The Ethical Implications of Machine-Led Content Moderation

Madelief Slaats - s1032615
Ivo Melse - s1088677
Pieter-Jan Lavaerts - s1151391
Erik Oosting - s1136456

Introduction

Content moderation is the practice of evaluating user-generated content (UGC) to determine whether it is appropriate in the context of the website or (social media) platform (Roberts, 2017). Over the last two decades, the amount of UGC has grown exponentially, and thus has the need for content moderation. One solution that has been proposed and is used by social media platforms is to automate (part of) the process of content moderation (Horta *et al.*, 2020).

This increasing reliance on automated methods, including machine learning (ML) for content moderation on social media platforms introduces a range of ethical challenges and philosophical questions. As Gomez *et al.* (2024) point out, while ML models offer an efficient means to manage the vast amount of content generated daily, they also risk making arbitrary and unexplainable decisions. These developments, much like those in other areas of AI, demand rigorous philosophical inquiry into their implications. Hernández *et al.* (2022) aptly argue that these systems are not mere technical tools but are embedded with epistemic and moral assumptions that have real-world consequences.

In this paper, we will focus on the ethical implications of using automated models to recognize and moderate UGC on social media. Specifically, we will examine how issues such as bias, accountability and societal harm arise in this context. We will try to answer the question: What are the ethical implications of delegating the recognition and moderation of user generated content on social media to automated methods?

We will begin this paper by exploring the historical evolution of content moderation, from its early volunteer-driven models to the rise of algorithmic systems in today's platform-driven internet. Next, we will analyze how machine learning-based moderation mediates societal norms, highlighting issues of bias, accountability, and the influence of platform owners' political and economic interests. We will then delve into the design of AI moderation tools, discussing strategies to embed values, ensure transparency, and foster collaboration with human moderators. Finally, we will consider the interaction between AI-generated content and AI-moderated systems, reflecting on the broader societal and ethical challenges posed by these technologies.

A brief history of content moderation

Despite what is commonly believed, content moderation was already present in some of the first internet communities in the early 90s. At the time, content moderation was mainly carried out through volunteers, and the moderation policies differed vastly between communities (Roberts, 2017). Over the last decades, the internet has grown exponentially in size and has also changed in nature. We have seen a trend of “platformization” of the internet, a term first coined by Helmond (2015) that refers to the expansion of social media platforms to encompass the most important infrastructural and economic model of the social web.

As a consequence, the internet is increasingly hosted by major US based public companies that need to carry out content moderation because they are responsible to their shareholders, yet at the same time have an economic incentive to do this as efficiently as possible. Furthermore, these companies have ample capital and technical expertise to automate content moderation. This sets the stage for the current algorithmic moderation that we will consider in this paper.

Analysis

A mediation perspective

The mediation approach focuses on the dynamic interplay between humans, technology and societal norms. It presents machine learning systems not as fully autonomous nor entirely under human control, but as mediators that actively shape, and are shaped by, cultural contexts and human values. This perspective is particularly relevant to analyzing the ethical implications of ML-driven content moderation, as it emphasizes the mutual influence between technology and its users.

As Hernández *et al.* (2022) argue, ML models inevitably embed the cultural and ethical assumptions of their creators. These assumptions can have profound consequences, particularly when they misrepresent or oversimplify complex issues like extremism. For instance, algorithms trained to flag "extremist" content may rely on datasets that predominantly represent certain cultural or linguistic norms, leading to biased enforcement. Gomez *et al.* (2024) illustrate this with examples from platforms where moderation algorithms disproportionately flagged content from non-Western languages or cultures as harmful, even when the context was benign. This overgeneralization risks silencing marginalized voices while failing to address content that is actually harmful. Addressing such harms requires ongoing refinement of the technology and deliberate inclusion of diverse perspectives during the design phase.

Chekkee (2024), a company dedicated to content moderation, highlights the limitations of ML systems in making nuanced ethical decisions, stressing the indispensable role of human judgment. For example, AI may flag satirical posts, activist content or historical references as extremist because it lacks the contextual understanding to differentiate between harmful intent and legitimate expression. This underscores the mediated relationship between humans and machines in moderation processes. Similarly, Binns *et al.* (2017) discuss how decisions made during the training of ML systems, such as the selection of training data or the prioritization of certain features, fundamentally shape their behavior. These decisions reveal the mutual reinforcement between human judgment and technological capabilities, demonstrating that AI cannot operate in isolation without risking significant ethical missteps.

