BIG DATA & SOCIETY

Commentary

Big Data & Society
July–December: 1–5
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2053951720943234
journals.sagepub.com/home/bds
SAGE

# Content moderation, AI, and the question of scale

Tarleton Gillespie

## Abstract

AI seems like the perfect response to the growing challenges of content moderation on social media platforms: the immense scale of the data, the relentlessness of the violations, and the need for human judgments without wanting humans to have to make them. The push toward automated content moderation is often justified as a necessary response to the scale: the enormity of social media platforms like Facebook and YouTube stands as the reason why AI approaches are desirable, even inevitable. But even if we could effectively automate content moderation, it is not clear that we should.

## Keywords

Artificial intelligence, bias, content moderation, platforms, scale, social media

This article is a part of special theme on The Turn to AI. To see a full list of all articles in this special theme, please click here: https://journals.sagepub.com/page/bds/collections/theturntoai

> Chris Anderson: “How many people do you have working on content moderation to look at [white nationalist activity on Twitter]?”
>
> Jack Dorsey: “It varies. We want to be flexible on this, because we want to make sure that we’re, number one, building algorithms instead of just hiring massive amounts of people, because we need to make sure that this is scalable, and there are no amount of people that can actually scale this. So this is why we’ve done so much work around proactive detection of abuse that humans can then review. We want to have a situation where algorithms are constantly scouring every single tweet and bringing the most interesting ones to the top so that humans can bring their judgment to whether we should take action or not....”[1] (2019)

## The promise of “the promise of AI”

As social media platforms have grown, so has the problem of moderating them—posing both a logistics challenge and a public relations one. Twenty years ago, online communities needed to answer only to their own users after incidents of harassment or trolling (Dibbell, 1993; Kiesler et al, 2011). For the now dominant social media platforms, community-scale techniques have become increasingly untenable and unconvincing (Gillespie, 2018). The quantity, velocity, and variety of content is stratospheric; users are linked less by the bonds of community and more by recommendation algorithms and social graphs; the consequences of online harms now extend beyond the platform on which they occur; and criticism of these platforms and their failures to moderate has exploded, catalyzed by Gamergate, Myanmar, revenge porn, the 2016 U.S. presidential election, Alex Jones, and Christchurch.

In response to public scrutiny, the CEOs of these platform companies needed to promise moderation techniques to match the enormity of these problems. Not surprisingly, many offered up AI as the solution.

Microsoft Research New England, Cambridge, MA, USA

**Corresponding author:**
Tarleton Gillespie, Microsoft Research New England, 1 Memorial Drive, Cambridge, MA 02142, USA.
Email: tarleton@microsoft.com

While a few claimed that AI would solve the problem entirely, and was just around the corner, others more modestly suggested that automation would relieve human moderators of some of their burden. But all the major platforms dream of software that can identify hate speech, porn, or threats more quickly and more fairly than human reviewers, before the offending content is ever seen. This idea, and the celebration of AI more broadly, is a product of a "mindset prevalent in Silicon Valley, which sees these problems as technological ones requiring technological solutions" (Geiger, 2016: 791). Such promises were not just aimed at users and lawmakers; they are also for their investors, who worry that platforms are teetering on the edge of toxicity.

In March and April of 2020, Twitter and many of the major social media platforms were forced to shift almost entirely to automated moderation, after sending most of their human moderators home in response to the coronavirus pandemic (Magalhães and Katzenbach, 2020). Unlike the outsized promises of just a year before, many platforms apologized for the uptick in errors they anticipated:

> We want to be clear: while we work to ensure our systems are consistent, they can sometimes lack the context that our teams bring, and this may result in us making mistakes...We appreciate your patience as we work to get it right – this is a necessary step to scale our work to protect the conversation on Twitter.[2]

As I write this, it is too soon to know how this abrupt shift has affected moderation practices or the experience of users. And it is far too soon to know whether the old arrangements will ever return, or whether the coronavirus has forever shifted how platform moderation works. It is either a vital moment, or a few months too late, to note that while it may be an appealing idea, automated content moderation is not a panacea for the ills of social media, and may be contrary to the principles of governance that platforms should be pursuing.

## Scale is not the same thing as size

The dream of automating content moderation is often justified as a response to scale. The immense amount of the data, the relentlessness of the violations, the need to make judgments without demanding that human moderators make them, all become the reason why AI approaches sound so desirable, inevitable, unavoidable to platform managers. Lawmakers threatening to impose stricter obligations on platform moderation also point to their scale as justification. And in response, platform spokespeople double down—warning that only AI can be fast enough, precise enough, and sensitive enough to meet any heightened obligations. This link between platforms, moderation, and AI is quickly becoming self-fulfilling: platforms have reached a scale where only AI solutions seem viable; AI solutions allow platforms to grow further.

