Not Everyone Feels the Same: Engagement Profiles and User Reactions to Reddit's Great Ban

Short Paper

Ye Fan
Brunel Business School, Brunel University of London,
Kingston Lane, Uxbridge, London, UB8 3PH, UK
Ye.Fan@brunel.ac.uk

Jung Min Jang
Brunel Business School, Brunel University of London,
Kingston Lane, Uxbridge, London, UB8 3PH, UK
Jungmin.Jang@brunel.ac.uk

Habin Lee
Brunel Business School, Brunel University of London,
Kingston Lane, Uxbridge, London, UB8 3PH, UK
Habin.Lee@brunel.ac.uk

Abstract

Online platforms increasingly function as digital societies, where top-down moderation interventions like subreddit bans aim to regulate user behavior. However, user responses vary widely, and prior research offers mixed evidence of effectiveness. Guided by theories of psychological reactance and rationalization, and drawing on the Social Media Engagement Behavior framework, this study examines how pre-ban behavioral engagement metrics explain changes in toxicity following subreddit bans. Using a dataset of 1,798 disruptive users across 15 banned subreddits, we analyze pre–post toxicity changes via Google's Perspective API and multiple regression. We conceptualize engagement on Reddit across three levels: platform, subreddit, and activity. Findings reveal that users with higher pre-ban engagement intensity and greater behavioral consistency tend to increase toxicity, while those with exclusive subreddit focus are more likely to reduce it. These results demonstrate the explanatory value of engagement profiles and offer implications for more targeted, data-driven moderation strategies in online communities.

Keywords: Online communities, platform governance, moderation strategies, subreddit bans, user behavioral metrics, toxicity analysis, behavioral adaptation

Introduction

Online platforms increasingly function as digital societies, where norms, values, and rules are shaped not only by users but also by platform governance mechanisms that structure user interactions. As these ecosystems grow in complexity, maintaining constructive engagement while curbing anti-social behavior has become a core challenge in platform governance (Seering et al., 2020; Chandrasekharan et al., 2017). Among various moderation strategies, community bans represent one of the strictest forms, aimed at dismantling environments that normalize hate speech, harassment, or misinformation.

Reddit, a pseudonymous platform comprising approximately 3.8 million user-created subreddits - of which around one million are currently active and moderated - exemplifies these governance challenges. In June

2020, Reddit enacted one of its most extensive moderation efforts to date, including subreddit quarantines, content removals, and, in severe cases, outright subreddit bans. This initiative, known as "The Great Ban," resulted in the removal of over 2,000 subreddits that consistently violated Reddit's content policies. These included controversial communities such as r/The_Donald and r/ChapoTrapHouse, which were frequently cited for incivility and hate speech. The intervention marked a critical evolution in Reddit's role - from reactive moderation to proactive governance, positioning Reddit as a platform capable of enforcing behavioral norms akin to institutional governance in offline societies.

Despite their prominence, the effectiveness of subreddit bans as a governance intervention remains contested. While some studies report that bans reduce platform-wide toxicity and successfully remove harmful users (Chandrasekharan et al., 2017), others observe spillovers, persistence among core members, or migration to alternative venues (Habib et al., 2019; Cima et al., 2024; Trujillo & Cresci, 2022). These mixed findings mirror heterogeneous user responses in reality: while some users disengage or reduce toxic behavior, others escalate their harmful activity or shift to alternative venues.

To better understand this phenomenon, we argue that it is crucial to move beyond aggregate group-level observations and examine individual-level behavioral differences. Prior work in Information Systems has investigated user responses to algorithmic moderation and governance mechanisms (Seering et al., 2020; Trujillo & Cresci, 2022), but has largely focused on aggregate outcomes without explaining why behaviors change or which user characteristics or behavioral metrics drive those changes. Similarly, computer-science research on toxicity has emphasized detection models and technical content moderation tools, often overlooking the behavioral mechanisms that explain user adaptation or resistance (Abbasi et al., 2022).

Taken together, these limitations underscore a critical knowledge gap: we still lack a robust understanding of how individual users' behavioral histories shape their post-ban trajectories. As digital platforms increasingly rely on large-scale governance interventions to manage harmful content, this gap prompts timely and important research questions: Which types of users resist or comply with subreddit bans? How do distinct pre-ban behavioral patterns influence post-ban outcomes? Understanding these dynamics is critical for advancing IS research on digital governance and for informing more adaptive, user-sensitive moderation strategies in complex online ecosystems.

