6-14-2024

Guilty Until Proven Innocent?: The Effects Of Shadowbanning On Social Media User Perceptions

Mattie Bryant
University of Colorado Boulder, mattiefbryant@gmail.com

Jason B. Thatcher
University of Colorado at Boulder, jason.b.thatcher@gmail.com

Marten Risius
The University of Queensland, m.risius@business.uq.edu.au

Recommended Citation

Bryant, Mattie; Thatcher, Jason B.; and Risius, Marten, "Guilty Until Proven Innocent?: The Effects Of Shadowbanning On Social Media User Perceptions" (2024). *ECIS 2024 TREOS*. 29.
https://aisel.aisnet.org/treos_ecis2024/29

GUILTY UNTIL PROVEN INNOCENT?: THE EFFECTS OF SHADOWBANNING ON SOCIAL MEDIA USER PERCEPTIONS

TREO Paper

Mattie Bryant, University of Colorado Boulder, Boulder, CO, USA
mattie.bryant@colorado.edu

Marten Risius, School of Business, The University of Queensland, Brisbane, Queensland, Australia, m.risius@business.uq.edu.au

Jason B. Thatcher, University of Colorado Boulder, Boulder, CO, USA
jason.thatcher@colorado.edu

Abstract

Social media platforms use content moderation practices to prevent or contain the dissemination of harmful content online. These practices can be initiated by the user (e.g. reporting, blocking) or by the platform (e.g. content or account removal, visibility reduction) and can be carried out overtly or covertly (i.e., with or without notifying users). This paper examines the unknown implications of platform-initiated covert moderation methods, specifically shadowbanning, on SM user behavior, experience, and perceptions. We apply moral intuition theory to investigate how other users perceive and respond to users who claim to be shadowbanned, as well as to platforms that shadowban users.

Keywords: Shadowbanning, platform governance, content moderation, moral intuition theory.

1 Introduction

Content moderation practices are '*governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse*' (Grimmelmann, 2015). This can be achieved through overt or covert methods, which involve either humans, machine actors, or both (Jiang et al., 2023). Overt moderation is disclosed to users and is triggered by a breach of platform policy, while covert moderation is not shared with users and does not specify a reason for the action. While there are good reasons for platforms not to disclose the shadowban (e.g., malicious spammers cannot adjust to moderation practices), the side effects for regular users that are shadowbanned have not yet been considered.

Although overt content moderation (e.g., flags, nudges) has received a lot of attention (Gerrard, 2018), relatively less is known about shadowbanning, a covert form of content moderation. Shadowbanning refers to a platform deprioritizing or suppressing content posted for public consumption, without notifying the users and groups who created the content (Gillespie, 2022). There is often "constructive ambiguity" around shadowbanning, meaning that the banned content creator does not know they are banned. If a user is shadowbanned, they may not be aware of the reason for it, whether it was due to a breach of platform policy, their behavior on or off the platform, or algorithmic error. Shadowbanning is used by SM platforms to contain content or behavior, rather than to correct it. The offending content is simply hidden, and the user receives no further information about why.

While many SM platforms appear to use shadowbanning to contain content and behavior that violates platform policy, this practice is not consistently applied to all users or content (Duffy & Meisner, 2023). When asked about shadowbanning, SM platforms reluctantly admit to engaging in the practice – both intentionally and erroneously – but rarely offer explanations for their motivations, beyond shielding users from inappropriate or harmful content or behavior (Nicholas, 2022). Marginalized communities, such as black, LGBTQ+, or handicapped users, have reported being shadowbanned on SM platforms; platforms often claim that reducing people's visibility is due to algorithmic error (Nicholas, 2022).

1.1 Research Agenda

Therefore, while we know shadowbanning exists, its causes and implications are subject to interpretation by the banned users, other platform users, and the broader public. In a recent Pew study, 61 percent of US Americans reported that they prefer tech companies to take steps and restrict content even at the expense of their access to information (Stocking et al., 2022). However, we do not know how SM users respond to platforms that fail to explain why they shadowbanned a user. For example, parts of the public were upset when Meta reported a technical glitch was responsible for shadowbanning pro-Palestinian voices (Luu, 2023). Nor, for that matter, do we know if SM users respond negatively to platforms that ban users for legitimate policy breaches (Myers West, 2018). Hence, to explore shadowbanning's implications, we ask: *How do SM users respond to platforms that they learn use shadowbanning to manage users and content?*

