Child and Adolescent Mental Health

Child and Adolescent Mental Health 29, No. 1, 2024, pp. 104–106 doi:10.1111/camh.12689

Debate: Social media content moderation may do more harm than good for youth mental health

Cindy C. Zhang[1,2], Grayden Zaleski[1,2], Jaya N. Kailley[1,2], Katelyn A. Teng[1,2], Mahala English[1,2], Anna Riminchan[1,2] & Julie M. Robillard[1,2]

[1]*Division of Neurology, Department of Medicine, The University of British Columbia, Vancouver, BC, Canada*
[2]*British Columbia Children's and Women's Hospital, Vancouver, BC, Canada*

Most social media platforms censor and moderate content related to mental illness to protect users from harm, though this may be at the expense of potential positive outcomes for youth mental health. Current evidence does not offer strong support for the relationship between censoring mental health content and preventing harm. In fact, existing moderation strategies can perpetuate negative consequences for mental health by creating isolated and polarized communities where at-risk youth remain exposed to harmful content, such as pro-eating disorder communities that use lexical variants to evade censorship. Social media censorship of content related to mental illness can also silence positive discourse about mental health, create barriers to accessing online support and resources, and hinder research efforts on youth well-being. Social media content about mental health can have important positive impacts on youth mental health by facilitating help-seeking, depicting positive coping strategies, and promoting a sense of belonging for struggling youth, but these benefits are minimized under existing moderation and censorship practices. This article presents a call to action for evidence-based social media policies and for practitioners to consider the clinical implications of social media engagement when connecting with young patients.

Key Practitioner Message

- Current social media moderation policies for mental health content are ineffective at preventing harm in youth.
- In fact, they can perpetuate negative outcomes and create barriers for accessing online support.
- This article presents a call to action for evidence-based social media policies that engage youth mental health practitioners in their development.
- Practitioners should consider the clinical implications of social media engagement when connecting with youth patients.

Keywords: Mental health; Social policy; Adolescence; Internet usage; Anorexia nervosa

Social media has become ubiquitous in the lives of young people today. Meanwhile, rates of mental health challenges among children and adolescents have soared to the extent of being declared a national emergency (American Academy of Pediatrics, 2021). Social media platforms are increasingly moderating mental health content to reduce user harm, but moderation policies should not come at the expense of silencing positive online dialogue about mental health. A close examination of censorship and moderation of mental health content on social media is urgently needed to ensure that current policies work to benefit youth mental health.

There are concerns that posting about mental illness and self-harm on social media will perpetuate negative outcomes—for instance, by putting users at risk of cyberbullying, being triggered, or imitating harmful behavior. Young people on Instagram who struggle with self-harm have reported wanting to self-injure when seeing others post pictures of their self-harm, and those who post their own self-harm have reported facing harassment from users (Brown, Fischer, Goldwich, & Plener, 2020). However, mental health dialogue on social media can also increase support and facilitate connections between youth with a shared experience—these same youth have also reported offering and being offered help and feeling a sense of belonging (Brown et al., 2020). Are social media moderation policies effective at balancing these benefits and harms? Although current strategies for moderating mental health content may be well-intended, evidence shows they may be dangerous to youth by failing to reduce harmful content while discouraging positive discourse.

Moderation strategies vary across social media platforms, often working to limit the reach of posts related to mental illness. Words such as 'depression', 'suicide', and 'pro-ana' (pro-anorexia) are banned as hashtags across many platforms that are popular among adolescents. On some platforms such as Instagram, users are presented with warnings and self-help resources when they search for these words. Another strategy employed by some platforms is to artificially limit the visibility of certain types of content by preventing personalization

algorithms from recommending it. Instagram removes self-harm images and posts promoting suicide, and other posts, such as healed self-harm scars, are hidden from search results and the Explore page but remain on the platform (Mosseri, 2019).

While this moderation may appear to be successful in reducing harmful content, the evidence does not offer strong support for this practice. In fact, censoring is increasingly shown to work against its intended purposes by creating isolated and polarized communities. Research by Chancellor, Pater, Clear, Gilbert, and De Choudhury (2016) found that online pro-eating disorder communities have developed lexical variants of moderated words to evade censorship, such as by adding or deleting letters (e.g., 'anorexia'), or using euphemisms (e.g., 'unalive' instead of 'suicide'). These tactics lead at-risk communities to remain exposed to harmful content. Findings further suggest that while communities that use lexical variants to avoid censorship of pro-anorexia tags were smaller than those using the banned tags, users were more engaged and active, and posts contained more triggering and self-harm content (Chancellor et al., 2016). In addition, despite blanket bans on Instagram, content promoting disordered eating is still suggested and visible to users that already engage in that community on the platform (Gerrard, 2018).