This study addresses this gap by investigating how granular, pre-ban behavioral engagement metrics influence post-ban changes in user toxicity, drawing on psychological theories of reactance and rationalization. Leveraging a large-scale dataset from fifteen subreddits banned during Reddit's "Great Ban" and applying Google's Perspective API, we evaluate user-level toxicity before and after the intervention, enabling us to trace individual behavioral trajectories over time.

Specifically, we ask:

RQ1: Does banning toxic subreddits lead to measurable changes in users' anti-social behavior?

RQ2: Which user types, identified through pre-ban behavioral metrics, are more likely to exhibit increased (reactance) or decreased (rationalization) toxicity post-ban?

By answering these questions, this study advances research on online toxicity and digital governance in three important ways. First, it provides theoretical advancement by introducing the Social Media Engagement Behavior (SMEB) framework as an overarching structure to examine behavioral engagement. By integrating this framework with psychological theories, specifically, psychological reactance (Brehm, 1966) and rationalization (Kay et al., 2002), we move beyond descriptive accounts to develop a more nuanced conceptual model that explains divergent user responses to governance interventions. Second, we contribute through granular user profiling across multi-level engagement. Specifically, this study makes a novel methodological contribution by being the first to operationalize behavioral metrics across three structural levels—platform-wide (Reddit), community (subreddit-level), and individual activity—within the context of Reddit. This multi-level engagement framework enables granular user profiling that captures the architectural complexity of the platform. By adopting this approach, we provide a more sophisticated analytical lens to identify heterogeneous user behaviors and examine user responses, addressing the limitations of prior studies that treated users as behaviorally uniform. This approach opens new avenues for understanding the behavioral dynamics of moderation interventions and delivers deeper, more contextualized insights into how and why user responses to bans diverge. Third, we generate actionable

insights for digital governance by empirically uncovering the causal relationships between pre-ban engagement profiles and post-ban toxicity outcomes. These findings reveal which user types are more likely to escalate or reduce anti-social behavior following intervention, offering practical guidance for designing more adaptive, personalized, and equitable moderation strategies in large-scale online platforms.

Background

The Effectiveness of Subreddit Bans

Previous research offers mixed findings on the impact of subreddit bans. Chandrasekharan et al. (2017) reported that Reddit's 2015 bans of r/fatpeoplehate and r/CoonTown were largely effective, with a significant number of account suspensions and an 80% reduction in hate speech among remaining users. Saleem and Ruths (2018) found that members of r/fatpeoplehate reduced their engagement and commenting activity, with some leaving Reddit entirely. Others attempted to recreate banned communities or infiltrate related subreddits, but these efforts were suppressed by coordinated actions from administrators and moderators.

More recent research presents a nuanced or skeptical view. Trujillo and Cresci (2022) found that although user activity dropped temporarily after r/The_Donald was quarantined, toxicity levels rebounded shortly thereafter. Cima et al. (2024) examined fifteen subreddits banned during the 2020 "Great Ban," reporting that 15.6% of users left the platform, while the remaining users reduced their toxicity by 6.6% on average. However, 5% of users exhibited significantly increased toxicity, illustrating the risk of unintended consequences.

Other research questions the strategy's overall effectiveness. Habib et al. (2019), in a longitudinal study using a dataset comprising the 3,000 most active subreddits, including 38 banned or quarantined subreddits, 118 identified as hateful, and 152 related communities, found no significant reduction in offensive language post-ban, noting that users often migrated to alternative hateful communities. Russo et al. (2023) documented similar patterns of cross-platform migration, with users returning to Reddit from less-moderated platforms, often displaying even more toxic behavior.

While these studies provide valuable insights, several gaps remain. First, most research focuses on one or two specific communities, which may introduce community-specific bias. Although Cima et al. (2024) included fifteen subreddits, their analysis treated all users as a single group, thereby overlooking meaningful individual-level variation. Second, and more importantly, prior work has yet to explain why user responses to bans diverge. Although previous studies document what happens after a ban, such as reduced toxicity, increased defiance, or user migration, they often fall short in explaining the underlying behavioral mechanisms. We argue that incorporating pre-ban engagement metrics provides explanatory leverage in distinguishing users who are more likely to resist moderation (reactance) from those who adapt their behavior (rationalization).

