The Importance of Looking Closer: Understanding Motivations for Bad Behavior Online

Joseph Seering
Carnegie Mellon University
Pittsburgh, PA 15232, USA
jseering@cs.cmu.edu

Geoff Kaufman
Carnegie Mellon University
Pittsburgh, PA 15232, USA
gfk@cs.cmu.edu

CHI 2018, April 21-26, Montreal, Canada.

Abstract
Harassment, spam, hate speech, and a wide variety of other destructive behaviors abound in online communities. In this paper we argue that, in order to address them, we must move beyond labeling of "Bad Actors" and instead into in-depth explorations of their motivations. This will require continued work in domains from ethnography to experimental psychology. In this paper we identify three challenges to understanding these motivations and offer three theory-based hypotheses to inform future explorations.

Author Keywords
Harassment, Hate Speech, Moderation, Governance

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Introduction
Defining and labeling "Bad Actors" in any area is a difficult and very dangerous thing to do. As various historical work has noted, these definitions have often been imposed by those in positions of power; definitions of acceptable and "normal" behaviors have been used to ostracize or punish political dissidents, non-mainstream sexual practices, and the mentally ill, among many other groups [10][11][12]. Per Blackwell et al., [4] "Classification Has Consequences".

However, the challenges inherent in this process should not be used as an excuse not to act. There are many behaviors found in online communities and social media that can be considered at minimum problematic; Danielle Keats Citron proposes a definition for unacceptable behavior that, while rooted in possibly-problematic technoliberal conceptions of the internet's role in facilitation of free speech, offers a useful baseline: any type of behavior or speech that significantly discourages others from participating in the conversation merits intervention [7]. Based on this definition, behaviors like targeted harassment of minorities and distribution of disruptive spam can be considered "bad behaviors," though classifying their perpetrators as "Bad Actors" is a more complex.

We argue here that it is extremely important to understand the motivations of people exhibiting these types of behaviors, both in order to better inform future classifications of "good" and "bad" and to guide social and technical responses. We present three challenges to doing this effectively as well as three hypotheses for the types of situations in which people are more likely to exhibit the types of behaviors associated with "Bad Actors". Finally, we situate ourselves and our research in this space.

Challenges

In this section, we highlight three challenges to investigating "Bad Actors": *methodological challenges*, *philosophical critiques*, and a broadly *technical focus* that, intentionally or unintentionally, may overlook social responses in favor of more effective automation.

Interviewing "Bad Actors" is difficult. Potential participants might reasonably be reluctant if approached explicitly to explain a suspect behavior. Moreover, perceptions of academics as very liberal or elitist may deter potential participants because they feel that their perspectives will not be treated fairly. Anti-intellectual trends growing recently in many online spaces may also contribute to this.

Jhaver, Chan, and Bruckman's recent work exploring the KotakuInAction subreddit encountered these challenges, and they note a strong initial skepticism among subreddit commenters about their motives [14]. They chose to build rapport with the community through open discussion of their methods and goals, which they report to have been a fairly successful approach. However, as they note, this approach does not guarantee that a representative sample of users will volunteer to be interviewed, nor does it guarantee that interviewees will be truthful (which, to be fair, are challenges in any interview study). The researchers take steps to verify the truthfulness of interviewees' representations of their perspectives and behavior, but qualify this by characterizing their research as an aggregation of subjective perspectives.

This leads to another challenge - *philosophical critiques* from the community. Among these critiques are the arguments that research and publication about these types of communities are problematic in that they legitimize and amplify hateful voices instead of using the power of the academic platform to amplify the voices of the vulnerable groups who are targeted. We believe that these critiques are reasonable and should be considered through the full process of any such research project. However, the avoidance of the study of social groups of major public importance because of these critiques is in itself problematic.

