we argue that platforms should consider how their moderation strategies, particularly bans, can be refined to guide users away from harmful ideologies and toward supportive environments. As platforms navigate the challenge of balancing free speech with user safety, these findings suggest that combining hard moderation strategies with targeted support mechanisms may offer a more comprehensive and effective approach to fostering recovery.

Broader Impact. While our findings suggest that platform moderation, particularly banning, may support deradicalization efforts, they also raise ethical questions about the potential consequences of deplatforming. Restricting users' ability to engage with certain content may lead to migration toward more radical and unregulated spaces (Horta Ribeiro et al. 2021b), where extremism may further intensify. Platforms must, therefore, balance the benefits of sanctions with the risk of pushing users to more harmful environments.

Limitations and Future Work. While we took steps to mitigate potential confounders, such as analyzing real-world events, unobserved factors may still influence our results. Future studies could address these issues by incorporating more detailed user activity data, including passive engagement, and exploring the effects of moderation across platforms with different community structures. An important avenue for future research is to assess the long-term efficacy of recovery community participation after moderation actions. It remains unclear whether users who join recovery communities remain active, undergo genuine deradicalization, or eventually regress to their previous beliefs. Longitudinal studies focusing on user retention and shifts in ideological content would provide valuable insights into the sustainability of the recovery process. Last but not least, future work could focus on platforms different from Reddit, or fringe communities other than those within the Manosphere.

Last, we stress that, while we do find evidence that content moderation interventions may act as catalysts for deradicalization, our estimates represent a lower bound for three key reasons. First, our study captures only active contributors to r/exredpill, not passive participants (lurkers), meaning that even more users may have moved away from fringe communities. Second, we only consider users who engaged with recovery communities, which is likely only a fraction of users moving away from fringe communities. Third, individuals redefining their identities may adopt new usernames, which our methodology cannot track. We argue that these limitations do not decrease the importance of our findings, as a lower bound can still help us understand the consequences of content moderation interventions.

References

Aho, J. A. 1988. Out of Hate: A Sociology of Defection from Neo-Nazism. *Current Research on Peace and Violence*, 11(4).

Anti-Defamation League. 2020. ADL Statement on Facebook's Decision to Finally Ban QAnon Content From Platform. https://www.adl.org/news/press-releases/adl-statement-on-facebooks-decision-to-finally-ban-qanon-content-from-platform.

Baele, S.; Brace, L.; and Ging, D. 2024. A diachronic cross-platforms analysis of violent extremist language in the incel online ecosystem. *Terrorism and political violence*, 36(3): 382–405.

Balsamo, D.; Bajardi, P.; Morales, G. D. F.; Monti, C.; and Schifanella, R. 2023. The Pursuit of Peer Support for Opioid Use Recovery on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 12–23.

Basu, T. 2020. The "Manosphere" is getting more toxic as angry men join the incels.

BBC. 2017. Reddit Bans' Involuntarily Celibate' Community. https://www.bbc.com/news/blogs-trending-41926687.

Bernal, J. L.; Cummins, S.; and Gasparrini, A. 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International journal of epidemiology*, 46(1): 348–355.

Bérubé, M.; Scrivens, R.; Venkatesh, V.; and Gaudette, T. 2019. Converging Patterns in Pathways in and out of Violent Extremism. *Perspectives on Terrorism*, 13(6): 73–89.

Binder, M. 2021. Reddit changes its harassment policy and bans major incel community.

Brodersen, K. H.; Gallusser, F.; Koehler, J.; Remy, N.; and Scott, S. L. 2015. Inferring causal impact using Bayesian structural time-series models.

CBS. 2018. Twitter suspends Proud Boys, Gavin McInnes accounts ahead of Unite the Right rally. https://www.cbsnews.com/news/proud-boys-gavin-mcinnes-twitter-suspension-today-unite-the-right-2018-08-10/.

CCDH. 2023. Public Support for Social Media Reform. https://counterhate.com/research/public-support-for-social-media-reform-star/.

Chandrasekharan, E.; Jhaver, S.; Bruckman, A.; and Gilbert, E. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(4): 1–26.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW): 1–22.