User Response to Community Ban: Engagement Profiles, Reactance, and Rationalization

To address the lack of individual-level analysis in prior moderation studies, our study shifts the analytical focus to user-level engagement profiles as a lens to explain heterogeneous behavioral trajectories. In social media research, engagement is commonly defined as communication or interaction among users. This concept has been shaped by theoretical insights across disciplines, including organizational behavior and marketing (Brodie et al., 2011; Saks, 2006). Brodie et al. (2013) further describe engagement as a psychological state and process linked to loyalty. It is frequently understood as a motivational construct that varies in intensity, involves both a subject and an object, and carries a valence (positive or negative) (Brodie et al., 2011; Hollebeek et al., 2022).

To operationalize this construct in a Reddit context, we adopt the Social Media Engagement Behavior (SMEB) framework developed by Dolan et al. (2016). In social media contexts, engagement behaviors typically include content consumption, contribution, and creation (Muntinga et al., 2011). These behaviors can be positioned along a continuum of intensity - from passive (e.g., browsing), to active (e.g.,

commenting), to highly active participation (e.g., content creation or moderation) (Malthouse et al., 2013; Muntinga et al., 2011).

User Engagement Behavioral Metrics	Description	Measurement	Related data
At the platform level:			
Account tenure	Length of time a user has been active on Reddit	Number of days from the first comment on Reddit to the fixed baseline date (June 29, 2020)	Timestamp of first comment
Posting frequency	Frequency of commenting on Reddit	Average number of comments per day	Comment timestamps
Average comment length	Verbosity of user's contributions	Mean number of words per comment	Comment text
At the subreddit level:			
Subreddit tenure	Length of time user has been active in a specific subreddit	Number of days from the first activity on Reddit to the fixed baseline date (June 29, 2020)	Timestamps of first comment/ post in subreddit
Exclusivity of subreddit participation	Degree to which user concentrates activity in the banned subreddit	Inverse of the number of unique subreddits participated in	Comment metadata
Relative dedication to the subreddit	Proportion of total Reddit comments posted in the banned subreddit	Ratio of comments in banned subreddit to total Reddit comments	Comment metadata
At the activity level:			
Toxicity direction over time	Evolution of toxicity in user behavior	F_score (Mall et al. 2020): average weighted difference in toxicity across adjacent comments, incorporating time decay	Comment timestamps and toxicity scores
Magnitude of toxicity fluctuation	Consistency of toxicity behavior	F_score - G_score * G_score is defined similarly to F_score but excludes absolute value in the equation (Mall et al. 2020).	Comment timestamps and toxicity scores
Average sentiment tone	Overall emotional tone of user comments	Mean VADER sentiment score across comments	Comment text

Table 1. Behavioral Metrics to Measure User Engagement Level in Reddit

Specifically, we construct a three structural level engagement profile:

Platform level engagement: Captures overall user investment in Reddit through (1) account tenure (duration of Reddit membership), (2) posting frequency, and (3) average comment length.

Subreddit level engagement: Reflects the degree of user identification with a specific community via (1) subreddit tenure (length of participation in the subreddit), (2) exclusivity of subreddit participation (how

exclusively users engage with that subreddit), and (3) relative dedication to the subreddit (the proportion of effort dedicated to that subreddit compared to others).

Activity level engagement: Assesses consistency and emotional tone of user behavior through (1) toxicity direction over time (the general trend of comment toxicity), (2) magnitude of toxicity fluctuation (variation in toxicity across posts), and (3) average sentiment tone (the overall emotional valence of user contributions).

We interpret these engagement patterns through two social-psychological lenses. Complementing this perspective, Hollebeek et al. (2022) show that externally imposed influence, such as subreddit bans, can elicit compliance or reactance, depending on how users internalize the intervention. This behavioral insight reinforces our theoretical framing of user responses as either resistance or adaptation.

Building on this, psychological reactance theory (Brehm, 1966), individuals may respond to perceived threats to their autonomy with defiance—manifesting as increased toxic behavior. Users with high engagement intensity or consistent expression are more likely to experience bans as identity threats, prompting resistance. In contrast, rationalization theory (Kay et al., 2002) suggests that when restrictions are perceived as legitimate or inescapable, individuals may reinterpret them as justified. This is more likely when users are heavily embedded in a single community and lack clear alternatives (Laurin et al., 2012).