ML-based moderation systems do more than enforce existing rules; they actively mediate and shape social norms by interpreting the boundaries of acceptable speech. As the Knight First Amendment Institute (2021) notes, these systems can inadvertently worsen societal polarization when their design prioritizes engagement metrics over ethical considerations. For example, algorithms optimized for user interaction often amplify sensationalist or divisive content, creating echo chambers that reinforce extremist ideologies. Conversely, the suppression of content deemed harmful can also have unintended consequences. During the COVID-19 pandemic, automated moderation systems mistakenly flagged and removed legitimate discussions about public health due to overly broad keyword filters, as reported by Gomez *et al.* (2024). These examples illustrate the double-edged nature of ML as a mediator; it both reflects and reshapes the discourse it is tasked to moderate.

From the mediation perspective, the ethical responsibility for content moderation is distributed across multiple stakeholders. Designers, platform operators and policymakers all play roles in shaping the interaction between humans and technology. As the Knight Institute (2021) emphasizes, addressing the unintended effects of algorithmic systems requires coordinated efforts. Policymakers must establish regulations that incentivize fairness and transparency, while designers and platform operators must ensure their systems are aligned with societal values. Chekkee (2024) reinforces this by arguing that the balance between free speech and harm prevention requires a nuanced understanding of how ML systems mediate societal values. Effective moderation, therefore, depends not just on refining algorithms but also on fostering collaborative and inclusive decision-making.

Content moderation, politics and accountability

In this section, we will investigate who is accountable for failing moderation on social media platforms, how the political opinions of their owners influences the content moderation, and how automated content moderation influences this accountability.

We have many examples of hate speech on social media corroborating violence or even genocide. We will consider the example of the Rohingya. The Rohingya are a Muslim ethnic group mostly located in Myanmar. For many years, they have been the victims of systemic oppression and ethnic cleansing in Myanmar (O'Brien & Hoffstaedter, 2020). A report by Amnesty international (2022) reveals that Facebook played a major role in a brutal campaign of ethnic cleansing in 2017, by providing an anti-Rohingya echo chamber that influenced military leadership. They point out that Facebook did not only fail to act, but actively profited from the situation. Activists have demanded remuneration from Meta that would go towards education of young Rohingya, but the company refused, stating that “Facebook doesn't directly engage in philanthropic activities.”.

From this example, it is clear that Meta acted immorally. From a consequentialist perspective, if Facebook had censored the hate speech against Rohingya, killings, sexual violence, and the displacement of 700.000 Rohingyas into neighboring Bangladesh might have been prevented. From a more deontological perspective, allowing hate speech against a marginalized ethnic group and allowing freedom of expression are two rules that could be in conflict here. However, we believe that the harm that is inherent in hate speech against a marginalized group outweighs the freedom of expression in this case.

We consider it clear that Meta is at least partially responsible for the ethnic cleansing. We may apply the ideas underlying tort law: As a consequence of Meta's inaction to prevent ethnic cleansing in 2017, they caused the Rohingyas harm, and therefore they should provide remuneration.

This is not an isolated case, there is a *pattern* of failing moderation by a big tech companies that leads to disastrous consequences, ranging from violence, the undermining of democracy, defamation and even possible genocide. However, since these major companies are hosted in the US, they are protected by Communications Decency Act, section 230. This law protects online services from liability regarding third party content. In practice, the interpretation of this law by US judges has largely given big tech companies *carte blanche* when it comes to most kinds of user-created content (Calo, 2024) (Johnson & Castro, 2021).

The discussion about content moderation is very much tied to political leaning, at least in the US. A recent study has found that democrats are more likely to think that false headlines should be removed than republicans (69% vs. 34%), while republicans are more likely to consider this removal censorship (49% vs. 29%) (Appel *et al.,* 2023).

It is then perhaps unsurprising that after the takeover of Twitter by the libertarian "free speech absolutist" Elon Musk, the moderation policy has been relaxed to allow many posts that would be considered hate speech by the previous ownership to proliferate. A study by Montclair State University (Benton, *et al.*, 2022) shows that hate speech on the platform immediately spiked after Musk's takeover. More recently, Meta CEO Mark Zuckerberg has announced that they would replace independent fact-checkers by "community notes", a system that is also in place on X/Twitter. A recent study has found that this system has no significant effect on user's engagement with false information (Chuai *et al.*, 2024), most likely because it is too late to intervene when the tweet has already been read by many people. These developments show that the owners of social media platforms determine the content moderation policy. This means that in practice, they are the ones who arbitrate what is *hate* speech and what is *free* speech on the internet. In response to Zuckerberg, UN rights chief Volker Türk has argued that lack of moderation leads to the silencing of marginalized voices, and that the world should call for accountability and governance in the online space (Unicef, n.d.).

