
Article:
Gerrard, Y. (2020) Social media content moderation : six opportunities for feminist intervention. Feminist Media Studies, 20 (5). pp. 748-751. ISSN 1468-0777

https://doi.org/10.1080/14680777.2020.1783807

Social media content moderation: Six opportunities for feminist intervention

Words: 1071

Social media content moderation – making and enforcing rules about the content that can (and cannot) be posted to a social media platform – is anything but neutral. While this form of governance has long lied *in the shadows* (Roberts, 2019:222), public concerns about content moderation have recently exploded. Contrary to early excitement about its participatory potentials, social media is no longer viewed as an empty shelf to fill as we like. Social media companies make value-laden, largely opaque decisions about what we are allowed to put on that shelf. As Roberts (2018:n.p.) notes, content moderation's logic of opacity is a form of 'depoliticization': it allows platforms to disavow evidence that women – mainly women of colour – face social media's harshest rules. For example, Twitter knowingly permits a disproportionate level of abuse to Black women (Dreyfuss, 2018), Instagram previously restricted hashtags related to women of colour, like #MixedGirls and #MexicanGirls (Drewe, 2016), and Tumblr no longer allows images of 'female-presenting nipples' (Paasonen et al, 2019).

These examples highlight moments when social media content is removed or hidden from view, but actions like removal are the ends points of human-algorithmic systems. This short essay explains why we get to those end points, outlining six stages of the content moderation process that are most fallible to human intervention (and therefore bias, subjectivity, intolerance). Somewhat optimistically, it also explains how these ideology-laden spaces might also be *opportunities* for feminist intervention. In short, they are the spaces we need to target if we want to enact change in the system.

1. Content moderation policies: As Gillespie explains, content moderation policies for some of the world's most popular social media platforms are written by workers who are 'overwhelmingly white, overwhelmingly male, overwhelmingly educated, overwhelmingly liberal or libertarian, and overwhelmingly technical in skill and worldview' (2018:12). This means decisions about what counts as 'problematic' content are not wholly attuned to the needs of social media's diverse userbase. Rule-setting reflects the worldview of rule-makers, but feminist scholarship is uniquely positioned to lay bare the biases of content moderation policies.

2. Public-facing community guidelines: Most social media platforms have a set of community guidelines: public-facing documents that lay out, in 'deliberately plainspoken language' (Gillespie, 2018:46), what content platforms do and do not allow. While platforms have long emphasised their neutrality (Gillespie, 2010), community guidelines undo this careful discursive work by revealing biases, politics and normativities. This means they can actually help us to scrutinise a company's corporate ethos. In short, community guidelines tell us everything we need to know about a social media company's values.

3. 'Flagging' (or, social media's language of complaint): Social media companies rely on users to 'flag' posts to send them for human review (Crawford and Gillespie, 2016). Users are given limited options to explain why they think a post should be removed from a platform, but one person's reason for complaint might differ from another's. Ahmed (2019:n.p.) explains that complaint means *committing* 'yourself, your time, your energy' to something. But the problem with flagging is that users are entirely removed from the process that occurs after the tick-box complaint is complete. In fact, they might not even learn the outcome (Crawford and Gillespie, 2016). Plenty of content is also wrongly taken down, restarting the cycle of complaint. This is a problem. We need to demand more transparent channels for complaints, and to challenge and help to prevent wrongly-imposed takedowns.

4. Human content moderation: Until fairly recently, 'Commercial Content Moderator' (CCM) was not a job title many people were familiar with, but we now know that humans do most of the dirty work of keeping platforms clean. CCMs use tightly-guarded rulebooks to make decisions (Hopkins, 2017), making them an urgent subject of concern for feminists. We know these rules are a problem. For example, a *ProPublica* investigation found that Facebook used to train its censors to 'delete hate speech against "protected categories," including white males, but to allow attacks on "subsets" such as female drivers and Black children' (Angwin and Grassegger, 2017:n.p.). The problem is that we do not have access to up-to-date moderator handbooks across different companies, and it is incredibly difficult to get hold of them. How do the rulebooks differ between social media companies? Who decides what

goes into them? How can we influence those decisions? Why are the rulebooks so opaque in the first place? These questions need feminists' urgent attention.

