Algorithmic folk theories and peer review: on the importance of valuing participant expertise (commentary)

Dr Carolina Are,

Innovation Fellow at the Centre for Digital Citizens, Northumbria University

carolina.are@northumbria.ac.uk

Abstract

Through this brief commentary piece I discuss the challenges and opportunities of submitting academic work on hidden forms of content moderation such as shadowbanning and malicious flagging, and the difficulty of balancing dated notions of academic rigour with investigating an issue that is connected with an absence of data from powerful stakeholders. In doing so, I address how peer reviews can inadvertently reinforce the inequalities of content moderation, aiding platform companies in the discrediting, victim-blaming and gaslighting of their users by replicating unequal and patriarchal behaviours adopted by various authorities when victims come forward to report violence and injustice.

Keywords:

'This is a valid paper, but the methodology is flawed,' read one of the reviewer comments on one of my most recent rejected manuscripts, a paper on the impact of malicious or misused flagging on social media content by marginalised users. Yet, as I continued reading, eager to take in reviewers' comments to improve my methods

for future submissions, they went on to say my methods would be better if I spoke to platform workers enforcing flags or to those who misuse them – demographics to which, incidentally, I had no access. In short, my paper would be better… if it was another paper entirely; if I chose not to believe participants. This experience, and previous experiences of peer review on studies that focus on invisible content moderation practices such as malicious flagging – or the practice of reporting content to a social media platform not because it violates community guidelines, but because users disagree with it and/or wish to de-platform it as a form of harassment (Are, ; Crawford & Gillespie, 2016) – and shadowbanning – or platforms' decision to hide from or avoid recommending content to their main feeds (Are, 2021) – have led me to reflect on the challenges of peer review. As a result, through this brief commentary piece I discuss the challenges and opportunities of submitting academic work on hidden forms of content moderation such as shadowbanning and malicious flagging, and the difficulty of balancing dated notions of academic rigour with investigating an issue that is connected with an absence of data from powerful stakeholders. In doing so, I wish to address how peer reviews can inadvertently reinforce the inequalities of content moderation, aiding platform companies in the discrediting, victim-blaming and gaslighting of their users by replicating unequal and patriarchal behaviours adopted by various authorities when victims come forward to report violence and injustice (Are & Gerrard, 2023).

Peer review is a fundamental component of academic scholarship, based on experts evaluating the standard of scientific work in their field to encourage rigour and impartiality (Lee et al., 2013). Yet, when studying platform governance, or the 'questions pertaining to the implications and impact of platform features, functions and rules' and 'the international regulatory dynamics that currently delineate the

freedoms, responsibilities and liabilities of platform companies' (Tiidenberg, 2021:2), we are faced with a conundrum: how can we ensure researchers provide tangible proof of processes and behaviours that platforms do not notify users of? How can we evaluate their findings on what are known as algorithmic black boxes (Cotter, 2021) when platforms actively restrict access to that specific information (Hatmaker, 2021)?

Platform governance studies therefore become an example of how ineffective peer review can be in situations where knowledge communities regulate themselves. Lee et al. found that factors like author nationality and prestige of institutional affiliation, reviewer nationality, gender, and discipline, and reviewer agreement with submission hypotheses (Lee et al., 2013) influence paper acceptance or rejection. Further, in the old struggle between quantitative and qualitative methods, there a tendency to believe that methods grounded in experience such as autoethnography and ethnography, and by extension methods that largely rely on participants' re-telling of their experiences on platforms, are not rigorous or valid (Poole, 2022).

In the face of inscrutable platform governance and lacking transparency from major Big Tech conglomerates, researchers and activists alike have found that 'folk theories' (Eslami et al., 2016) about algorithms and 'algorithmic gossip' among users (Bishop, 2019) about social media companies' capricious content sorting can aid both professional content creators and the public in understanding and unveiling black boxes (Cotter, 2021). These stories outlining governance quirks and the specific experiences of creators are often the only type of information we - as academics and members of civil society - have about platform processes (Are, 2021). The same stories have been found to have brought platforms to admit moderation practices such as shadowbanning were taking place (ibid; Leybold & Nadegger, 2023), and they are beginning to be used to investigate the impact of

malicious flagging (Silverman & Fortis, 2023). Yet, in my experience both as a researcher and a censored social media user (Are, 2021), these stories are often discredited by platforms (Cotter, 2021) and peer review alike, mirroring societal disdain against gossip – an important safeguarding tool that women and marginalised communities often use to protect themselves and thrive under patriarchal, structurally oppressive systems (Bishop, 2019).

We already know that platform governance and content moderation replicate offline inequalities by disproportionately affecting marginalised users such as sex workers, LGBTQIA+, disabled and/or BIPOC accounts, those posting nudity and sexuality and profiles from the Global South (Are, 2021; Paasonen et al., 2019; Haimson et al., 2021 etc.). We know that the demographics of Big Tech's workforce can replicate offline power structures (Tomaskovic-Devey and Han, 2018) by treating different elements of expression as deviant (Are, 2022; Are and Paasonen, 2021), and that social media companies tend to restrict content they deem unsavoury in accordance with their financial and reputational interests (e.g. Paasonen et al., 2019; Tiidenberg & van der Nagel, 2020). In short, we know platforms' design and their governance reproduce the matrix of domination (white supremacy, heteropatriarchy, capitalism, and settler colonialism) (Costanza-Chock, 2018; 2020).

We do not need peer review and academic research to beat platforms at their game.

