Selected Papers of #AoIR2021:
The 22nd Annual Conference of the
Association of Internet Researchers
Virtual Event / 13-16 Oct 2021

INCELS ON REDDIT: A STUDY IN SOCIAL NORMS AND DECENTRALISED MODERATION

Rosalie Gillett
ARC Centre of Excellence for Automated Decision-Making and Society and Digital Media Research Centre, QUT

Nicolas Suzor
ARC Centre of Excellence for Automated Decision-Making and Society and Digital Media Research Centre, QUT

Introduction

The social news website Reddit has a long history of hosting communities ('subreddits') that advocate or encourage white supremacy (Gillespie 2018), disparagement of minority groups (Topinka 2017), and violence against women (Massanari 2017). As a platform that relies heavily on volunteer moderators ("mods") to self-govern the subreddits (Matias 2016), Reddit has been criticised for failing to adequately enforce its site-wide rules (Gillespie 2018). Incels—an internet subculture that ascribes to deeply misogynistic beliefs—grew in visibility when they developed subreddits on Reddit. After ongoing criticism and media attention about harmful behaviour of incels both on and off the platform, Reddit imposed escalating sanctions and ultimately banned the most visible of these subreddits over a period of several years. In this paper, we focus on the interaction between formal rules and social norms in incel and related subreddits.

This paper aims to improve understanding about how problematic norms are contested in (partially-) decentralised systems of content moderation. We examine discourse about moderation to better understand the role of moderation teams in maintaining and changing social norms in their communities (Fiesler et al. 2018) and to examine the interaction between these norms and both site-wide and subreddit-specific rules. Our study shows how the moderators of major incel subreddits resisted external pressure by adopting techniques to neutralize their behaviour and by enforcing boundaries to insulate their communities from critics. Our analysis suggests that the threat of prohibition alone is unlikely to be sufficient to drive cultural change in problematic subreddits. We argue that content moderation is an insufficient frame to understand the regulation of harmful communities; real change requires addressing the underlying cultural norms rather than focusing on individual pieces of content.

Gillett, R., Suzor, N. (2021, October). *Incels on Reddit: A study in social norms and decentralised moderation*. Paper presented at AoIR 2021: The 22nd Annual Conference of the Association of Internet Researchers. Virtual Event: AoIR. Retrieved from http://spir.aoir.org.

Examining decentralised moderation on Reddit

In this study, we use archived Reddit data to examine how four subreddits responded to external pressure and criticism and how they worked to develop and maintain their rules and norms. We compare the discourse around moderation in the now-banned r/Incels and r/Braincels subreddits with that in r/ForeverAlone, where users discuss similar themes but do not generally identify as incels. We also include r/IncelsWithoutHate, a subreddit that its creator framed as a space for less hateful incels, which Reddit banned from the platform in March 2021 for violating the platform's rules against promoting hate. We draw on Sykes and Matza's (1957) neutralization theory to examine the techniques that Reddit's incel communities used to justify abusive speech in the face of heavy external pressure and criticism.

Findings and discussion

Our findings show how subreddits adapt to and resist pressure from Reddit and other users to comply with the site-wide rules. Ultimately, stricter enforcement of the formal site-wide rules did not challenge or displace the underlying ideologies that foster toxic communities. Reddit's threats to quarantine (a step that makes communities less visible to those who have not subscribed) or ban subreddits did not carry sufficient weight to substantially change the social norms in the incel subreddits. After several warnings from Reddit, the subreddits failed to make a serious effort to reduce the level of harmful content in their communities. Instead, as the incel subreddits became more defensive, the moderators and participants rigorously policed their boundaries to keep critics out. In the face of heavy external criticism, incels reinforced problematic norms by minimising their culpability for harmful actions. For instance, they often denied that their behaviour was harmful and instead refocused discussion on condemning the behaviour of their critics. This enabled them to justify their subreddits as online support groups for bullied and oppressed men. It was also common for incels to draw on baseless pseudo-scientific theories to provide intellectual legitimacy for their claims (Baele, Brace, and Coan 2019). This "turbocharged genetic determinism" (Ging 2017, 650) enabled incels to deny responsibility for their lack of romantic relationships and provided opportunities to justify their misogynistic attitudes. While some subreddits attempted to comply with the letter of externally imposed rules, we found that participants adopted similar neutralization techniques to justify their behaviour. The result, we argue, is that the communities were able to maintain a set of norms that continued to be highly problematic.

By contrast, our findings demonstrate how moderators of less misogynistic communities continuously work to reinforce prosocial norms, despite dealing with very similar issues and sharing some subscribers with the main incel communities. We found that r/ForeverAlone's moderation team worked consistently over time to foster a respectful community through the rules and moderation practices. For instance, these moderators more actively deleted harmful comments, posts, and banned those who published them. The moderators' responses importantly demonstrated the bounds of acceptable behaviour in the community and created a space where other participants could question and contest toxic claims.

Conclusion

To foster prosocial subreddits, our study suggests that Reddit must tackle the underlying identity and ideology that makes people feel hateful. Content moderation, by itself, is an insufficient tool to change harmful social norms. Driving real change when moderation is decentralized requires the committed and supported participation of moderators who can undertake the extensive work of tackling the underlying identity and ideology that brings hateful communities together.

References

Baele, Stephane J., Lewys Brace, and Travis G. Coan. 2019. "From 'Incel' to 'Saint': Analyzing the Violent Worldview behind the 2018 Toronto Attack." *Terrorism and Political Violence*. 1–25. https://doi.org/10.1080/09546553.2019.1638256.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* New Haven, CT: Yale University Press.

Ging, Debbie. 2017. "Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere." *Men and Masculinities* 22 (4): 638–57. https://doi.org/10.1177/1097184X17706401.

Fiesler, Casey, Jialun "Aaron" Jiang, Joshua McCann, Kyle Frye, and Jed R. Brubaker. 2018. "Reddit Rules! Characterizing an Ecosystem of Governance." In Twelfth International AAAI Conference on Web and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17898.

Massanari, Adrienne. 2017. "#Gamergate and The Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures." *New Media & Society* 19(3): 329–46. https://doi.org/10.1177/1461444815608807.

Matias, J. Nathan. "Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout." In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 1138–51. San Jose California USA: ACM, 2016. https://doi.org/10.1145/2858036.2858391.

Sykes, Gresham M., and David Matza. 1957. "Techniques of Neutralization: A Theory of Delinquency." *American Sociological Review* 22(6): 664–70. https://doi.org/10.2307/2089195.

Topinka, Robert J. 2018. "Politically Incorrect Participatory Media: Racist Nationalism on r/ImGoingToHellForThis." New Media & Society 20(5): 2050–69. https://doi.org/10.1177/1461444817712516.