The final challenge we identify is a core tendency in the HCI field to value technological solutions. Of particular note is the continued development, refinement, and deployment of algorithms designed to detect hate speech, e.g., [9]. While some measure of detection at scale is probably necessary in media sites like YouTube, we argue for the importance of continued

development of technology informed by an understanding of social behaviors. In particular, we note the importance of developing technology *with users* in addition to technology *for platforms*. HeartMob [4] and Squadbox [16] are good examples of the former; both take an understanding of existing user experiences, behaviors, preferences, and needs, informed by interviews, workshops, and speculative design exercises, to develop better tools for combating harassment. Ashktorab and Vitak's work takes a similar approach through participatory design sessions with teenagers [2]. These examples stand in stark contrast to tools developed purely from researchers' intuitions or to industry specifications.

We have learned this lesson in our ongoing interview work; while we initially aimed to examine moderators' utilization of technical tools to combat negative behaviors, we were quickly informed by our interviewees that this should not be our area of focus. Existing, platform-developed moderation tools were frequently used by the moderators, but only as a baseline way to remove the most egregious misbehavior. Our interviewees have educated us about a wide variety of nuanced social strategies that they (and their communities) use to deal with more complex behaviors. While we do hope that our work will in the future aid in development of better tools, our approaches both now and in the future will reflect a greater respect for the importance of social practices in regulating behaviors.

Hypotheses

We present three hypotheses for what conditions lead to "bad behaviors", and what implications these hypotheses have in terms both of prevention of harassment and hate speech and mitigation of its impact and spread. Note that there are many reasons people act badly, and none of these hypotheses is intended to exclude other explanations.

1. Theories of both personal and personal identity [8] suggest that when individuals' personal or social identities are threatened ("*ego threats*"), prejudicial attitudes and behaviors are increased. These individuals may act to diminish, slander, or even harm threatening individuals or groups in order to maintain their status. This type of scenario plays out frequently online as, for example, women join spaces traditionally dominated by men and men respond via exclusion and harassment. While efforts focused on the "we're just like you" approach have struggled, research suggests two psychologically driven remedies: first, *perspective-taking* involves helping people understand what it's like to be someone else via exercises designed to help them imagine others' perspectives [3]. Second, developing *intergroup identities* [13] can help groups work together better than imposing a singular identity; when groups are allowed to maintain their identities but also are shown their separate roles in achieving a shared goal, they are much more tolerant of each other.

2. Recent CSCW literature, mixing backgrounds of information cascades [5] and social learning [1], has identified strong patterns of imitation in negative behaviors; exposure to a negative or unproductive comment makes commenters more likely to write such a comment [6]. Exposure to spam in a chatroom makes users who might never have spammed at all much more likely to do so, especially if the example behavior was performed by a user with status or authority [17]. Per this work, mobs of "Bad Actors" may be the result of a small number of authoritative users demonstrating bad behaviors in a very visible way. This matches our qualitative understanding of, for example the impact of the visibility of high-profile white supremacists on Twitter on the activity of their followers. Two ques-

tions arise from this - how do we identify and deal with these core “Bad Actors” who drive a substantial amount of bad behavior? Also, what can be done to minimize their influence if they cannot or should not be removed entirely? The above work suggests that visibility of high-influence, well-behaved users can mitigate the spread of bad behaviors, as can thorough cleanup of examples of bad behaviors. The idea, for example, that Twitch chat at esports events will always be toxic and therefore there’s no point in trying to clean it up is actually self-reinforcing; Twitch chat may be that way in part *because* organizations put minimal effort into moderating it.

3. Finally, over the course of our recent interviews with moderators, we heard much about one particular perceived category of rule violators - the group that simply didn’t know how to behave properly in the space. Moderators that this category of users included those who had come from a very different community with very different expectations (here 4chan was frequently referenced), and didn’t yet understand that it was not okay to behave that way in this new space. This results from the over-generalization of social norms from other platforms or channels such as the idea that the internet as a whole is a rude, caustic space where one has no right to a “safe space” and must develop a thick skin. Clear posting of rules has been found to be effective in minimizing this confusion [15], but much work remains to be done in teaching users that different spaces have different norms for behavior.