To advance this line of inquiry, our study leverages these theoretical foundations alongside fine-grained engagement metrics to develop exploratory models of user response. Specifically, we examine how pre-ban behavioral characteristics, such as posting frequency, sentiment tone, and consistency, shape post-ban changes in toxicity. By anchoring our approach in Reddit's multi-level engagement architecture and grounding it in social-psychological theory, our study aims to move beyond descriptive analysis and offer a robust, theory-driven explanation for the divergent behavioral outcomes observed in response to community bans.

Data and Methods

Core User Selection

To investigate user behavioral trajectories following governance interventions, we constructed a dataset from fifteen subreddits removed during Reddit's "Great Ban" on June 29, 2020. These subreddits were selected based on the criteria developed by Cima et al. (2024). We extracted historical user activity using the Arctic Shift API (ArthurHeitmann, 2024), an open-source, research-friendly archive of Reddit comment data. Our initial data collection included all posted comments from November 1, 2019, to April 10, 2020, across the banned subreddits. To ensure behavioral consistency and minimize noise from infrequent participants, we identified core users as those who posted at least once per month in one or more of the targeted subreddits during the collection period. This filtering yielded an initial cohort of 6,897 core users, contributing 2,720,207 comments.

Removal of Deleted and Bot Accounts

To improve data validity, we implemented two layers of account filtering. First, we removed one anonymous account labeled '[deleted]', which pooled activities from all deleted users, thus their individual-level activity is indistinguishable and untraceable. This account alone contributed 901,802 comments, which were excluded. Second, we filtered out likely automated accounts (bots); accounts were removed if they (1) posted five or more identical comments in the same subreddit or (2) posted more than two comments within two seconds. Short, generic expressions (e.g., "Yes," "Nice," "NSFW") were exempted to reduce false positives from human behavior. This process eliminated 355 suspected bot accounts, including AutoModerator, with 620,699 comments. The final filtered dataset consisted of 6,541 human core users and 1,197,706 comments.

Identification of Disruptive Users

Building on prior work, reactance and rationalization are both motivational processes (see Kay et al., 2002; Wortman & Brehm, 1975) that are especially likely to emerge when individuals perceive restrictions as personally relevant. Our analysis focuses on disruptive users—those most likely to be influenced by governance interventions and thus theoretically relevant to models of reactance and rationalization. In our study, a disruptive user is defined as an individual who posted at least three toxic comments during the pre-ban period. Toxicity was assessed using the toxicity attribute of Google's Perspective API (Jigsaw, n.d.),

which assigns each comment a score from 0 to 1 based on the likelihood that it would be perceived as toxic – that is, rude, disrespectful, or likely to discourage participation.

Perspective API was selected for its scalability, interpretability, and empirical validation across multiple peer-reviewed studies on online toxicity (Mudambi & Viswanathan, 2022; Fan et al., 2024). It has been trained on millions of human-labeled texts from diverse online platforms, enabling consistent application across large-scale datasets. While alternative tools exist, Perspective API's transparent scoring and proven reliability make it especially appropriate for longitudinal behavioral analysis.

Following best-practice guidelines from the API documentation (Jigsaw, n.d.), we applied a threshold of 0.9 to classify comments as toxic. Applying this rule, we identified 1,798 users who met the criteria for disruptive behavior based on their pre-ban activity, which is for the final regression analysis.

Creation of Pre-Ban Behavioral Metrics and Toxicity for Disruptive Users

To analyze behavioral change, we collected Reddit-wide activity for each of the 1,798 disruptive users across two three-month windows: before the ban (March 29 - June 29, 2020) and after the ban (June 30 - September 29, 2020). For each user, comments were retrieved via the Arctic Shift API, resulting in 829,417 pre-ban and 146,628 post-ban comments. Toxicity scores for each user in each period were computed as the proportion of toxic comments (toxicity ≥ 0.9) to total comments, consistent with prior operationalizations. To ensure causal interpretability, all behavioral metrics were derived from pre-ban activity only and are detailed in Table 1. These variables represent user engagement across three structural levels (platform, subreddit, and activity) and serve as independent variables in the subsequent analysis. During this process, four users were excluded due to missing data, resulting in a final sample of 1,794 disruptive users.

