Does the Punishment fit the “crime”? Online harassment policies and the case of self-harm

Jessica Pater
Georgia Institute of Technology
Atlanta, GA 30308
pater@gatech.edu

Casey Fiesler
University of Colorado, Boulder
Boulder, Colorado 80309
Casey.Fiesler@colorado.edu

Abstract

All harassment is not equal. What happens when expressions of mental health issues are classified as harassment? Can individuals technically harasses themselves? In this paper, we discuss the case of self-harm, how platforms view these behaviors via formal and informal policy, and the potential punishment for breaking these policies by posting content related to their mental health. We conclude with a discussion about the ethical challenges that this design tension raises for designers and regulators of online communities.

Author Keywords

Harassment, online community, social media, self-harm, policy.

ACM Classification Keywords

K.4 Computers and Society

Introduction

What happens when you are accused of and punished for harassment of individual that ends up being yourself in an online community? What does it even mean to harass oneself online? This is the situation that adversely impacts a group of individuals that use social media platforms like Reddit, Twitter, Instagram, and Tumblr to express aspects of mental health attributes associated with self-harm. The manner in which we currently classify "bad actors" or deviance as it relates to online harassment actually could have serious negative implications for this vulnerable population.

Currently, social media platforms lump together harassing activities focused on others (bullying, hate speech, threats, stalking, abuse, racism) with activities that they deem harassment to oneself (eating disorders, self-harm, self-mutilation, and self-injury) into a general "harassment" category [6]. Because of this governance structure, people that use hate speech can be charged with and have the same sanctions and consequences as those who post about their mental health state. Is this ethical? Should designers of online communities take into consideration these type of nuances when developing policies and punishments related to mental health activities and behaviors?

Who are these bad actors?

Self-harm is a term that is used to describe certain behaviors associated with individuals who cause pain or injury to oneself [10] and most notably include cutting and wrist slashing [9] and eating disorders [5]. While there is no research on prevalence rates of self-harm activity on specific platforms or across multiple platforms, we do have prevalence rates for the US. In 2014, the World Health Organization found that 20% of 15 year-olds surveyed reported having self-harmed within the last 12 months [2].

Research into the types of media created and/or shared related to self-harm includes thinspiration, the self-harm "journey", diet, cutting, suicidal ideation, and other co-morbidities [7]. Individuals share content through centralized and more formal channels like sub-Reddits and Facebook groups [11] or through decentralized channels like the use of self-harm hashtags and variants across different platforms [1,7].

It should be noted that the potential harm is not *just* to themselves, but by making this type of content more visible to the general public it has the potential to negatively impact the community as a whole. The actual visibility of this content to the larger community is unknown, thus the negative impacts of this content are largely speculative.

Current Policy and Sanctions

Harassment policies are not solely outlined in formal community documentation like Terms of Service (TOS), Privacy Policy, or Acceptable Use Policy (AUP). They are also prevalent in informal documentation like safety guides, community guidelines and guides for specific types of users like parents, teen/youth, and law enforcement [6]. Pater et al. recently characterized the different behaviors or activities that were classified as "harassment" across many social media platforms. Table 1 highlights a subset of platforms analyzed and the behaviors associated with harassment policies.

The sanctions that are associated with these policies also varied in the severity: everything from the mild of restricting accounts and sending warning to users to the most severe of deleting accounts and working with law enforcement. While this makes sense if someone is stalking, sending threats, or bullying other members in the community, does this make sense when the intent behind the post is not directed externally? It could be

	Abuse	Attack	Bullying	Defamation	Eating Dis.	Harm	Hate	Impersonate	Intimidate	Libelous	Racist	Self-harm	Self-injury	Self-mutilation	Stalking	Threats	Torture	Vulgarity
Facebook		X	X		X		X						X	X		X		
Twitter	X	X																
Instagram	X		X	X	X		X	X	X				X		X	X		
Tumblr																		
Pintrest		X	X		X	X	X					X				X		
Flickr	X			X		X	X	X	X	X	X				X	X	X	X

Table 1. Terms associated with harassment within policy documents [Fiesler/Pater]

argued that these punishments do not fit the “crime” that has been committed against oneself.

Ethical Considerations

There are other considerations that should be taken into consideration in this discussion. Through the act of criminalizing the expressions of self-harm on these platforms it is possible that these policies actually causing secondary or indirect harm to those individuals through the addition of stress associated with the sanctions of these. As designers of these tools and the policies that drive their use, what are our obligations to these vulnerable communities?

Author’s Connection to the Field

We have extensive connections to the fields of online policies surrounding content creation [3,4], behavioral health presentations in online communities [1,7,8], and online community policies as it relates to harassment of self-harm [6]. Additionally, both Pater and Fiesler have written about and organized the HCI community to think about ethical considerations and the evolution of our current shared norms surrounding best practices as it relates to online community research.

Conclusion

Our previous research highlights the tension between needing to protect the general community from harassment and the potential risk of these policies when operationalized to address activities deemed harassment towards oneself. Harassment policies should not be a one-size-fits all endeavor. Yet, in the age of online communities with over a billion members, nuanced policy is difficult to scale. As we design the next generation of online platforms and as we conduct research on these platforms, we should be reminded that labels have power. Labeling community members with mental health issues as deviants or bad actors could have serious implications in not just our formal policies, but also the informal policies / community norms that govern most of our day-to-day online interactions.

References

1. Stevie Chancellor, Jessica Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thygapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*, ACM, 1201–1213.

2. Candace Currie, Cara Zanotti, Anthony Morgan, et al. 2012. Social determinants of health and well-being among young people. *Health Policy for Children and Adolescents* 6.

3. Casey Fiesler, Jessica Feuston, and Amy S Bruckman. 2015. Understanding Copyright Law in Online Creative Communities. *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*.

4. Casey Fiesler, Cliff Lampe, and Amy S. Bruckman. 2016. Reality and Perception of Copyright Terms of Service for Online Content Creation. *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*.

5. E David Klonsky, Thomas F Oltmanns, and Eric Turkheimer. 2003. Deliberate self-harm in a nonclinical population: prevalence and psychological correlates. *The American Journal of Psychiatry* 160, 8: 1501–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12900314

6. Jessica A. Pater, Moon Kim, Elizabeth D. Mynatt, and Casey Fiesler. 2016. Governing Online Harassment: Characterizing Policies Across Social Media Platforms. *Proceedings of the ACM GROUP Conference*, ACM, 369–374. http://doi.org/10.1145/2957276.2957297

7. Jessica Pater, Oliver Haimson, Nazanin Andalibi, and Elizabeth D Mynatt. 2016. "Hunger Hurts but Starving Works:" Characterzing the Presentation of Eating Disorders Online. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ACM, 1185–1200.

8. Jessica Pater and Elizabeth D Myantt. 2017. Defining Digital Self-Harm. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ACM Press, 1501–1513.

9. Caron Zlotnick, Jill I Mattia, and Mark Zimmerman. 1999. Clinical Correlates of Self-Mutilation. *The Journal of Nervous and Mental Disease* 5: 296–301.

10. Self-Harm. *SANE Australia*, 2016. Retrieved from https://www.sane.org/mental-health-and-illness/facts-and-guides/self-harm

11. 2018. Self Harm. *Reddit*.