Moderation practices on social media should be evidence-based. However, censorship can present a circular barrier to the study of this very phenomenon and its implications. Referring to mental health issues using lexical variants creates variation in how this content is presented online, making it challenging for researchers to search for and observe online communities that engage in this content using both traditional methods of search and machine-learning-based methods.

In a quest to remove harmful content, censoring words associated with mental health can also limit helpful content. For instance, consistent with the Papageno Effect, engaging with posts depicting positive coping strategies for depression on Twitter can provide psychosocial benefits (Yuan, Saha, Keller, Isometsä, & Aledavood, 2023). Yet these benefits remain unrealized so long as posts about depression are censored. The lack of positive discourse may also contribute to stigmatizing attitudes toward mental illness and prevent youth from accessing support and resources online.

Although the goals of censorship are rooted in a desire to get youth the help they need, it does not appear to be an effective solution. Current social media censorship policies are concerning because they silence important conversations about mental health and fail to eliminate exposure to harmful content. Rather than banning words linked to mental health, social media platforms should emphasize mental health discussions in alignment with research showing that having honest conversations about mental health on social media can benefit rather than harm users (Yuan et al., 2023). Platforms could also consider a community moderation approach, which has already proven useful on platforms such as Reddit and Twitch to reduce online toxicity (Cullen & Kairam, 2022). In such systems, standout community members make moderation decisions with added context and nuance provided by an intimate understanding of the community and the content they are moderating. Additionally, content labeling has become more widely used to combat misinformation and could possibly be applied to mental health content to provide context to users (Morrow, Swire-Thompson, Polny, Kopec, & Wihbey, 2022). Youth mental health practitioners should be engaged in the development of these policies to ensure that content related to mental health on social media works to the benefit of youth that consume it.

It is also important for practitioners to consider the clinical implications of content moderation on social media. The role of social media in youth's lives cannot be understated, and their engagement with mental health content is an important social factor in their mental wellness. Understanding moderation policies, and what content youth may be engaging with, can help practitioners connect with their patients and provide relevant clinical care and advice. Transparency from platforms about their policies will allow practitioners to engage in social media safety education with patients and families. Discussions about maintaining privacy, identifying harmful content, and navigating interactions on social media are important to empower youth to engage with mental health content in a productive and positive manner.

Acknowledgements

This work was funded by the BC Children's Hospital Foundation. All contributors are listed as authors.

Conflict of interest statement

The authors have declared that they have no competing or potential conflicts of interest.

Ethics statement

This work did not require ethics approval.

Correspondence

Julie M. Robillard, British Columbia Children's and Women's Hospital, B404-4480 Oak Street, Vancouver, BC V6H 3N1, Canada; Email: jrobilla@mail.ubc.ca

References

American Academy of Pediatrics. (2021). *AAP-AACAP-CHA Declaration of a national emergency in child and adolescent mental health*. Available from: https://www.aap.org/en/advocacy/child-and-adolescent-healthy-mental-development/aap-aacap-cha-declaration-of-a-national-emergency-in-child-and-adolescent-mental-health/ [last accessed 9 August 2023].

Brown, R.C., Fischer, T., Goldwich, D.A., & Plener, P.L. (2020). "I just finally wanted to belong somewhere"—Qualitative analysis of experiences with posting pictures of self injury on Instagram. *Frontiers in Psychiatry*, *11*, 274.

Chancellor, S., Pater, J., Clear, T., Gilbert, E., & De Choudhury, M. (2016). #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings for the 19th ACM conference on computer-supported cooperative work & social computing* (pp. 1201–1213). https://doi.org/10.1145/2818048.2819963

Cullen, A.L.L., & Kairam, S.R. (2022). Practicing moderation: Community moderation as reflective practice. *Proceedings of the ACM on Human-Computer Interaction*, *6*, 1–32.

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

Morrow, G., Swire-Thompson, B., Polny, J.M., Kopec, M., & Wihbey, J.P. (2022). The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, *73*, 1365–1386.

Mosseri, A. (2019). *Changes we're making to do more to support and protect the most vulnerable people who use Instagram.* Instagram Blog. Available from: https://about.instagram.com/blog/announcements/supporting-and-protecting-vulnerable-people-on-instagram [last accessed 17 April 2023].

Yuan, Y., Saha, K., Keller, B., Isometsä, E.T., & Aledavood, T. (2023). Mental health coping stories on social media: A causal-inference study of Papageno effect. *Proceedings of the ACM Web Conference, 2023*, 2677–2685.

Accepted for publication: 14 November 2023