Content Moderation

Zeerak Talat

For the past decade, the planetary scale of online platforms has resulted in more content created by humans than could possibly be subjected to manual human verification. To address this issue, researchers and content platforms have sought automated means for content moderation, often powered by machine learning technologies, to augment human content moderation efforts (Roberts 2019). While commercial content moderation, as defined by Sarah T. Roberts (2019), primarily considers the human labor in content moderation, contemporary practices often construct content moderation as a shared effort between humans and machines that seeks to determine the aesthetics of acceptability for a public, mediated society. For instance, Meta Platforms, Inc.'s Facebook platform reports that large quantities of content are "actioned" using machine learning, or artificial intelligence (AI) (Facebook n.d.a, n.d.b). The question of how machine learning affects the power dynamics within content moderation has remained underexplored in the content moderation and machine learning literature. This entry provides an overview of content moderation with an emphasis on machine learning-based approaches. It discusses how content moderation infrastructures impact Stuart Hall's (1973) encoder-decoder model of communication, how machine learning distributes power (Kalluri 2020), and how it impacts the construction of acceptability in society, and thereby the everyday aesthetics of mediated spaces online. The entry examines these questions and concludes that automating content moderation relies on the exploitation of poor, marginalized, and systematically excluded communities and individuals in their development and use.

Although content moderation has recently come to be at the center of popular discourse around the Internet, content has been moderated continuously throughout Internet history. For instance, UseNet lists, an early example of Internet mailing lists, had active moderation to ensure that topics discussed and content shared remained useful to list members and did not contain harmful software. However, as access to content and the ability to create content have shifted from scales that allow human content moderation (e.g. by members of a community) to a planetary scale, with the associated increase in the amount of content created, content moderation has shifted toward being practiced by external parties (Gillespie 2020; Roberts 2019) and is increasingly reliant on automated methods (Facebook n.d.a). Contemporary content moderation practices allow us to understand commercial social media platforms as mediators of content that determine which aesthetics are deemed "acceptable" and which are externalized. By connecting content moderation to Douglas's (1978) work on the cultural politics of dirt and sanitization, Lepawsky (2019) argues that moderation constitutes a sanitization effort. By viewing content moderation as a sanitization effort, we can understand it simultaneously as seeking to reinforce the community through negative removals of harmful content and as "a more fundamental and positive 're-ordering of our environment' through practices of classification and purification" (Thylstrup and Talat 2020). Indeed, content moderation, whether human or automated, involves processes that influence and reinforce cultural artifacts, thereby providing a cultural aesthetic "filter" (Lepawsky 2019) through which experiences of Internet platforms are mediated. The content that is to pass through that filter—i.e. all content that is created—far exceeds the capabilities of human reviewers, particularly given the rise of AI-generated content (Ables 2023). Machine learning-based content moderation systems have therefore been integrated into content moderation infrastructures, which are often applied in tandem with human content moderators. That is, automated systems may automatically remove content without human

review, as on the Facebook platform (Facebook n.d.a), or they may be used to highlight content for human review (Gillespie 2020). The result of such efforts is that content moderation infrastructures create two potential designations for each bit of content: that which is permissible, and that which *may* be sanctionable. In this way, content moderation comes to wield an influence on the everyday aesthetics (Saito 2017) of speech and culture.

Considering communication within the context of television programming and audiences, Stuart Hall (1973) proposed the encoder-decoder framework. Hall sought a theory that encompassed the complexities of the relationship between broadcaster, message, and viewer. Hall posited that broadcasters, i.e. encoders, embedded meaning within communicative messages subject to the political and material conditions within which the broadcaster existed; the audience, i.e. decoders, were subject to the material conditions within which they existed, in light of which they decoded the messages from a dominant, oppositional, or negotiated position. Information would then flow from decoders to encoders, closing the feedback loop between encoder and decoder. Seeking to account for the role of content moderation infrastructures, Nanna Bonde Thylstrup and I have elsewhere introduced a third party into Hall's framework: the mediating toxic content moderation infrastructures (Thylstrup and Talat 2020). We argue that platforms play a mediating role as a third party that is subject to coalitions between encoders and decoders. Within the scope of platforms that host content created and generated by humans and automatic methods respectively, private individuals and (commercial) organizations occupy the roles of both encoder and decoder. Although we have introduced the notion of the mediator, the issues of the mediator's role, its responsibilities, and how it engages in the political ecosystem of online platforms remain underexplored. To explore the role of the mediator, we must first consider the claim that content moderation is a form a censorship.