We should also consider that even in a scenario where political decisions do not influence the moderation policy of social media platforms, these platforms already have an economic interest in having controversial content on their platform, since this content drives user engagement, as Ghosh (2021) points out.

We have argued that big tech companies should be accountable for their content moderation, and we have shown that the opinions of platform owners and their economic interests shape the content moderation policy. Therefore, we conclude that platform owners should be held accountable for (the failure of) content moderation of the internet. Furthermore, we consider it fundamentally undesirable in a democratic society that a few affluent individuals have such a big influence on the question of free speech vs. hate speech.

We do not believe that the use of automated methods for content moderation changes anything significant about this accountability. It is true that if automated methods are used, engineers and lower level management also influence the way that moderation is carried out. Their biases will inevitably influence content moderation, and their mistakes could have disastrous consequences. However, it should ultimately be the responsibility of the owners of the companies and the leadership to put proper oversight mechanisms and safeguards in place to prevent harm to society.

Design of AI moderators

In his paper, Van de Poel (2020) studies whether the approach of value sensitive design (VSD) is applicable to AI systems. He defines sociotechnical systems as consisting of a technology together with agents abiding by a set of rules called an *institution*. These are the norms and values of the human agents. He argues that AI systems should abide by a different set of rules from non-AI systems, called a technical norm.

Technical norms are a robust mechanism for embedding values during the design phase and they can ensure system-wide compliance with ethical principles. For instance, norms could enforce fairness by requiring that certain moderation decisions be flagged for human review. Artificial agents, because of their self-learning behavior, risk not adhering to institution rules over time. This is why it is essential to monitor their behavior and adjust norms as needed.

The paper therefore emphasizes the importance of human oversight, arguing that despite AI's autonomy, human control is critical to ensure systems remain aligned with societal values. Monitoring systems for unintended consequences, redesigning components, and ensuring explainability are vital steps for maintaining meaningful human control.

Van de Poel stresses that AI systems require continuous monitoring and redesign due to their dynamic nature and capacity for unintended consequences. Values may need to evolve as societal priorities shift or unforeseen risks emerge. To embed values successfully, designers should focus on technical norms and ensure human oversight, and in doing so, develop AI systems that are adaptive and ethically robust.

Crosset and Dupont (2022) come to the same conclusion as Van de Poel in that iterative redesign and continuous monitoring of AI systems is crucial in embedding values. They study content moderation systems as a particular instance of AI systems. They highlight that current AI systems tend to give false positives when it comes to monitoring content pertaining to sensitive societal issues, thereby emphasizing the need for these systems to cooperate with human moderators. They argue that in content moderation, human moderators should provide contextual understanding that current algorithms lack.

Rieder and Skop (2021) analyze Perspective API, a tool for moderating online discourse developed by Google. A key insight from the paper is the conflict between openness and centralization. They warn against the centralization of moderation practices, which could suppress diverse perspectives in favor of uniform standards.

Many researchers seem to agree that continuous monitoring and transparency are features to strive for in AI content moderation tools. Lai*, et al.* (2022) have conducted research into the specificities of such interactive systems. They have developed novel ways in which human content moderators can work together with one or several LLM moderation tools. They theorize that content moderation tools that output a degree of toxicity instead of a binary decision are beneficial, a claim also made by the researchers investigating Perspective API. They argue that a non-binary output can help deploy the same AI tool in different settings. In a setting where content is particularly sensitive, only the most clear-cut cases of harmful content are handled completely autonomously. Content with some ambiguity is delegated to human moderators. A non-binary output also helps to clarify and justify the decision making, by helping us understand what the system deems a clear case of harmful content and in which cases the system detects nuance.

Earlier approaches teach AI systems desired behavior by feeding them labeled data in the form of a binary decision. In this novel approach, an AI moderator was fed data labeled with reasoning as to why the data was harmful, by highlighting harmful regions of text. As such, the system was taught combinations of words that create harmful contexts together, and was taught how to explain content moderation decisions as such creating transparency into its decision making.