Ethical Considerations

This study uses only publicly available Reddit data, adhering to ethical standards in computational social science. All variables were derived from users' behavioral histories, with no access to or use of personally identifiable information (e.g., usernames, demographics, IP addresses).

Data Analysis and Results

To evaluate post-ban behavioral change, we computed each user's toxicity change score as the difference between post-ban and pre-ban toxicity. Toxicity was calculated as the proportion of comments classified as toxic (toxicity ≥ 0.9). Positive values indicate increased toxicity after the ban, while negative values indicate decreased toxicity.

Overall Effectiveness of Subreddit Bans

Among the 1,798 identified disruptive users, the majority (1,461 users, 81.3%) exhibited a decrease in toxic commenting after the ban, suggesting behavioral adaptation consistent with rationalization. Notably, 880 users (48.9%) completely ceased posting toxic content. In contrast, 336 users (18.7%) either increased their toxic behavior, indicating potential psychological reactance in response to the intervention. One user showed no change in toxicity and was excluded from further analysis.

Toxicity Change Category	Proportion of Users		Operational Definition	Explanation
Rationalization (Toxicity Decreased)	81.3%		Toxicity change score < 0	Adaptive behavior
Reactance (Increased Toxicity)	18.7%		Toxicity change score > 0	Resistant behavior

Table 2. Post-Ban Toxicity Outcomes for Disruptive Users (N=1,797)

The Impact of Behavioral Engagement Metrics on Toxicity Change

To investigate whether pre-ban behavioral engagement contributes to changes in post-ban toxicity, we conducted a multiple linear regression analysis using SPSS (Table 3). Independent variables were derived from users' pre-ban activity, operationalized across three structural levels of engagement - platform, subreddit, and activity - to capture different dimensions of user commitment and behavioral tendencies. Before interpreting results, we verified the model's reliability. Correlations among independent variables were all below 0.6, and Variance Inflation Factor (VIF) values were well below the conventional threshold of 5, confirming no issues of multicollinearity. Additionally, Test for Linearity examining the relationship between each independent and dependent variable confirmed that the linearity was statistically significant (*ps* < .05), while deviation from linearity was not (*ps* > .10), indicating the linear model provides an appropriate fit.

	Unstandardized Coefficients				Collinearity Statistics	
	β	SE	t	*p*	Tolerance	VIF
Constant	-.0135	.0006	-22.8994	.0000		
Account tenure	.0005	.0006	.7045	.4812	.8306	1.2039
Posting frequency	.0036**	.0008	4.4293	.0000	.5137	1.9466
Average comment length	.0020**	.0006	3.4179	.0006	.9747	1.0259
Subreddit tenure	.0005	.0006	.7389	.4600	.8279	1.2078
Exclusivity of subreddit participation	-.0041**	.0007	-6.2992	.0000	.8148	1.2273
Relative dedication to the subreddit	.0031**	.0007	4.3941	.0000	.6724	1.4873
Toxicity direction over time	-.0005	.0006	-.8276	.4080	.9179	1.0895
Magnitude of toxicity fluctuation	-.0025**	.0008	-3.1269	.0018	.5544	1.8037
- Average sentiment tone	.0032**	.0006	5.1377	.0000	.9165	1.0911
Model summary	R2 = .0736, F(9,1783) = 15.7377, *p* < .001					

Table 3. Regression Analysis Results (N = 1,793, *p < .05, **p < .01)

Platform level engagement: Posting frequency (β = .0036, p < .001) and average comment length (β = .0020, p < .001) both have significant positive effects on post-ban toxicity, whereas account tenure is not significant. These results suggest that highly active users who post frequently and write longer comments were more likely to increase toxic behavior after the ban. This pattern aligns with psychological reactance theory (Brehm, 1966), which posits that individuals with stronger behavioral investments may perceive bans as threats to expressive autonomy, leading to oppositional responses. In this context, frequent contributors may interpret subreddit bans as unjust censorship, triggering reactance.