Questioning this received wisdom may require questioning the very idea of scale itself. While others have much more nuanced takes on "scale" than I do, the first necessary observation is not hard: scale is not the same thing as size. Too often, when platform representatives point to their scale, they mean little more than the enormous number of users or amount of content. But scale is something more than size. Scale is about how the small can be made to have large effects; or how a process can be proceduralized such that it can be replicated in different contexts, and appear the same. We do not talk about parenting happening at scale, even though there are an enormous number of people currently parenting. But when Instagram had millions of users but just 14 employees, that is more than size, that is scale. Scale is a specific kind of "articulation," in Jennifer Slack's (2006) understanding of the term: different components attached, so they are bound together but can operate as one—like two parts of the arm connected by an elbow that can now "articulate" their motion together in powerful but specific ways.

Content moderation on social media platforms often involves large and small components articulated into a single, functioning apparatus: small policy teams overseeing large populations of human moderators; short lists of guidelines fitted with large lists of procedures and exceptions; enormous populations of users attached to flagging mechanisms that produce tiny bits of data about many, many violations. AI, too, is a specific articulation of elements at different sizes: it turns an enormous amount of training data into a simple calculus that can then act on an enormous amount of content. Building on the work of Tsing, Seaver (2016) calls these "varieties of sociotechnical scalemaking."

The claim that moderation at scale requires AI is a discursive justification for putting certain specific articulations into place—like hiring more human moderators, so as to produce training data, so as to later replace those moderators with AI. In the same breath, other approaches are dispensed with, as are any deeper interrogations of the capitalist, "growth at all costs" imperative that fuels these massive platforms in the first place.

## The pitfalls of automating moderation

As some platforms seemed to admit when they moved toward automation during the coronavirus pandemic, the current precision of these tools has been embellished. First, some very different moderation tools tend to get lumped under the umbrella of "AI." While some platforms are using machine learning techniques to identify new instances of harassment, hate speech, or pornography, most are doing a sophisticated version of pattern matching: merely comparing new content to a blacklist of already known examples (Gorwa et al., 2020). This is automation, but it is hardly AI, except under the broadest possible definition.

This means that recent claims by platforms of successful automated moderation are overstated. To claim that "65.4% of [hate speech] content actioned was found and flagged by Facebook before users reported it"[3] sounds impressive. But, at least for now, the overwhelming majority of what is being automatically identified are copies of content that have already been reviewed by a human moderator. Stats like these are deliberately misleading, implying that machine learning (ML) techniques are accurately spotting new instances of abhorrent content, not just variants of old ones.

Then there are deeper problems with automating moderation, many of which resonate with familiar concerns about AI and data science more broadly, and that animate worries about automated policing, data-driven insurance assessments, hiring software, and automated medical diagnostics (Ajunwa, 2020; Benjamin, 2019; Brayne, 2017; Eubanks, 2018). First, ML tools must be trained on existing, labeled data. Platforms have plenty, in the corpus of moderation decisions they have made in the past. But this data is an artifact of the policies and judgments of the platform's existing moderation arrangements. An effective tool may learn to make the same kinds of distinctions as before (Binns et al., 2017; Gehl et al., 2017). But while consistency might sound like a good thing, these policies should actually adapt over time (Sinnreich, 2018). ML tools also have difficulty accounting for context, subtlety, sarcasm, and subcultural meaning (Duarte et al., 2017). Even the tools designed to identify duplicates may be insensitive to the use of the same content in a different context, like terrorist propaganda reposted in a journalistic context (Llanso, 2019).

Perhaps, automated tools are best used to identify the bulk of the cases, leaving the less obvious or more controversial identifications to human reviewers. Finding duplicates is certainly one of those bulk tasks, which is why platforms already use these tools to do exactly that. But beyond duplicates, it is not clear how to know in advance which areas are safely bulk and which are more controversial, as the landscape of controversy changes over time. Nor is it clear whether the certainty of the identification software aligns with the obviousness of the violation.