Davidson, T. R.; Surkov, V.; Veselovsky, V.; Russo, G.; West, R.; and Gulcehre, C. 2024. Self-recognition in language models. *arXiv preprint arXiv:2407.06946*.

De Choudhury, M.; and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 71–80.

D'Agostino, A. R.; Optican, A. R.; Sowles, S. J.; Krauss, M. J.; Lee, K. E.; and Cavazos-Rehg, P. A. 2017. Social networking online to recover from opioid use disorder: A study of community interactions. *Drug and alcohol dependence*, 181: 5–10.

Ebaugh, H. R. F. 1988. *Becoming an ex: The process of role exit*. University of Chicago Press.

Engel, K.; Phadke, S.; and Mitra, T. 2023a. Learning from the Ex-Believers: Individuals' Journeys In and Out of Conspiracy Theories Online. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–37.

Engel, K.; Phadke, S.; and Mitra, T. 2023b. Learning from the Ex-Believers: Individuals' Journeys In and Out of Conspiracy Theories Online. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–37.

Farrell, T.; Fernandez, M.; Novotny, J.; and Alani, H. 2019. Exploring Misogyny across the Manosphere in Reddit. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, 87–96. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362023.

Fettach, Y.; and Benhiba, L. 2019. Pro-eating disorders and pro-recovery communities on Reddit: text and network comparative analyses. In *Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services*, 277–286.

Friedl, P.; et al. 2023. Toxic Bias: Perspective API Misreads German as More Toxic. *arXiv preprint arXiv:2312.12651*.

Gargee, S.; Gopinath, P.; Kancharla, S.; Anand, C.; and Babu, A. 2023. Analyzing and Addressing the Difference in Toxicity Prediction Between Different Comments with Same Semantic Meaning in Google's Perspective API. In *Proceedings of the International Conference on Computing and Network Communications (CoCoNet)*.

Gauthier, R. P.; Costello, M. J.; and Wallace, J. R. 2022. "I Will Not Drink With You Today": A Topic-Guided Thematic Analysis of Addiction Recovery on Reddit. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–17.

Gheorghe, R. M.; and Yuzva Clement, D. 2023. 'It's time to put the copes down and get to work': a qualitative study of incel exit strategies on r/IncelExit. *Behavioral Sciences of Terrorism and Political Aggression*, 1–21.

Ging, D. 2019. Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and masculinities*, 22(4): 638–657.

Guardian, T. 2019. Toronto van attack suspect says he was 'radicalized' online by 'incels'. https://www.theguardian.com/world/2019/sep/27/alek-minassian-toronto-van-attack-interview-incels. Accessed: 2024-08-01.

Harmon-Jones, E.; and Harmon-Jones, C. 2012. Cognitive dissonance theory. *Handbook of motivation science*, 71.

Harmon-Jones, E.; and Mills, J. 2019. An introduction to cognitive dissonance theory and an overview of current perspectives on the theory.

Harris, J. 2023. *The QAnon Infection: How Families Manage, Adapt to, and Abandon their QAnon-Infected Family Members*. Ph.D. thesis, Idaho State University.

Horta Ribeiro, M.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; Long, S.; Greenberg, S.; and Zannettou, S. 2021a. The evolution of the manosphere across the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 196–207.

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

Incel Wiki. 2024. Blackpill. https://incels.wiki/w/Blackpill. Accessed on 2024-06-15.

Jigsaw. 2021. Conspiracy Theories. https://jigsaw.google.com/the-current/conspiracy-theories/#new-perspectives. The Current 005.

Jigsaw. 2022. Perspective API. https://perspectiveapi.com/.

Jones, C.; Trott, V.; and Wright, S. 2020. Sluts and soyboys: MGTOW and the production of misogynistic online harassment. *New media & society*, 22(10): 1903–1921.

Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; and Tamura, K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*, 35(6): 1547.

Latona, G. R.; Ribeiro, M. H.; Davidson, T. R.; Veselovsky, V.; and West, R. 2024. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *arXiv preprint arXiv:2405.02150*.

Liguori, J. B. 2021. *Recovering from Racism: Why and How White Supremacists Quit Hate*. Master's thesis, Arizona State University.