Crosset and Dupont (2022) argue that we should strive for automated systems which enable collaboration between humans and tools, and transparency. This is because automated systems are deemed unfit to understand the nuances of content in context by themselves. Molina and Sundar (2022) highlight a different reason why these features are necessary. In their research they have surveyed people about their trust in AI content moderation systems and have found that people generally trust AI systems less than human moderators.

The previous studies deem the deployment of AI for content moderation a threat to free speech and claim that understanding human content is a task AI will never be good at due to the nuanced nature of understanding sensitive content. A symptom specific to a distrust within the user base toward AI moderation, is the creation of so-called "*folk theories"*. These are discussions among users, where they try to understand the inner workings of the AI moderators.

These problems of distrust can be alleviated by having an AI system be able to explain why a particular piece of content was taken down. Another feature the authors claim is beneficial in creating trust by the users, is allowing the users to give feedback and allowing the users to call to the help of a human moderator.

A Strange loop

We live in an unprecedented age, where AI not only moderates content, but also creates it. The effects that these two processes have on each other has not been accurately investigated, but with the rise of the concept of the "Dead Internet Theory" we are looking at an era where social media has been, as it were, automated.

To disambiguate: Dead Internet Theory is a conspiracy theory which claims that the internet "died" in around 2016-2017 (Tiffany, 2021). The internet you're looking at, and are interacting with, now, is presumably not made by humans, but by content-generating bots. The conspiracy has recently taken an eerie turn toward reality with the rise of generative AI (Walter, 2024). Walter also argues that the "Dead Internet" gives rise to "a concerning convergence of consumerism and artificiality in spaces once dedicated to human expression".

What this means for content *moderation* however, remains unclear. Perhaps we could see the interactions of this systems as a sort of "Generative Adversarial Network" (Goodfellow *et al.* 2014), in which the content generation AI tries to make content good enough to "fool" the content moderation AI, and the moderation AI tries to identify AI-generated content. This assumes, of course, that moderation AI is inherently adversarial towards generative AI in the first place, and experience has shown that this cannot be assumed.

Ghosh (2021) also points out that the companies on whose platforms this AI-generated content is hosted stand to gain from it existing, as the sheer volume of AI-generated content helps drive engagement of (human) users. In this way, we can relate the papers of Walter (2024) and Ghosh (2021) to each other, explaining that this increase in artificial content is basically a consequence of a platform's existence, and desire towards commodification of user experiences. What this means then, is that AI content *moderation* is not at odds with AI content *generation*, as the platform's main goal for both of these systems is to maximize user engagement.

This is, of course, problematic. If we choose to embrace automated social media, it is important to prevent it from simply maximizing user engagement and consumerism. Instead, we must focus on regulating both content moderation and generation "to ensure that technological advancement enhances, rather than detracts from, the human experience in the digital domain." (Walter, 2024).

Conclusion

In this paper, we examined the ethical implications of delegating the recognition and moderation of user-generated content to automated methods. Through mediation theory, we explored the interplay between automated systems, societal norms, and the human stakeholders who design and use these technologies. This perspective underlines how automated systems both reflect and shape the cultural contexts in which they operate, often amplifying biases and introducing new ethical challenges.

Our investigation into accountability highlighted that while engineers and lower-level management influence content moderation decisions, ultimate responsibility should rest with the owners and leadership of social media platforms. They must implement effective supervision mechanisms and safeguards to mitigate societal harm.

The application of value-sensitive design demonstrated how embedding ethical principles during the development of automated moderation systems could ensure alignment with societal values. Continuous monitoring, transparency, and collaboration with human moderators were identified as essential components to address the limitations of current AI systems, including their inability to grasp nuanced or context-sensitive content fully.

In the last section, we investigated the possibility of an internet that is both AI-generated and AI-moderated. This scenario illustrates the need to regulate both content moderation and content generation to preserve the integrity of digital spaces.

We believe that in general, human intervention is necessary to keep automated moderation on the right track. Rigorous supervision should be implemented by those accountable. To this end, mediation theory can help the people responsible for supervision in understanding the subtle interactions between humans and technology at play, while value sensitive design can provide concrete guidelines for development. If we can guarantee that automated moderation tools are aligned with societal values, possibly by government intervention, then we might be able to prevent the scenario of a truly Dead Internet.