Further, however well implemented, there is a fundamental contradiction in using data-centric techniques for content moderation. Machine learning techniques shift our understanding of societal phenomena: from instances among collectives, to patterns among populations. As Ananny (2019) noted, "Probabilistic ideas – about chance, likelihood, normalcy, deviance, confidence, thresholds – underpin many of the sociotechnical infrastructures and institutions that regulate online speech platforms." Statistical accuracy often lays the burden of error on underserved, disenfranchised, and minority groups. The margin of error typically lands on the marginal: who these tools over-identify, or fail to protect, is rarely random (Buolamwini and Gebru, 2018).

And, there will always be mismatch between an immense moderation apparatus and the affective experience of being on the receiving end of it (West, 2018). Moderation may be felt as an injustice—that the platform failed to understand the legitimacy of my post, failed to protect my right to speak. It may be a sense of absurdity, as when automated systems are wildly off the mark (Matsakis, 2018). It may be a feeling of suspicion, that these inscrutable methods only hide naked corporate self-interest. It may be a feeling of insignificance, when stilted, bureaucratic warnings offer a stark reminder that these megamachines (Hill, 2019) have little concern for their users as individuals. This gap between data-scale approaches and the individual experiences of them is slowly undermining the legitimacy of content moderation itself—a risk amplified by a move to pandemic conditions, where the human moderator is removed entirely.

## Maybe we should not automate

Even if we could effectively automate content moderation, it is not clear that we should. Perhaps, the kind of judgment that includes the power to ban a user from a platform should only be made by humans. Penalizing someone for violating the rules is in fact one of the ways we as communities and societies discover, test, and reassert our shared values. There is no stable, widely shared value system that simply must be implemented. Calling something hate speech is not an act of classification, that is either accurate or mistaken. It is a social and performative assertion that something should be treated as hate speech, and by implication, about what hate speech is. And, undoubtedly, it will be disagreed with. Competing, even incommensurate values should not be papered over, nor should they be our undoing (Mouffe, 2016). A society is a society in part because it articulates and grapples with its purported values; societies hold together not by reaching perfect consensus, but by keeping their values under constant and legitimate reconsideration. "The public good cannot be fixed in advance because the public itself is always in a process of reconstitution. Our debates about what is good for us are always, in part, debates about whom we want to be" (Calhoun, 1998: 33). Debating the hard cases, publicly, asserting what is unacceptable and why, is essential and without end.

None of this is to suggest that automated tools should play no role in content moderation. While they wait for their AI to improve, social media companies have hired thousands of human moderators to fill the gap. As Roberts (2019) has documented, these moderators are doing psychologically scarring work, in sometimes intolerable conditions, often under precarious labor arrangements. In fact, the strongest argument for the automation of content moderation may be that, given the human costs, there is simply no other ethical way to do it, even if it is done poorly.

It is good that platform CEOs are beginning to acknowledge that AI should not entirely replace human judgment, even as the coronavirus pandemic forced their hand. And it is good that platforms use automated tools to spot duplicates. I think there are two other ways we might think about the possible partnership between human and automated moderation. First, ML tools could be designed and optimized for identifying the most egregious, scarring content, in order to protect human moderators from having to view it at all, and letting them focus on content that is not so damaging to encounter. Users should accept less precision in these areas, so that human moderators might not have to endure the beheadings, violent pornography, and child abuse. Second, we might think about designing ML tools to support human teams rather than supplant them. For instance, by providing more contextual data about specific violations (what is the likelihood that they will do harm or cause offense?) human moderators might be better sensitized to the variety of norms across subcommunities.

Still, every time we try to "solve" this problem, we pass up the opportunity to ask a much more profound question. It is not just that moderating an enormous platform is harder. The persistent failure of social media platforms to build moderation architectures that work should tell us something. Perhaps, if moderation is so overwhelming at this scale, it should be understood as a limiting factor on the "growth at all costs" mentality. Maybe some platforms are simply too big. This is the hardest question to raise, of course, in a capitalist context. But it is one that should be raised, especially alongside growing calls for antitrust efforts against Silicon Valley (Hughes, 2018; Wu, 2018). The head of Instagram has argued that separating Instagram from Facebook would not alleviate their moderation problems, because Instagram depends on Facebook's enormous content moderation apparatus.[4] This may be true. Or it may be that, only then, would Instagram find itself able to pursue more innovative, artisanal approaches.

It is size, not scale, that makes automation seem necessary. Size can be changed. But it may require a profound act of countercapitalist imagination to find new ways to fit new ML techniques with new forms of governance—not as a replacement for repetitive human judgments, but to enhance the intelligence of the moderators, and the user communities, that must undertake this task.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was conducted with the financial support of Microsoft Research. It was not directed by, nor does it express the opinions of, Microsoft.