Authors’ Experience

The authors both have a research background considering social behaviors in virtual contexts. The first author has published on social behaviors on Twitch through the lens of applying social psychology to facilitate the development of more positive communities. In his current work he has interviewed more than 50 moderators across several platforms to understand the social practices they use in their moderation. He is also working on a quantitative analysis of how users respond to hate speech.

The second author has published on the impact of fictional narratives and games on increasing perspective-taking and reducing prejudice. This work has contributed to the design and release of two commercial tabletop games shown to reduce social biases and produced a set of accompanying strategies and best practices for the embedding of persuasive content and the utilization of psychological theory in the design of interventions for social impact.

The first author notes that he has only rarely experienced harassment or hate speech tied to his identity, but that he is committed in his work to amplifying the voices of those for whom they are a constant problem. Many of his colleagues, peers, friends, and family are constantly targeted by an exhausting amount of hate, and he learns from them daily what it is like to live constantly under attack.

The second author’s experiences with (offline) harassment and hate speech during his youth and adolescence directly inspired his choice to select social psychology as his field of study and continues to drive his involvement in a variety of outreach efforts to mitigate unconscious bias and counteract its impact on identity and well-being.

REFERENCES

1. R. L. Akers. 1977. *Deviant behavior: A social learning approach*. Wadsworth Pub. Co.
2. Z. Ashktorab and J. Vitak. 2016. Designing cyberbullying mitigation and prevention solutions

through participatory design with teenagers. In *Proc. CHI 2016*. ACM Press.

3. C. D. Batson, S. Early, and G. Salvarani. 1997. Perspective taking: Imagining how another feels versus imaging how you would feel. *Personality and social psychology bulletin* 23, 7 (1997).
4. L. Blackwell, J. Dimond, S. Schoenebeck, and C. Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. In *Proceedings of the ACM on Human-Computer Interaction*. ACM Press.
5. J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. 2014. Can cascades be predicted?. In *Proc. WWW 2014*.
6. J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. *CSCW '17* (2017). DOI: http://dx.doi.org/10.1145/2998181.2998213
7. D.K Citron. 2014. *Hate Crimes in Cyberspace*. Harvard University Press.
8. J. Crocker, L. L. Thompson, K. M. McGraw, and C. Ingerman. 2018. Downward comparison, prejudice, and evaluations of others: Effects of self-esteem and threat. *Journal of personality and social psychology* 52, 5 (2018).
9. T. Davidson, D. Warmsley, M. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proc. ICWSM 2017*. AAAI Press.
10. M. Foucault. 1961. *Madness and Civilization*. Vintage Books.
11. M. Foucault. 1975. *Discipline and Punish*. Vintage Books.
12. M. Foucault. 1984. *The History of Sexuality, Vol. 1: An Introduction*. Vintage Books.
13. M. A. Hogg and D. E. Van Knippenberg, D. andRast. 2012. Intergroup leadership in organizations: Leading across group and organizational boundaries. *Academy of Management Review* 37, 2 (2012).
14. S. Jhaver, L. Chan, and A. Bruckman. 2018. The border between controversial speech and harassment on Kotaku in Action. *First Monday* 23, 2 (2018). DOI: http://dx.doi.org/10.5210/fm.v23i2.8232
15. S. Kiesler, A. Kittur, R. Kraut, and P. Resnick. 2012. *Building successful online communities: Evidence-based social design*. MIT Press, Chapter Regulating behavior in online communities.
16. K. Mahar, A. X. Zhang, and D. Karger. 2018. Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation. In *Proc. CHI 2018*. ACM Press.
17. J. Seering, R. E. Kraut, and L. Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. ACM Press. DOI:http://dx.doi.org/10.1145/2998181.2998277