Lin, J. L. 2017. *Antifeminism Online: MGTOW (Men Going Their Own Way)*, 77–96. Bielefeld: transcript Verlag. ISBN 9783839434970.

Mamié, R.; Horta Ribeiro, M.; and West, R. 2021. Are anti-feminist communities gateways to the far right? Evidence from Reddit and YouTube. In *Proceedings of the 13th ACM Web Science Conference 2021*, 139–147.

Marwick, A. E.; and Caplan, R. 2018. Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Studies*.

NBC. 2020. Facebook bans QAnon across its platforms. https://www.nbcnews.com/tech/tech-news/facebook-bans-qanon-across-its-platforms-n1242339.

O'Donnell, C.; and Shor, E. 2022. 'This is a political movement, friend': Why 'incels' support violence. *The British Journal of Sociology*, 73(2): 336–351.

Okholm, J. C. S.; Fard, A. E.; and Thij, M. t. 2024. Debunking and exposing misinformation among fringe communities: Testing source exposure and debunking anti-Ukrainian misinformation among German fringe communities. *Harvard Kennedy School Misinformation Review*.

Pedroja, C. 2021. Reddit bans "men going their own way" forums for violating hate speech rule.

Phadke, S.; Samory, M.; and Mitra, T. 2021. Characterizing Social Imaginaries and Self-Disclosures of Dissonance in Online Conspiracy Discussion Communities. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–35.

Reddit. 2024. Quarantined Communities. https://support.reddithelp.com/hc/en-us/articles/360043069012-Quarantined-Communities. Accessed on 2024-08-01.

Rieger, D.; Kümpel, A. S.; Wich, M.; Kiening, T.; and Groh, G. 2021. Assessing the extent and types of hate speech in fringe communities: a case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media+ Society*, 7(4): 20563051211052906.

Russo, G.; Horta Ribeiro, M.; Casiraghi, G.; and Verginer, L. 2023a. Understanding online migration decisions following the banning of radical communities. In *Proceedings of the 15th ACM Web Science Conference 2023*, 251–259.

Russo, G.; Ribeiro, M. H.; and West, R. 2024. Stranger Danger! Cross-Community Interactions with Fringe Users Increase the Growth of Fringe Communities on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1342–1353.

Russo, G.; Verginer, L.; Ribeiro, M. H.; and Casiraghi, G. 2023b. Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 742–753.

Russo Latona, G.; Gote, C.; Zingg, C.; Casiraghi, G.; Verginer, L.; and Schweitzer, F. 2024. Shock! Quantifying the Impact of Core Developers' Dropout on the Productivity of OSS Projects. In *Companion Proceedings of the ACM Web Conference 2024*, 706–709.

Russo[1], G.; Stoehr[1], N.; and Ribeiro, M. H. 2024. Automatic Conspiracy Theory Identification Task Overview. In *EVALITA Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop: Parma, Italy, September 7-8th, 2023*, 252. Accademia University Press.

Samory, M.; and Mitra, T. 2018. Conspiracies online: User discussions in a conspiracy community following dramatic events. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Schmitz, M.; Muric, G.; and Burghardt, K. 2022. Quantifying how hateful communities radicalize online users. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 139–146. IEEE.

Schulze, H.; Hohner, J.; Greipl, S.; Girgnhuber, M.; Desta, I.; and Rieger, D. 2022. Far-right conspiracy groups on fringe platforms: a longitudinal analysis of radicalization dynamics on Telegram. *Convergence*.

The Guardian. 2015. Can data analysis reveal the most bigoted corners of Reddit? https://www.theguardian.com/technology/2015/mar/23/can-data-analysis-reveal-the-most-bigoted-corners-of-reddit. Accessed on 2024-06-15.

Thorburn, J. 2023a. The (de-) radical (-ising) potential of r/IncelExit and r/ExRedPill. *European Journal of Cultural Studies*, 26(3): 464–471.

Thorburn, J. 2023b. Exiting the Manosphere. A Gendered Analysis of Radicalization, Diversion and Deradicalization Narratives from r/IncelExit and r/ExRedPill. *Studies in Conflict & Terrorism*, 1–25.