## ORCID iD

Tarleton Gillespie https://orcid.org/0000-0002-2601-6073

## Notes

1. "How Twitter needs to change," conversation with Jack Dorsey, Chris Anderson, Whitney Pennington Rodgers, *TED*, April 2019, https://www.ted.com/talks/jack_dorsey_how_twitter_needs_to_change/transcript.
2. Vijaya Gadde and Matt Derella, "An update on our continuity strategy during COVID-19." *Twitter blog*, 16 March 2020. Available at: https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html (accessed August 3 2020).
3. Facebook "Community Standards Enforcement Report, January-March 2019," https://transparency.facebook.com/community-standards-enforcement#hate-speech.
4. "Instagram's Adam Mosseri, Facebook's Andrew Bosworth, and former tech insiders on the outside (Live at Code 2019)." *Recode Decode*, 24 June 2019, https://www.vox.com/recode/2019/6/10/18660514/adam-mosseri-andrew-bosworth-casey-newton-facebook-instagram.

## References

Ajunwa I (2020) The paradox of automation as anti-bias intervention. *Cardozo Law Review* 41: 54.

Ananny M (2019) Probably speech, maybe free: Toward a probabilistic understanding of online expression and platform governance. *Knight First Amendment Institute*, 21 August.

Benjamin R (2019) *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity.

Binns R, Veale M, Van Kleek M, et al. (2017) Like trainer, like bot? Inheritance of bias in algorithmic content moderation. *Social Informatics* 2017: 405–415.

Brayne S (2017) Big Data surveillance: The case of policing. *American Sociological Review* 82(5): 977–1008.

Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research* 81: 1–15.

Calhoun C (1998) The public good as a social and cultural project. In: Powell W and Clemens E (eds) *Private Action and the Public Good*. New Haven: Yale University Press, pp.20–35.

Dibbell J (1993) A rape in cyberspace, or how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *The Village Voice*, 23 December.

Duarte N, Llanso E and Loup A (2017) Mixed messages? The limits of automated social media content analysis. Center for Democracy and Technology.

Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Gehl R, Moyer-Horner L and Yeo SK (2017) Training computers to see internet pornography: Gender and sexual discrimination in computer vision science. *Television & New Media* 18(6): 529–547.

Geiger SR (2016) Bot-based collective blocklists in twitter: The counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19(6): 787–803.

Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press.

Gorwa R, Binns R and Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 3(1): 1–15.

Hill S (2019) Empire and the megamachine: Comparing two controversies over social media content. *Internet Policy Review* 8(1): 1–18.

Hughes C (2018) It's time to break up Facebook. *New York Times*, 9 May. Available at: https://www.nytimes.com/2019/05/09/opinion/sunday/chris-hughes-facebook-zuckerberg.html (accessed August 3 2020).

Kiesler S, Kraut R, Resnick P, et al. (2011) Regulating behavior in online communities. In: Kraut RE and Resnick P (eds) *Building Successful Online Communities: Evidence-Based Social Design*. Cambridge: MIT Press, pp.77–124.

Llanso E (2019) Platforms want centralized censorship. That should scare you. *Wired*, 18 April. Available at: https://www.wired.com/story/tumblr-porn-ai-adult-content/ (accessed 3 August 2020).

Magalhães JC and Katzenbach C (2020) Coronavirus and the frailness of platform governance. *Internet Policy Review*. Available at: https://policyreview.info/articles/news/coronavirus-and-frailness-platform-governance/1458 (accessed 3 August 2020).

Matsakis L (2018) Tumblr's porn-detecting AI has one job—And it's bad at it. *Wired*, 5 December. Available at: https://www.wired.com/story/tumblr-porn-ai-adult-content/ (accessed August 3 2020).

Mouffe C (2016) Democratic politics and conflict: An agonistic approach. *Política Común* 9: 17–29.

Roberts ST (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press.

Seaver N (2016) *Care and Scale: Size and Sensibility in Music Discovery*. Minneapolis: American Anthropological Association.

Sinnreich A (2018) Four crises in algorithmic governance. *Annual Review of Law and Ethics* 26: 181–190.

Slack JD (2006) Communication as articulation. In: Shepherd G, St. John J and Striphas T (eds) *Communication as . . . : Perspectives on Theory*. Thousand Oaks: SAGE Publications, pp.223–231.

West SM (2018) Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20(11): 4366–4383.

Wu T (2018) *The Curse of Bigness: Antitrust in the New Gilded Age*. New York: Columbia Global Reports.