Selected Papers of #AoIR2023:
The 24th Annual Conference of the
Association of Internet Researchers
Philadelphia, PA, USA / 18-21 Oct 2023

ALGORITHMIC FOLK THEORIES OF ONLINE HARASSMENT: HOW SOCIAL MEDIA ALGORITHMS ENABLE ONLINE HARASSMENT AND PREVENT INTERVENTION

Cait Lackey
University of Illinois Chicago

Samuel Hardman Taylor
University of Illinois Chicago

The prevalence of online harassment is undeniable, with 64% of U.S. citizens under the age of 30 having experienced some form of online harassment, and people with marginalized identities reporting being victimized most often (Lenhart & Zickuhr, 2016; Vogels, 2021). Enabled by networked Internet technologies, online harassment is a variety of abusive online behaviors, which target a specific person or group (Blackwell et al., 2017). Research suggests people are often harassed by groups connected on platforms who are motivated by moral outrage (Marwick, 2021). Social media algorithms, or computerized systems that curate, detect, and filter social media content based on pre-programmed model specifications, are often suggested as a scalable solution to reducing networked online harassment (Al-Garadi et al., 2019; Rosen, 2021; Taylor & Choi, 2022). Despite the vast resources dedicated to creating algorithms to mitigate harassment, online harassment continues to grow in frequency and severity (Vogels, 2021).

Research raises concerns about the inadequacies and inequalities of algorithms as content moderators because these systems are blind to the social issues and concerns contextualizing online harassment (Musgrave et al., 2022). This critique of algorithms introduces questions about what is missed at the intersection of online harassment and algorithms. Investigating victims, perpetrators, and bystanders' perceptions of algorithmic failures related to online harassment provides insight into their role in perpetuating online harassment.

Suggested Citation (APA): Lackey, C. & Taylor, S. H. (2023, October). *Algorithmic folk theories of online harassment: How social media algorithms enable online harassment and prevent intervention.* Paper presented at AoIR2023: The 24th Annual Conference of the Association of Internet Researchers. Philadelphia, PA, USA: AoIR. Retrieved from http://spir.aoir.org.

Algorithmic folk theories are a framework for investigating perceptions of social media algorithms because folk theories can shed light on how victims, preparators, and witnesses behave during online harassment episodes (DeVito et al., 2018). Algorithmic folk theories are the "intuitive, informal theories that individuals develop to explain the outcomes, effects, or consequences of technological systems" (Devito et al., 2017, p. 3165). Research suggests that algorithmic folk theories inform experiences of online harassment (DeVito, 2022; Karzait et al., 2021; Ytre & Moe, 2021). Actionable folk theories inform how users adapt their behavior to meet their goals within the confines of algorithms, whereas demotivational theories leave no clear avenue for action (DeVito, 2022). Given the importance of folk theories in understanding people's behavior on platforms, there remains important questions about how people who have been victims, perpetrators, or bystanders of online harassment theorize the role of algorithms in their experience and how these folk theories influence their behavior. Thus, the goal of this research is to explore:

RQ1: What folk theories and folk theorization do victims, perpetrators, and bystanders of online harassment have of algorithms?
RQ2: How does folk theorization impact people's behaviors during online harassment?

Method

We conducted semi-structured grounded theory interviews focusing on perceptions of social media algorithms and their relation to online harassment. Participants were recruited in a large U.S. Midwest city and required to be over 18 years old. Our participants identified as victims, instigators, and/or witnesses of online harassment on social media. We interviewed 19 individuals: 16 victims, 19 witnesses, and 6 preparators. Our sampling strategy focused on identity categories most likely to be victims of online harassment, such as LGBT individuals, young adults, women, and racial or ethnic minorities. Online harassment episodes were reported across a spread of social media platforms. After completing a brief introductory survey, qualifying participants completed Zoom interviews lasting between 60 to 80 minutes and were compensated $20.

There were multiple rounds of data collection to allow flexibility for analytic work and the development of emergent theory (Charmaz, 2006). Interviews were transcribed verbatim for coding. After each interview, we conducted initial grounded theory coding to identify gaps in research, allowing space to adjust for analytic leads based on the indications of emergent theoretical categories (Glazer, 1978). Following saturation, we utilized focused coding to synthesize data followed by a theoretical coding process to invoke comprehensible and grounded conceptual findings.

Findings

Participants provided folk theories of online harassment they witnessed, experienced, or instigated as a consequence of algorithmic design. We identify four algorithmic folk theories: (1) *the critical mass intervention theory*, (2) *the harassment amplifier theory*,

(3) *the algorithmic virus theory*, and (4) *the biased protection theory*. Each folk theory is associated with behavioral outcomes.

First, the *critical mass intervention theory* finds algorithmic intervention was only possible when a large number of people flag a harassment incident. Conversely, if victims personally flag their harassment on platforms, without soliciting additional intervenors, they failed to receive algorithmic attention. As bystander Brianna experienced, *"my friend reached out to me because she couldn't get her nudes taken down. She needed a bunch of people to flag it to get TikTok's attention."* Because of this theory, some victims worked to solicit large groups of support to receive intervening algorithmic moderation, whereas others without large social networks described helplessness.