Trujillo, A.; and Cresci, S. 2022a. Make reddit great again: assessing community effects of moderation interventions on

r/the_donald. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–28.

Trujillo, A.; and Cresci, S. 2022b. Make reddit great again: assessing community effects of moderation interventions on r/the_donald. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–28.

Vox. 2017. Reddit's TheRedPill, notorious for its misogyny, was founded by a New Hampshire state legislator. https://www.vox.com/culture/2017/4/28/15434770/red-pill-founded-by-robert-fisher-new-hampshire. Accessed on 2024-06-15.

Wacquant, L. J. 1990. Review essay: Exiting roles or exiting role theory? critical notes on ebaugh's becoming an ex. *Acta sociologica*, 33(4): 397–404.

Wagner, A. K.; Soumerai, S. B.; Zhang, F.; and Ross-Degnan, D. 2002. Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*, 27(4): 299–309.

Waller, I.; and Anderson, A. 2021. Quantifying social organization and political polarization in online platforms. *Nature*.

Xiao, S.; Cheshire, C.; and Bruckman, A. 2021. Sensemaking and the Chemtrail Conspiracy on the Internet: Insights from Believers and Ex-believers. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–28.

Zannettou, S. 2021. "I Won the Election!": an empirical analysis of soft moderation interventions on Twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 15, 865–876.

Zuckerman, E.; and Rajendra-Nicolucci, C. 2021. Deplatforming Our Way to the Alt-Tech Ecosystem. *Knight First Amendment Institute at Columbia University*, 11.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? No
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (e) Did you address potential biases or limitations in your theoretical framework? Yes
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? No, given that the license is ill-defined.
   (c) Did you include any new assets in the supplemental material or as a URL? No

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Prompts

```
You are tasked with determining whether a comment posted by a migrating user
on the r/exredpill subreddit violates the community guidelines of r/exredpill.

I will provide (1) the community guidelines and (2) the comment posted on r/exredpill.

Your answer must be formatted as follows:
1) "Answer: Yes" if the comment posted on r/exredpill violates the community guidelines

2) "Answer: No" if the comment does not violates the community guidelines


** Community Guidelines of r/exredpill ** : [...]

** Comment posted on r/exredpill **: [...]

Does the comment posted on r/exredpill violate the community guidelines?
```

```
You are tasked with evaluating whether a comment posted
by a user in r/exredpill aligns with the ideologies of fringe communities
such as r/Braincels, r/MGTOW, and r/TheRedPill.
You will be provided descriptions of these ideologies and the comment to analyze.

Your answer must be formatted as follows:
1) "Answer: Yes" if the comment posted on r/exredpill violates the community guidelines

2) "Answer: No" if the comment does not violates the community guidelines

** Description of incels ideology  ** : [...]
** Description of MGTOW ideology  ** : [...]
** Description of TheRedPill ideology  ** : [...]


** Comment posted on r/exredpill **: [...]


Is the comment posted on r/exredpill aligned with the description of incels,

mgtow, theredpill ideology?
```

Activity Volume				
	Braincels+TheRedPill		MGTOW	
	Quarantine	Banning	Quarantine	Banning
β_1	-0.03	0.02	0.15	0.02
	(0.494)	0.853	(0.014)	0.853
β_2	-17.4	24.06	0.91	86.49
	(0.06)	(0.001)	(0.881)	(0.001)
β_3	0.16	0.259	-0.0004	-0.63
	(0.002)	(0.182)	(0.988)	0.001
Others				
(Intercept)	10.737	22.124	9.344	98.432
	(0.002)	(0.000)	(0.000)	(0.001)
R^2	0.017	0.025	0.047	0.014
Newcomers				
β_1	-0.004	3.788	0.002	-0.216
	(0.771)	(0.306)	(0.114)	(0.623)
β_2	-2.12	5.30	-0.69	11.74^{***}
	(0.261)	(0.001)	(0.663)	(0.001)
β_3	0.03	0.04	0.04	-0.219^{*}
	(0.130)	(0.042)	(0.988)	(0.089)
Others				
(Intercept)	4.281^{***}	5.621	3.222	19.439
	(0.001)	(0.002)	(0.001)	(0.001)
Adj. R^2	0.034	0.038	0.033	0.033
$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05; ^{\cdot}p < 0.1$				
Migrants from Fringe Communities				
β_1	0.167	3.788	0.005	-0.216
	(0.267)	(0.306)	(0.172)	(0.623)
β_2	-0.01	1.24	-0.031	21.10
	(0.318)	(0.003)	(0.058)	(0.001)
β_3	0.044	-0.012	0.029	0.042
	(0.020)	(0.318)	(0.094)	(0.441)
Others				
(Intercept)	2.859	3.206	6.408	18.123
	(0.007)	(0.001)	(0.002)	(0.003)
Adj. R^2	0.023	0.012	0.027	0.018
$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05; ^{\cdot}p < 0.1$				

