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

MENTAL HEALTH & THE DIGITAL CARE ASSEMBLAGE: USER & MODERATOR PRACTICES

Anthony McCosker
Swinburne University of Technology

Jane Farmer
Swinburne University of Technology

Peter Kamstra
University of Melbourne

Introduction and background

This paper examines the socio-technical ecosystems that shape the moderation of mental health content. To explore how care is formulated across and between different actors and automated systems, I focus on the experiences of moderators and users of three peer-based mental health support platforms. The analysis is framed by the notion of the 'digital care assemblage' to delineate the interactions between goal-oriented moderation policies, automated systems, human content moderators or platform managers, and users seeking or giving help in relation to mental ill-health. Each of these actors contribute to the supportive capacity of the platforms for addressing mental health issues in the community.

Early uses of the internet saw the potential to revolutionise healthcare through increased participation and community support (McCosker, 2018). The push toward peer-to-peer, participatory or personalised modes of digital health has only intensified in the rush toward online services during Covid-19 (Sorkin et al., 2021). Meanwhile, as mental health support groups have proliferated across commercial platforms and non-profit forums to fill some of these gaps, the need for moderation to ensure safety and care has deepened.

While mental health content often falls into the category of 'problematic', 'borderline' and ambiguous content for commercial platforms, it poses different challenges to other moderated material such as misinformation, 'obscenity', or hate speech (Gillespie and

Suggested Citation (APA): McCosker, A., Farmer, J., Kamstra, P. (2023, October). Mental Health and the Digital Care Assemblage: User and Moderator Practices. Paper presented at AoIR2023: The 24th Annual Conference of the Association of Internet Researchers. Philadelphia, PA, USA: AoIR. Retrieved from http://spir.aoir.org.

Aufderheide, 2020). Gerrard, for instance, raises the question of whether Instagram should allow currently banned images of healed self-cutting, under the guise of supporting recovery (2022, p. 86-87). To curtail these 'risks', platforms increasingly rely on algorithmic moderation of various kinds (Gorwa et al., 2018), alongside teams of human moderators.

This paper contributes to the growing body of work on content moderation as a socio-technical system, and to the goal of improving moderation practices for better mental health support. Examining three dedicated mental health platforms, we present the digital care assemblage as a concept for understanding – and optimising – digital mental health care and support as an assemblage of actors, practices, governance mechanisms and technical components (Fox, 2011; LaMarre and Rice, 2021).

Methods

The three project partners deliver community-oriented, non-profit mental health support programs and services, funded through government grants and charity donations – SANE Australia, Reach Out and Beyond Blue. This paper focuses on qualitative research with forum moderators and managers (n=7), lived experience volunteer moderators or 'community guides' (n=4), and people living with mental health issues who make use of the platforms for support (n=35). Data was collected throughout 2021 and 2022, through online focus group workshops with moderators and managers, and video interviews with users.

Findings and Analysis

Participants spoke at length about how human and automated moderation, community guidelines and user actions interacted to create a safe online space for mental health support – and where tensions arise within what we refer to as the overall 'digital care assemblage'. That is, we found moderation practices consisted of interactions between institutional goals, community guidelines and negotiated policies, through a set of routines and feedback loops, and in conjunction with automated moderation systems. We characterise these processes as both *reactive* and *adaptive* – they were dynamic, but circumscribed by a distribution of routines, actions and often contested decisions (McCosker et al., 2023).

Platform managers and moderators described the mix of broad goals for creating supportive, peer-led spaces in contrast to the 'cluster-mess' (as one platform manager put it) of commercial platforms like Facebook. They saw their professional role as juggling a set of service goals and policies, working with automated moderation systems, through collaborative and reflexive moderation practices, to manage unpredictable user actions and behaviours.

In some respects, the automated moderation systems contribute to structuring the routine tasks moderators must attend to (Jhaver et al., 2019), but the influence moves in both directions with collaborative and distributed forms of decision-making in responding to problematic content and interactions. And this included the actions and responses of platform users. For example, on one platform, users responded to the Covid-19 pandemic and lockdowns through extensive 'venting' and engaging in off-topic 'social-posts', which breach community guidelines designed to focus on mental health challenges more directly. Moderators and managers eventually adapted the guidelines to allow for these uses of the platform.