Next, the victims, bystanders, and perpetrators of harassment described the *harassment amplifier theory* to perceive that algorithms are designed to amplify harassment content because it increases engagement. As perpetrator Asher describes, *"[Twitter's algorithm] ranks hate or hateful comments or negative constantly over positive because that will get clicks."* Thus, there was widespread belief that algorithms actively amplify harassment, which invigorated perpetrators to increase the toxicity of their comments.

Third, many victims were harassed by individuals outside of their personal networks. *The algorithmic virus theory* describes perceptions that algorithms intentionally network profiles and content within and across platforms, creating opportunities for harassment. As Alex describes, *"algorithms can spread and share my profile to other people pretty fast, just based off of similar interests. Kind of like a virus."* Victims experienced algorithmic spread of their social media interactions to potential harassers, and stopping the spread of the hate was beyond their control.

Last, *the biased protection theory* was described by victims and bystanders, which was the perception that algorithms fail to contextualize the harassment experienced by marginalized communities. As Jennifer laments, *"I should be able to report disability hate speech. I always report those and almost never get it removed. The algorithm comes back and says that they review it, but it didn't violate anything."* Despite victim and bystanders' efforts, algorithms often fail to reduce harm to these groups, which increase feelings of marginalization on social media.

Conclusion

Algorithmic folk theories described by victims, witnesses, and perpetrators of online harassment suggest that social media algorithms perpetuate online harassment. Our findings contribute to algorithmic folk theories suggesting that online harassment elicits novel folk theories associated with communicative actions from victims, bystanders, and perpetrators of online harassment. Some victims and bystanders used these folk

theories to curb online harassment; whereas perpetrators used their theories to increase the visibility and harm of their attacks. However, many reported their perceptions of algorithms can lead to unactionable behaviors, such as censorship or no longer reporting harassment, suggesting whether folk theories of online harassment are actionable and demotivational depends upon one's positionality in online harassment: victim, bystander, or perpetrator.

These folk theories also build theory on networked harassment suggesting that social media algorithms are experienced as amplifiers of harassment, in addition to popular accounts or communities (Marwick, 2021). Furthermore, victims and bystanders described leveraging algorithms to reach networked audiences of bystanders to end the harassment. Our research, ultimately, counter narratives from social media companies about algorithms as a scalable solution to harassment, as the lived experience of algorithms discourages intervention while encouraging in new types of online harassment and reinforcing marginalization.

References

Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., ... & Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges. *IEEE Access*, *7*, 70701-70718. http://doi.org/10.1109/ACCESS.2019.2918354

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction, 1*(CSCW), 1–19. https://doi.org/10.1145/3134659

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

DeVito, M. A., Gergle, D., & Birnholtz, J. (2017, May). " Algorithms ruin everything" #RIPTwitter, folk theories, and resistance to algorithmic change in social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 3163-3174). https://doi.org/10.1145/3025453.3025659

DeVito, M. A., Birnholtz, J., Hancock, J. T., French, M., & Liu, S. (2018, April). How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-12). https://doi.org/10.1145/3173574.3173694

DeVito, M. A. (2022). How transfeminine TikTok creators navigate the algorithmic trap of visibility via folk theorization. In *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1-31. https://doi.org/10.1145/3555105

Glaser, B. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. The Sociology Press.

Karizat, N., Delmonaco, D., Eslami, M., & Andalibi, N. (2021). Algorithmic folk theories and identity: How TikTok users co-produce knowledge of identity and engage in algorithmic resistance. In *Proceedings of the ACM on human-computer interaction*, *5*(CSCW2), 1-44. https://doi.org/10.1145/3476046

Lenhart, A., & Zickuhr, K. (2016). Online harassment, digital abuse, and cyberstalking in America. *Data & Society Research Institute.*

Marwick, A. E. (2021). Morally motivated networked harassment as normative reinforcement. *Social Media+ Society*, *7*(2). http://doi.org/20563051211021378

Musgrave, T., Cummings, A., & Schoenebeck, S. (2022, April). Experiences of harm, healing, and joy among Black women and femmes on social media. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-17). https://doi.org/10.1145/3491102.3517608

Rosen, G. (2021, November 1). *Hate speech prevalence has dropped by almost 50% on Facebook*. Meta. Retrieved February 23, 2023, from https://about.fb.com/news/2021/10/hate-speech-prevalence-dropped-facebook/

Taylor, S. H., & Choi, M. (2022). An Initial Conceptualization of Algorithm Responsiveness: Comparing Perceptions of Algorithms Across Social Media Platforms. *Social Media+ Society*, *8*(4). https://doi.org/10.1177/20563051221144322

Vogels, E. A. (2021). *The state of online harassment.* Pew Research Center.

Ytre-Arne, B., & Moe, H. (2021). Folk theories of algorithms: Understanding digital irritation. *Media, Culture, & Society*, *43*(5), 807-824. https://doi.org/10.1177/0163443720972314