Table 4: Summary of results. The ITS coefficient estimates (top) activity volume, (mid) newcomers, (bottom) number of migrants. Coefficient estimates and p-values in parenthesis.

banning event	outcome variable	p-value	relative effect	95% CI	absolute effect (s.d.)
Braincels ban	activity volume	0.032	88.4%	[12.5%, 179.2%]	1241 (883)
	new users	0.004	174.3%	[10.4%, 388.2%]	87.2 (41.07)
	migrating users	0.001	94.6%	[35.4%, 207.2%]	45.71 (10.32)
MGTOW ban	activity volume	0.001	64.5%	[45.6%,91.3%]	5161 (564)
	new users	0.001	31.7%	[17.7%,48.2%]	166.8 (32.76)
	migrating users	0.006	22.8%	[3.3%,56.9%]	20.73 (12.0)

Table 5: BSTS results of bans effects on volume activity, newcomers, and migrating users within r/exredpill.

quarantine event	outcome variable	p-value	relative effect	95% CI	absolute effect (s.d.)
Braincels & TRP[1] quarantine	activity volume	0.345	15.4%	[-34.3%, 65.9%]	42.1(138.9)
	new users	0.167	15.3%	[-19.6%, 59.2%]	17.43 (21.69)
	migrating users	0.045	24.6%	[4.7%, 53.1%]	13.03 (12.66)
MGTOW quarantine	activity volume	0.093	38.5%	[-9.3%,73.6%]	797(264.9)
	new users	0.223	13.7%	[-29.2%, 47.8%]	58.76(82.5)
	migrating users	0.468	3.4%	[-25.3%, 48.4%]	0.77 (20.05)

Table 6: BSTS results of quarantine effects on volume activity, newcomers, and migrating users within r/exredpill.

banning event	outcome variable	p-value	relative effect	95% CI	absolute effect (s.d.)
Braincels ban	activity volume	0.105	10.5%	[-5.3%, 22.3%]	124 (88.3)
	new users	0.122	15.3%	[-4.2%, 38.2%]	8.72 (4.11)
	migrating users	0.132	9.6%	[-3.4%, 20.7%]	4.57 (1.03)
MGTOW ban	activity volume	0.114	6.5%	[-4.6%, 9.1%]	516 (56.4)
	new users	0.125	3.7%	[-1.7%, 4.8%]	16.7 (3.27)
	migrating users	0.148	2.8%	[-3.3%, 5.6%]	2.07 (1.2)

Table 7: Placebo: BSTS results of banning effects on volume activity, newcomers, and migrating users within r/exredpill.

quarantine event	outcome variable	p-value	relative effect	95% CI	absolute effect (s.d.)
Braincels & TRP[1] quarantine	activity volume	0.345	1.5%	[-3.4%, 6.5%]	4.21 (13.89)
	new users	0.367	1.3%	[-1.9%, 5.9%]	1.74 (2.17)
	migrating users	0.412	2.4%	[-0.7%, 5.3%]	1.30 (1.27)
MGTOW quarantine	activity volume	0.393	3.5%	[-0.9%, 7.3%]	7.97 (26.49)
	new users	0.423	1.3%	[-2.9%, 4.7%]	5.88 (8.25)
	migrating users	0.468	0.4%	[-2.5%, 4.8%]	0.08 (2.01)

Table 8: Placebo: BSTS results of quarantine effects on volume activity, newcomers, and migrating users within r/exredpill.