To understand content moderation simply as censorship relies on a very narrow view of both content moderation and censorship. For instance, in addition to removal, content moderation practices include reordering content by ranking and hiding it from view, a practice also known as shadow banning (Gillespie 2022). Indeed, Thylstrup and I have argued that content moderation infrastructures can also serve as methods for positively restructuring content within communities, thereby fortifying the sense of cohesion within those communities (Thylstrup and Talat 2020). If one regards content moderation not simply as the removal of content, but as content that is removed, hidden, or rearranged, this becomes more readily apparent. It is through such practices that we can come to understand how content moderation influences the cultural aesthetics of the permissible, desirable, and unwanted. For instance, content moderation heavily moderates the expression of women's bodies (Gerrard and Thornham 2020), creating a narrow space in which they can exist. At one extreme, the display of nipples *not belonging* to bodies coded as male is prohibited and moderated. At the other extreme, content with fat bodies is often ranked lower or moderated more heavily (Williams 2021). In both instances, content moderation practices create a narrow cultural aesthetic within which female-coded bodies are permissible on the Internet, namely, as thin and deemed to be desirable. However, even within this narrow scope, the aesthetics of how to present such bodies is coded. For instance, content about sex workers, and workers in adjacent areas such as exotic dancers, is often moderated. The aesthetics to which such moderation practices thus aspire are those that police women's bodies (Gerrard and Thornham 2020). Such practices make clear that the mediator takes an active role in developing and maintaining online ecosystems. However, it is unclear for whom this work is being undertaken, and while such maintenance work *could* be constructed as care work, I argue that contemporary content moderation practices do not amount to care work, as the work is predominantly outsourced to algorithmic third parties and to underpaid workers in

Southeast Asia and Africa (Roberts 2019). Indeed, the algorithmic third parties themselves rely on human intervention—implicitly by providing data through interaction patterns, and explicitly through the creation and curation of training data for machine learning algorithms. Yet, while inherently reliant on human efforts, the algorithms are developed as defensive tools with which platforms insulate themselves from potential criticism.

Content moderation infrastructure as mediation thus acts on behalf of the platform, which inserts itself as a third party into Hall's encoder-decoder framework. Mediators seek to referee the relationship between encoders and decoders: on one hand by determining for encoders what content is appropriate, so that they will not be penalized; on the other, by determining on behalf of decoders what content is surfaced to them. While the role of the mediator is often publicly described as being conducted on behalf of the decoders—i.e. to ensure their "safety, dignity and authenticity" (Facebook n.d.a)— in practice the mediator wrests power from both encoders and decoders to ensure that content conforms to the mediator's perspective and position. That is, the mediator acts in concordance with a belief about what decoders *should* see, rather than what decoders *wish* to see or what encoders *aim* to communicate. The content that is viewable and easily accessible to decoders is that which has remained unsanctioned. Yet mediators must remain within bounds of what encoders and decoders deem acceptable content moderation interventions in order to avoid antagonist relationships with their communities. Such conflicts between mediators, encoders, and decoders can be expressed through the evasion of content moderation, e.g. disguising the keyword "white people" as "wypipo" or blurring body parts in order to avoid moderation (Thylstrup and Talat 2020). Such evasive actions are indicative of political failures of machine vision to distinguish between the semantics and surface form of content for moderation (i.e., text, images, videos, and audio). Considering the demographic disparities (Dias Oliva, Antonialli, and Gomes 2021) and the goal to remove "toxicity" (Wulczyn,

Thain, and Dixon 2017) in content moderation, mediators act as agents that uphold hegemony through respectability politics (Thylstrup and Talat 2020).