Interactions between moderation and user practices are complex and cause some tension and are integral to the way the care assemblage is constructed and maintained. While our interviews with platform members did not initially focus on their experience of moderation, it was a theme that often arose during discussions – usually when reflecting on things that did not work well for them. Their experiences, however, contribute to the functioning of the care assemblage.

When posts are flagged by moderators and changed, deleted, or given warnings, some users feel reluctant to continue posting, or modulate how they post: ‘I’ve become less willing to put certain things out here’ (SANE, 1). In this case, anonymity was a positive factor, enabling their sharing and openness. But a warning from a moderator about potential legal issues related to one post made them less willing to share at all afterwards. Variations on this experience have been understood as forms of ‘self-censorship’ in negotiations with platform moderation and algorithmic curation (Gillespie, 2018). We argue that the negotiations were more than self-censorship; they were as intricately involved in the work of the care assemblage as the moderation practices and systems they ran up against.

Managing safety and care is experienced by some as a form of co-production. For example, one person reflected on initially being upset when she had a post removed, but she came to understand and appreciate the moderation processes: ‘now I’m fine with it, like I know they need to do that and it’s good they do’; ‘they email me the post and they let me know why it’s been moderated. And usually, they are just changing a word or two.’ She elaborated: ‘Like you don’t mean to write something that could be upsetting but sometimes you just don’t know.’ (Beyond Blue, 2).

In a different way, recalling a time when a distressed and suicidal member disappeared from the forum, one user emphasised the need for more transparency, and hence involvement in the care assemblage. They noted feeling upset at the time and their belief that the person had ‘ended her life’:

> it’s a bit of a fine line. […] I think it would be helpful that if someone’s been banned … there should be some kind of notification … just for the peace of mind for the people that have been communicating with this person and genuinely care about them. (SANE, 15)

Conclusion

Improving moderation practices and systems to better support mental health requires firstly a more complete understanding of the digital care assemblage, and secondly, negotiation between moderators and managers, automated systems designers as well as users. Bringing users’ experiences into moderation goals and practices can add an important, underutilised dimension. This would align with the professionalisation and growth in dedicated platforms for community-based mental health support outside of the less equipped commercial social platforms.

References

De Cotta, T., Knox, J., Farmer, J., White, C., & Davis, H. (2021). Community co-produced mental health initiatives in rural Australia: A scoping review. *Australian Journal of Rural Health*, 29(6), 865-878.

Fox, N. J. (2011). The ill-health assemblage: Beyond the body-with-organs. *Health Sociology Review*, 20(4), 359-371.

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

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T., and Aufderheide, P. (2020). Introduction: Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, 9(4), 1–29.

Girrard, Y. (2022). What is content moderation? In D. Rosen (Ed.). *The Social Media Debate: Unpacking the Social, Psychological, and Cultural Effects of Social Media*. Routledge, 77-95.

Jhaver S, Birman I, Gilbert E, et al. (2019) Human-machine collaboration for content regulation: The case of Reddit automoderator. *ACM Transactions on Computer-Human Interaction* (TOCHI), 26(5), 1-35.

Kang, Y. B., McCosker, A., Kamstra, P., & Farmer, J. (2022). Resilience in web-based mental health communities: Building a resilience dictionary with semiautomatic text analysis. *JMIR formative research*, 6(9), e39013.

LaMarre, A., & Rice, C. (2021). The eating disorder recovery assemblage: collectively generating possibilities for eating disorder recovery. *Feminism & Psychology*, 31(2), 231-251.

McCosker, A., Kamstra, P., & Farmer, J. (2023). Moderating mental health: Addressing the human–machine alignment problem through an adaptive logic of care. *New Media & Society*, 14614448231186800.

McCosker, A. (2018). Engaging mental health online: Insights from beyondblue's forum influencers. *New Media & Society*, 20(12), 4748-4764.

McCosker, A., & Gerrard, Y. (2021). Hashtagging depression on Instagram: Towards a more inclusive mental health research methodology. *New Media & Society*, 23(7), 1899-1919.

Sorkin DH, Janio EA, Eikey EV, et al. (2021) Rise in use of digital mental health tools and technologies in the United States during the COVID-19 pandemic: survey study. *Journal of Medical Internet Research*, 23(4), e26994.