Bibliography

Amnesty International. (2022). Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya. Retrieved https://www.amnesty.org/en/documents/ASA16/5933/2022/en/

Appel, R. E., Pan, J., & Roberts, M. E. (2023). Partisan conflict over content moderation is more than disagreement about facts. Science Advances, 9(44), eadg6799. https://www.science.org/doi/10.1126/sciadv.adg6799

Arive, L. (2024). *The Ethics of Content Moderation: Balancing free speech and Harm Prevention*. Chekkee. https://chekkee.com/the-ethics-of-content-moderation-balancing-free-speech-and-harm-prevention/

Benton, B., Choi, J. A., Luo, Y., & Green, K. (2022). Hate speech spikes on Twitter after Elon Musk acquires the platform. *School of Communication and Media, Montclair State University*.

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *Lecture notes in computer science* (pp. 405–415). https://doi.org/10.1007/978-3-319-67256-4_32

Birch, K., & Bronson, K. (2022). Big tech. *Science as Culture*, *31*(1), 1-14.

Calo, R. (2024). Courts Should Hold Social Media Accountable — But Not By Ignoring Federal Law. *Harvard Law Review.* Retrieved https://harvardlawreview.org/blog/2024/09/courts-should-hold-social-media-accountable-but-not-by-ignoring-federal-law/

Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter?. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW2), 1-52.

Crosset, V., & Dupont, B. (2022). Cognitive assemblages: The entangled nature of algorithmic content moderation. Big Data & Society, 9(2). https://doi.org/10.1177/20539517221143361

Ghosh, D. (2021). The Future of Platform Power. *Journal of Democracy*. Retrieved https://muse.jhu.edu/article/797794

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2), 2053951720943234.

Gomez, J. F., Machado, C. V., Paes, L. M., & Calmon, F. P. (2024).. *Algorithmic arbitrariness in content moderation*. arXiv.org. https://arxiv.org/abs/2402.16979

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, .C., & Bengio, Y. (2014). Generative Adversarial Nets. Neural Information Processing Systems.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1). https://doi.org/10.1177/2053951719897945

Helmond, A. (2015). The Platformization of the Web: Making Web Data Platform Ready. Social Media + Society, 1(2). https://doi.org/10.1177/2056305115603080

Hernández, A. D., Owen, R., Nielsen, D. S., & McConville, R. (2023). Ethical, political and epistemic implications of machine learning (mis)information classification: insights from an interdisciplinary collaboration between social and data scientists. *Journal of Responsible Innovation*, *10*(1). https://doi.org/10.1080/23299460.2023.2222514

Johnson A., Castro D. (2021). Overview of Section 230: What It Is, Why It Was Created, and What It Has Achieved. *Information Technology & Innovation Foundation*. https://itif.org/publications/2021/02/22/overview-section-230-what-it-why-it-was-created-and-what-it-has-achieved/

Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., & Tan, C. (2022). Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-18). https://doi.org/10.48550/arXiv.2204.11788

Larrauri, J. S. R. I. &. H. P. (2023). *The algorithmic management of polarization and violence on social media*. Knight First Amendment Institute. https://knightcolumbia.org/content/the-algorithmic-management-of-polarization-and-violence-on-social-media

Molina, M.D., Sundar, S.S. (2022). When AI moderates online content: effects of human collaboration and interactive transparency on user trust, *Journal of Computer-Mediated Communication*, Volume 27, Issue 4, zmac010, https://doi.org/10.1093/jcmc/zmac010

O'Brien, M., & Hoffstaedter, G. (2020). "There We Are Nothing, Here We Are Nothing!"—The Enduring Effects of the Rohingya Genocide. *Social Sciences*, *9*(11), 209. https://doi.org/10.3390/socsci9110209

Rieder, B., & Skop, Y. (2021). The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. Big Data & Society, 8(2). https://doi.org/10.1177/20539517211046181

Roberts, S. T. (2017). *Content moderation*.

Tiffany, K. (2021). *Maybe You Missed It, but the Internet "Died" Five Years Ago.* The Atlantic. https://www.theatlantic.com/technology/archive/2021/08/dead-internet-theory-wrong-but-feels-true/619937/

Unicef. (n.d.). It's not censorship to stop hateful online content, insists UN rights chief. Retrieved https://news.un.org/en/story/2025/01/1158886 van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds & Machines* **30**, 385–409 https://doi.org/10.1007/s11023-020-09537-4

Walter, Y. (2024). Artificial influencers and the dead internet theory. *AI & Soc* https://doi.org/10.1007/s00146-023-01857-0