5. Automated content moderation: Some low-level content moderation can be done automatically and without direct human intervention (Gerrard, 2018). But a reliance on automated moderation will likely never – and nor should it be – fully realised. Automation it a 'blunt tool' (Mozilla Insights, 2020:n.p.): it makes mistakes, misses context and nuance, and in many cases might be used unethically. Researchers like Eubanks (2018) and Noble (2018) have explored the dangers of other automated systems, like search engines and police profiling tools. Such systems are often criticised for relying on harmful stereotyping along the lines of race, gender, sexual orientation and other identity markers. How can we continue to leverage feminist scholarship and activism to create change in automated content moderation systems? And how can we avoid an increasing reliance on 'flawed' (Mozilla Insights, 2020:n.p.) content filtering technologies?

6. In-platform content restrictions: Social media companies often rely on quick fixes instead of overhauling a whole content moderation policy. For example, restricting search results for certain hashtags (Chancellor et al, 2016) or shadowbanning users (Myers-West, 2018). Some of these fixes can be helpful, but because they are often opaque and inconsistently applied (Suzor, 2016) – especially shadowbanning (see Joseph, 2019) – they allow social media companies to deepen inequalities in a way that evades public critique. What sort of transparency should we be demanding here, and how do we get it?

This short essay has outlined the processes of social media content moderation that are perhaps most vulnerable to human intervention. It reminds readers that content moderation is anything but a neutral process, and argues towards greater transparency and oversight of decisions that affect a good chunk of world's population. This essay has offered a blueprint for enacting real change in content moderation systems, highlighting six places where humans are at their most influential, both from within and outside of tech organisations. I am not saying this is an easy task, and nor am I demanding

perfection in moderation. But this is a starting point to push for fairer, more inclusive and significantly more transparent content moderation systems.

References

Ahmed, S. (2019, July 22). Why complain? *Feministkilljoys.* [Blog post]. Available at: https://feministkilljoys.com/2019/07/22/why-complain/.

Angwin, J. and Grassegger, H. (2017, June 28). Facebook's secret censorship rules protect white men from hate speech but not Black children. *ProPublica.* [Online]. Available at: https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

Chancellor, S., Pater, J.A., Clear, T., et al. (2016) #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In: Proceedings of the 19th ACM conference on computer–supported cooperative work & social computing, CSCW '16. Available at: http://www.munmund.net/pubs/cscw16_thyghgapp.pdf.

Crawford, K. and Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media and Society*. *18*(3): 410-428.

Drewe, N. (2016, 10 May). The hilarious list of hashtags Instagram won't let you search. *The Data Pack*. Available at: http://thedatapack.com/banned-instagram-hashtags-update/.

Dreyfuss, E. (2018, December 10). Twitter is indeed toxic for women, Amnesty report says. *WIRED*. Available at: https://www.wired.com/story/amnesty-report-twitter-abuse-women/.

Eubanks, V. (2018). *Automating inequality: how high-tech tools profile, police, and punish the poor*. New York, NY: St. Martin's Press,

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

Gillespie, T. (2010). The politics of 'platforms'. *New Media and Society. 12*(3): 347-364.

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

Hopkins, N. (2017, May 21). Revealed: Facebook's internal rulebook on sex, terrorism and violence. *The Guardian*. Available at: https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence.

Joseph, C. (2019, November 8). Instagram's murky 'shadow bans' just serve to censor marginalised communities. *The Guardian*. Available at: https://www.theguardian.com/commentisfree/2019/nov/08/instagram-shadow-bans-marginalised-communities-queer-plus-sized-bodies-sexually-suggestive.

Mozilla Insights. (2020, May 4). *When content moderation hurts*. [Online]. Available at: https://foundation.mozilla.org/en/blog/when-content-moderation-hurts/.

Myers-West, S. (2018). Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. *New Media and Society. 20*(11): 4366-4383.

Noble, S.U. (2018). *Algorithms of oppression: how search engines reinforce racism*. New York: NYU Press.

Paasonen, S., Jarrett, K and Light, B. (2019). *NSFW: sex, humor, and risk in social media*. Cambridge, MA: MIT Press.

Roberts, S.T. (2017, March 8). Social media's silent filter. *The Atlantic*. Available at: https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/.

Roberts, S.T. (2018). Digital detritus: 'error' and the logic of opacity in social media content moderation. *First Monday*. *23*(3). Available at: https://doi.org/10.5210/fm.v23i3.8283.

Roberts, S.T. (2019). *Behind the screen: content moderation in the shadows of social media*. New Haven: Yale University Press.

Suzor N (2016, September 17). How does Instagram censor hashtags? *Medium*. Available at: https://digitalsocialcontract.net/how-does-instagram-censor-hashtags-c7f38872d1fd.