Subreddit level engagement: Exclusivity of subreddit participation (β = –.0041, p < .001) has a significant negative effect on post-ban toxicity, while relative dedication to the banned subreddit (β = .0032, p < .001) had a significant positive effect. Subreddit tenure is not significant. These findings indicate that users who participated almost exclusively in a single subreddit were more likely to reduce toxic behavior post-ban, whereas those with high relative dedication—but not exclusivity—were more likely to escalate toxicity. This divergence can be explained by rationalization theory (Kay et al., 2002), which suggests users without viable alternatives are more likely to accept and rationalize imposed constraints. Conversely, users with high investment but available alternatives may reject the ban as illegitimate, intensifying their resistance.

Activity level engagement: The magnitude of toxicity fluctuation (β = –.0025, p < .01) negatively indicates post-ban toxicity, while average sentiment tone (β = .0032, p < .001) is positively associated. The toxicity direction over time is not significant. These findings imply that users with emotionally consistent commenting behavior (low fluctuation) and more positive sentiment expressions pre-ban were more likely to become toxic afterward. In line with psychological reactance theory (Brehm, 1966; Laurin et al., 2012; Proudfoot & Kay, 2014), such users may view the ban as a threat to their established behavioral identity, eliciting resistance. Stability in either prosocial or toxic patterns appears to signal internalized norms, making externally imposed constraints feel more intrusive.

Overall, these findings reveal that user response to subreddit bans is systematically shaped by their pre-ban engagement profiles. Reactance is most pronounced among users with high behavioral intensity, emotional consistency, and perceived expressive investment, while rationalization is more likely among users whose engagement context limits alternatives. These results substantiate the theoretical proposition that user behavior following platform interventions is not random but grounded in prior engagement dynamics and psychological mechanisms.

Conclusion

This study advances Information Systems research by demonstrating that pre-ban behavioral engagement metrics meaningfully explain post-ban behavioral change following governance interventions, specifically, subreddit bans. By operationalizing engagement across three structural levels (platform, subreddit, activity), we demonstrate that responses to bans are not uniform: users' engagement intensity and behavioral consistency prior to the ban provide explanatory leverage in explaining whether toxicity increases or declines afterward. Our findings highlight the practical value of user profiling for anticipating the outcomes of moderation and for informing adaptive, data-driven governance strategies in complex digital platforms.

Several limitations should be noted. First, our dataset is Reddit-specific, limiting generalizability to other platforms. Second, our behavioral measures may not capture all factors influencing post-ban behavior. Third, the analysis is based solely on observable actions, without direct insight into users' psychological states. Future research should expand to other platforms, incorporate additional behavioral indicators, and explore user perceptions to deepen understanding of post-intervention dynamics.

Acknowledgements

The authors gratefully acknowledge Hyunsang Lee, a graduate of Imperial College London, for his valuable assistance with the data collection process in this research.

References

Abbasi, A., Javed, A. R., Iqbal, F., Kryvinska, N., & Jalil, Z. (2022). Deep learning for religious and continent-based toxic content detection and classification. *Scientific Reports, 12*(1), 17478. https://doi.org/10.1038/s41598-022-22523-3

ArthurHeitmann. (2024). *arctic_shift* [Computer software]. GitHub. https://github.com/ArthurHeitmann/arctic_shift

Brehm, J. W. (1966). *A theory of psychological reactance*. Academic Press.

Brodie, R. J., Hollebeek, L. D., Jurić, B., & Ilić, A. (2011). Customer engagement: Conceptual domain, fundamental propositions, and implications for research. *Journal of Service Research, 14*(3), 252-271. https://doi.org/10.1177/1094670511411703

Brodie, R. J., Ilic, A., Juric, B., & Hollebeek, L. (2013). Consumer engagement in a virtual brand community: An exploratory analysis. *Journal of Business Research, 66*(1), 105-114. https://doi.org/10.1016/j.jbusres.2011.07.029

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, 1*(CSCW), Article 31, 1-22. https://doi.org/10.1145/3134666

Cima, L., Trujillo, A., Avvenuti, M., & Cresci, S. (2024, May). The Great Ban: Efficacy and unintended consequences of a massive deplatforming operation on Reddit. In *Companion Publication of the 16th ACM Web Science Conference*, 85-93. https://doi.org/10.1145/3630744.3663608

Dolan, R., Conduit, J., Fahy, J., & Goodman, S. (2016). Social media engagement behaviour: a uses and gratifications perspective. *Journal of Strategic Marketing, 24*(3-4), 261-277. https://doi.org/10.1080/0965254X.2015.1095222