Many automated content moderation infrastructures rely on machine learning. Machine learning is often cast as a technology that is apolitical, whereby any discriminatory outcomes (i.e. biases) that arise are simply inherited from the data. Indeed, machine learning systems can accurately be described as machines that identify patterns in data (Bishop 2006). For content moderation, the particular type of machine learning algorithm used is classification algorithms. Classification algorithms seek to identify patterns in data and correlate them with a set of given output classes. A common approach to identifying patterns in data is the frequentist approach, i.e. the more frequently a pattern appears in the data, the more salient it is. This formulation of salience results in machine learning becoming a site of power and privilege by centering that which is normative. Contemporary machine learning algorithms, such as the Transformer architecture that underlies technologies such as ChatGPT, require large amounts of data, which are often initially published on the Internet, prior to their inclusion in machine learning datasets. By combining Internet data—which is produced primarily in the European Union and the USA (Dunn 2020)—and the frequentist approach, machine learning models come to center hegemony and normative values. That is, salient patterns identified by machine learning are inherently conservative, owing to their preference for the normative. For content moderation, the consequence of this proclivity toward a normative aesthetics of acceptability is reflected in decision-making, i.e. the overemphasis on existing norms of acceptability comes to determine the acceptability of new content, thereby imposing potentially outdated and inappropriate norms onto that new content. Such conservative tendencies thus enact an inability for machines to see the development of cultural aesthetics in favor of a static view of culture that is grounded in hegemonic renderings of the past.

Lastly, I close with a consideration of the automation of content moderation and human labor. While automated methods bring a promise of the reduction, if not the outright removal, of humans from content moderation infrastructures, humans remain central to those infrastructures at every step, in part to correct the errors–with respect to content policies–that occur due to the inability of machines to see and correctly process the cultural developments upon which machines are enacted. Indeed, the promise of automation neglects the fact that content moderation is an inherently human endeavor. First, objections to content arise primarily from humans, e.g. when humans report hate speech. In this mode, people are exposed to content that they experience as objectionable, and they report it. Second, commercial content moderators adjudicate on the content. These two steps create training data for machine learning models. Moreover, efforts are also devoted to creating additional data by identifying data samples, labeling them, and validating the labels. This process is primarily, albeit not always entirely (see e.g., Wulczyn, Thain, and Dixon 2017), performed by humans. Once a machine learning model has been trained on the created resources, human moderators are tasked with validating and correcting the model's output and predictions to ensure surface-level conformity to norms set out by the organization (see e.g., Perrigo 2023). Humans are thus tasked to address the inability of machines to appropriately see content within the cultural and social aesthetics that govern the creation of content. As such, humans moderate the machine learning models. In addition to this moderation, platforms may create additional data to account for conformity to corporate values (Solaiman and Dennison 2021). Finally, models are deployed and used. However, as machine learning optimizes for hegemony, content moderation decisions that occur at the edge of the patterns identified by the model are subject to human reports and moderation. Thus, while machine learning promises automation, it necessitates a return to human moderation work. Moreover, the politics of which voices content moderation infrastructures are developed for creates a

disproportionate burden of labor—to protest decisions and insist upon equal protections—on the people and communities that are least protected.

Automated methods for content moderation serve to erase the human work that undergirds content moderation decisions—more so than serving to automate the infrastructure and work. For instance, training machine learning models requires large, manually labeled datasets. In commercial content moderation, the training data for machine learning models is created using data labeled by human moderators. As a consequence of their probabilistic basis, as well as the frequentist assumption, machine learning models cannot be error-free, particularly within the context of hate speech, which requires an understanding of the harms of hegemony. It is therefore necessary to continuously correct the predictions of machine learning models. The promise of automating content moderation infrastructures is thus an effort to displace work: from the poorly compensated labor of commercial content moderators, onto the disregarded and targeted communities that platforms have volunteered for content moderation work. As models are continuously updated, so are the requirements for the creation of data, and it is because of these continuous processes and the continued need for human engagement that the promise will remain a dream deferred.