In addition, we explore the effects of shadowbanning on banned SM users. Some users report experiencing 'black box gaslighting,' which refers to being told that their lack of understanding of SM content and promotion algorithms hinders them from accurately determining if they are banned or if others simply don't care for their content (Cotter, 2023). As a result, some users speculate publicly that they have been shadowbanned after a drop in engagement statistics (Duffy & Meisner, 2023) or when they feel isolated from their online connections (Lutz & Schneider, 2021). We are interested in exploring what determines how SM users view or interact with people who claim to have been shadowbanned. Hence, we ask: *How does self-disclosure of shadowbanning affect social media engagement?*

Our study of shadowbanning will draw on moral intuition theory, which suggests that the appraisal of a ban's rightness or wrongness is based on an individual's intuition (Haidt & Joseph, 2004), such as their feelings about the cause, the person, and the behavior that led to the ban. We anticipate that learning about a ban and its causes will lead to discrete responses from SM users, with some causes of bans resulting in sanctions on banned users and other causes having no effect.

1.2 Proposed Methodology

We will use a mixed methods approach to investigate these questions. Study 1 will evaluate how users respond to learning about shadowbans. We will identify platforms that shadowban users, based on news reports, and will assemble an archival dataset to evaluate the immediate and longer-term effect of shadow bans on user engagement, based on user behavior and engagement statistics. This analysis will provide more insight into user phenomena trends that occur in response to shadowbanning incidents.

Study 2 will offer further insight into why and when users respond to shadowbanning. We anticipate moral appraisals of shadow bans will be moderated by user perceptions of content moderation as a general practice. In addition, we suspect that if a user claiming to be shadowbanned states the reason for the shadowban, then the stated reason, the source of disclosure, and the appraising user's strength of conviction will further influence perceptions. To evaluate the impact of these factors, we will conduct a vignette experiment that allows us to emulate shadowbans while controlling for different characteristics of bans. We will measure resulting perceptions through close-ended and open-ended questions about perceived personal traits (e.g., warmth, competence, conflict propensity, reliability), social characteristics (e.g., identified groups, leadership ability, treatment of others), and judgment of actions.

2 Conclusion

Our studies will extend knowledge of shadowbanning impacts on SM platforms and on the people who disclose they have been shadowbanned. We will help to explain why and how shadowbanned users' and adjacent users' (e.g., bystanders) behaviors change after a shadowban claim. This will provide insight into the real-world repercussions of shadowbanning within SM platforms (e.g. SM engagement) and, potentially, outside of platforms (e.g. job screening). By doing this, we will shed light on how covert content moderation practices can have unintended consequences for SM platforms and banned users.

References

Cotter, K. (2023). “Shadowbanning is not a thing”: Black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*, *26*(6), 1226–1243.

Duffy, B. E., & Meisner, C. (2023). Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility. *Media, Culture & Society*, *45*(2), 285–304.

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12), 4492–4511.

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, *8*(3), 1-13.

Grimmelmann, J. (2015). The Virtues of Moderation. *Yale Journal of Law and Technology*, *17*, 42–109.

Haidt, J., & Joseph, C. (2004). Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus*, *133*(4), 55–66.

Jiang, J. A., Nie, P., Brubaker, J. R., & Fiesler, C. (2023). A Trade-off-centered Framework of Content Moderation. *ACM Transactions on Computer-Human Interaction*, *30*(1), 1–34.

Lutz, S., & Schneider, F. M. (2021). Is receiving Dislikes in social media still better than being ignored? The effects of ostracism and rejection on need threat and coping responses online. *Media Psychology*, *24*(6), 741–765.

Luu, J. (2023). *This musician says his pro-Palestinian posts were banned. Is social media being censored?* SBS News. Retrieved March 11, 2024, from https://www.sbs.com.au/news/the-feed/article/this-musician-says-he-was-shadowbanned-for-making-pro-palestinian-posts-is-social-media-being-censored/jipi1vvn3

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383.

Nicholas, G. (2022). Shedding light on shadowbanning. *Center for Democracy and Technology, 1-52.*

Stocking, G., Mitchell, A., Matsa, K. E., Widjaya, R., Jurkowitz, M., Ghosh, S., Smith, A., Naseer, S., & St Aubin, C. (2022). The role of alternative social media in the news and information environment. *Pew Research Center*. Retrieved March 11, 2024, from https://www.pewresearch.org/journalism/wp-content/uploads/sites/8/2022/10/PJ_2022.10.06_Alternative-Social-Media.pdf