Fan, L., Li, L., & Hemphill, L. (2024). Toxicity on social media during the 2022 Mpox public health emergency: Quantitative study of topical and network dynamics. *Journal of Medical Internet Research, 26*, e52997. https://doi.org/10.2196/52997

Habib, H., Musa, M. B., Zaffar, F., & Nithyanand, R. (2019). To act or react: Investigating proactive strategies for online community moderation. *arXiv preprint* arXiv:1906.11932. https://doi.org/10.48550/arXiv.1906.11932

Hollebeek, L. D., Sprott, D. E., Sigurdsson, V., & Clark, M. K. (2022). Social influence and stakeholder engagement behavior: Conformity, compliance, and reactance. *Psychology & Marketing, 39*(1), 90-100. https://doi.org/10.1002/mar.21577

Jigsaw. (n.d.). *Perspective API Developer Documentation*. Retrieved August 22, 2025, from https://developers.perspectiveapi.com/s/?language=en_US

Kay, A. C., Jimenez, M. C., & Jost, J. T. (2002). Sour grapes, sweet lemons, and the anticipatory rationalization of the status quo. *Personality and Social Psychology Bulletin, 28*(9), 1300-1312. https://doi.org/10.1177/01461672022812014

Laurin, K., Kay, A. C., & Fitzsimons, G. J. (2012). Reactance versus rationalization: Divergent responses to policies that constrain freedom. *Psychological Science, 23*(2), 205-209. https://doi.org/10.1177/0956797611429468

Mall, R., Nagpal, M., Salminen, J., Almerekhi, H., Jung, S. G., & Jansen, B. J. (2020, October). Four types of toxic people: Characterizing online users' toxicity over time. *In Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping experiences, shaping society,* Article 37, 1-11. https://doi.org/10.1145/3419249.3420142

Malthouse, E. C., Haenlein, M., Skiera, B., Wege, E., & Zhang, M. (2013). Managing customer relationships in the social media era: Introducing the social CRM house. *Journal of Interactive Marketing, 27*(4), 270-280. https://doi.org/10.1016/j.intmar.2013.09.008

Mudambi, M., & Viswanathan, S. (2022). Prominence reduction versus banning: An empirical investigation of content moderation strategies in online platforms. In ICIS 2022 proceedings (Paper 15). Association for Information Systems. https://aisel.aisnet.org/icis2022/social/social/15

Muntinga, D. G., Moorman, M., & Smit, E. G. (2011). Introducing COBRAs: Exploring motivations for brand-related social media use. *International Journal of Advertising, 30*(1), 13-46. https://doi.org/10.2501/IJA-30-1-013-046

Proudfoot, D., & Kay, A. C. (2014). Reactance or rationalization? Predicting public responses to government policy. *Policy Insights from the Behavioral and Brain Sciences, 1*(1), 256-262. https://doi.org/10.1177/2372732214550489

Russo, G., Verginer, L., Ribeiro, M. H., & Casiraghi, G. (2023, June). Spillover of antisocial behaviour from fringe platforms: The unintended consequences of community banning. In *Proceedings of the International AAAI Conference on Web and Social Media, 17*, 742-753. https://doi.org/10.1609/icwsm.v17i1.22184

Saks, A. M. (2006). Antecedents and consequences of employee engagement. *Journal of Managerial Psychology, 21*(7), 600-619. https://doi.org/10.1108/02683940610690169

Saleem, H. M., & Ruths, D. (2018). The aftermath of disbanding an online hateful community. *arXiv preprint* arXiv:1804.07354. https://doi.org/10.48550/arXiv.1804.07354

Seering, J., Kraut, R. E., & Dabbish, L. (2020). Shaping pro- and anti-social behavior on Twitch through moderation and example-setting. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), Article 187, 1-27. https://doi.org/10.1145/3415211

Trujillo, M. Z., & Cresci, S. (2022). Make Reddit Great Again: Assessing community effects of moderation interventions on r/The_Donald. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), Article 1, 1-14. https://doi.org/10.1145/3567557

Wortman, C. B., & Brehm, J. W. (1975). Responses to uncontrollable outcomes: An integration of reactance theory and the learned helplessness model. In *Advances in experimental social psychology* (Vol. 8, pp. 277-336). Academic Press.