Through the different facets of (automated) content moderation outlined here—i.e. content moderation as a sanitization effort, mediators' or content moderation infrastructures' role in Hall's communicative framework, the conservativism of machine learning when applied to content moderation, and the way automated moderation displaces rather than replaces careful human work—we can come to understand the ramifications of the political commitments of content moderation infrastructures, human or automated. Considering these aspects, it becomes apparent that machinic content moderation infrastructures are deeply committed to hegemony and the exploitation of labor as a mechanism to account for failures

of the machine to appropriately see and recognize content. Thus, while the promise of automation could alleviate the harms faced by marginalized communities, the political commitments of content moderation infrastructures create a "double problem" whereby marginalized communities are excessively policed while also carrying the burden of addressing the failures of content moderation infrastructures (Thylstrup and Talat 2020). Rather than a promise, contemporary content moderation infrastructures, then, constitute a capitalist fever dream: seeking the comfort of high-resourced individuals at the cost of exploiting poor and marginalized individuals and communities.

References

Ables, Kelsey. 2023. "Sci-Fi Magazine Clarkesworld Flooded with AI-Generated Work." *Washington Post*, February 22, 2023. https://www.washingtonpost.com/technology/2023/02/22/scifi-magazine-clarkesworld-artificial-intelligence/.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Dias Oliva, Thiago, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. "Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online." *Sexuality & Culture* 25 (2): 700–732. https://doi.org/10.1007/s12119-020-09790-w.

Douglas, Mary. 1978. *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo*. Reprint, London: Routledge.

Dunn, Jonathan. 2020. “Mapping Languages: The Corpus of Global Language Use.” *Language Resources and Evaluation* 54 (4): 999–1018. https://doi.org/10.1007/s10579-020-09489-2.

Facebook. n.d.a. “Community Standards Enforcement.” Accessed April 30, 2021. https://transparency.facebook.com/community-standards-enforcement#hate-speech.

Facebook. n.d.b. “How Does Facebook Use Artificial Intelligence to Moderate Content?” Accessed June 5, 2023. https://www.facebook.com/help/1584908458516247.

Gerrard, Ysabel, and Helen Thornham. 2020. “Content Moderation: Social Media’s Sexist Assemblages.” *New Media & Society* 22 (7): 1266–86. https://doi.org/10.1177/1461444820912540.

Gillespie, Tarleton. 2020. “Content Moderation, AI, and the Question of Scale.” *Big Data & Society* 7 (2): 205395172094323. https://doi.org/10.1177/2053951720943234.

Gillespie, Tarleton. 2022. “Reduction/Borderline Content/Shadowbanning.” *Yale Journal of Law & Technology* 476 (24). Accessed February 20, 2024. https://yjolt.org/reduction-borderline-content-shadowbanning.

Hall, Stuart. 1973. “Encoding and Decoding in the Television Discourse.” Discussion paper. Birmingham, UK: University of Birmingham.

Kalluri, Pratyusha. 2020. “Don’t Ask If Artificial Intelligence Is Good or Fair, Ask How It Shifts Power.” *Nature* 583 (7815): 169–169. https://doi.org/10.1038/d41586-020-02003-2.

Lepawsky, Josh. 2019. “No Insides on the Outsides.” *Discard Studies*. September 23, 2019.

https://discardstudies.com/2019/09/23/no-insides-on-the-outsides/.

Perrigo, Billy. 2023. “OpenAI Used Kenyan Workers on Less Than $2 Per Hour.” *Time*, January 18, 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/.

Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press.

Saito, Yuriko. 2017. *Aesthetics of the Familiar, Volume 1*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780199672103.001.0001.

Solaiman, Irene, and Christy Dennison. 2021. “Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets.” arXiv. Last modified November 23, 2021. http://arxiv.org/abs/2106.10328.

Thylstrup, Nanna, and Zeerak Talat. 2020. “Detecting ‘Dirt’ and ‘Toxicity’: Rethinking Content Moderation as Pollution Behaviour.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3709719.

Williams, Sherri. 2021. “Watch out for the Big Girls: Black Plus-Sized Content Creators Creating Space and Amplifying Visibility in Digital Spaces.” *Feminist Media Studies* 21 (8): 1360–70. https://doi.org/10.1080/14680777.2021.2004195.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 2017. “Ex Machina: Personal Attacks Seen at Scale.” In *Proceedings of the 26th International Conference on World Wide Web*, 1391–99. Perth, Australia: International World Wide Web Conferences Steering Committee.

https://doi.org/10.1145/3038912.3052591.