Since social media platforms often deny that censorship techniques are being used on specific content, engaging in a form of gaslighting against women, femme presenting and marginalised users and nude bodies (Are & Gerrard, 2023; Diaz and Hecht-Felella, 2021), we – as scholars and peer reviewers – should instead believe and centre the most marginalised users' needs, working with them not as passive

research subjects (Costanza-Chock, 2018; 2020), but recognising them as algorithmic experts (Bishop, 2019). As I and Ysabel Gerrard (2023) have previously argued, the ills of content moderation are structural and systemic, and the fact that women and marginalised users' experiences are denied and belittled by platform governance replicate the victim-blaming and gaslighting these users face offline when trying to report abuse or injustice (ibid).

Believing participants is more than a gateway to untapped research findings: it is a crucial act of radical questioning of societal and platform power structures by centring the stories, the gossip of those who have been wronged.

This act of change must come with the acknowledgement that we ourselves, as researchers and peer reviewers, occupy a gate-keeping role in determining who is an algorithmic professional (Bishop, 2019) and who deserves to be believed when discussing governance inequalities.

Of course, we cannot take users' experiences and beliefs at face value, and trusting our participants should not mean avoiding to research with rigour, or failing to conduct thorough investigations on the issues at hand and to look for opposing experiences to the ones we are trusting. But if we truly wish to hold Big Tech to account, we need to make room for methodologies that differ from the ones we are familiar with, and we need to believe participants to find the chinks in platform power's armour through a more constructive peer review process that does not believe tech workers are the only algorithmic experts. In this sense, research rigour and believing participants blends with peer review as a skill, making Nygaard's (2020) remarks on peer review a valid and important concluding statement:

'Few of us are taught how to review someone else's work: we are expected to be able to give relevant feedback on someone else's research simply on the basis of the research we have produced ourselves. But if we want to encourage greater diversity in the knowledge we create as scholars, and reduce social inequalities, perhaps we should treat peer reviewing as a separate skill that needs to be learned.'

Bibliography

Are, C. (2021). The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, *22*(8): 2002-2019.

Are, C. and Gerrard, Y. (2023). Violence and the feminist potential of content moderation. In: K. Boyle and S. Berridge (Eds.), *Routledge Companion on Gender, Media and Violence*. Routledge.

Costanza-Chock, S. (2018). Design justice: Towards an intersectional feminist framework for design theory and practice. *Design Research Society Conference 2018. University of Limerick*. https://doi.org/10.21606/drs.2018.679.

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. Cambridge: MIT Press.

Cotter, K (2021). 'Shadowbanning is not a thing': black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society, 26*(6):1226-1243.

Crawford, K and Gillespie, T. (2016). What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint. *New Media & Society*, *18*(3): 410–428.

Diaz, A. & Hecht-Felella, L. (2021). Double standards in social media content moderation. *Brennan Centre for Justice*.

https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf.

Eslami, M.; Karahalios, K.; Sandvig, C.; Vaccaro, K.; Rickman, A.; Hamilton, K.; Kirlik, A. (2016). First I "like" it, then I hide it: Folk Theories of Social Feeds. *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*: 2371–2382. https://doi.org/10.1145/2858036.2858494.

Haimson, O.; Delmonaco, D.; Nie, P. and Wegner, A. (2021). Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction*,5, (CSCW2), No.: 466. 1–35, https://doi.org/10.1145/3479610.

Hatmaker, T. (2021). 'Facebook cuts off NYU researcher access, prompting rebuke from lawmakers.' *TechCrunch*. https://techcrunch.com/2021/08/04/facebook-ad-observatory-nyu-researchers/.

Lee, C.J., Sugimoto, C.R., Zhang, G. and Cronin, B. (2013). Bias in peer review. *Journal for the American Society of Information Science and Technology*, *64*: 2-17. https://doi.org/10.1002/asi.22784

Leybold, M., & Nadegger, M. (2023). Overcoming communicative separation for stigma reconstruction: How pole dancers fight content moderation on Instagram. *Organization*, *0*(0). https://doi.org/10.1177/13505084221145635

Nygaard, L. P. (2020). Turned Away at the Gate: How Peer Review Can Reinforce Social Inequalities. *Peace Research Institute Oslo (PRIO)*.

https://blogs.prio.org/2020/09/turned-away-at-the-gate-how-peer-review-can-reinforce-social-inequalities/.

Paasonen, S., Jarrett, K. and Light, B. (2019). *#NSFW: Sex, Humor, And Risk In Social Media*. The MIT Press.

Poole, A. (2022). 'Methodologies don't hurt people, bad people wielding methodologies do: Autoethnography and *that* paper from *Qualitative Research*.' *British Educational Research Association.* https://www.bera.ac.uk/blog/methodologies-dont-hurt-people-bad-people-wielding-methodologies-do-autoethnography-and-that-paper-from-qualitative-research.

Silverman, C. & Fortis, B. (2023). A Scammer Who Tricks Instagram Into Banning Influencers Has Never Been Identified. We May Have Found Him. *ProPublica*. Available at: https://www.propublica.org/article/instagram-fraudster-ban-influencer-accounts.

Tiideberg, K. (2021). Sex, power and platform governance. *Porn Studies Forum*, *8*(4), p.381-393.

Tiidenberg, K. and van der Nagel, E. (2020). *Sex and Social Media*. Melbourne: Emerald Publishing.

Tomaskovic-Devey, D. and Han, J. (2018). Is Silicon Valley Tech Diversity Possible Now? *UMass-Amherst, Center for Employment Equity*, https://www.umass.edu/employmentequity/sites/default/files/CEE_Diversity+in+Silicon+Valley+Tech.pdf.