From civility to parity: Marxist-feminist ethics for context-aware algorithmic content moderation

Dayei Oh
Helsinki Institute for Social Sciences and Humanities, University of Helsinki
dayei.oh@helsinki.fi

Abstract

Algorithmic content moderation governs online speech on large-scale commercial platforms, often under the guise of neutrality. Yet, it routinely reproduces white, middle-class norms of civility and penalizes marginalized voices for unruly and resistant speech. This paper critiques the prevailing 'pathological' approach to moderation that prioritizes sanitization over justice. Drawing on Marxist-feminist ethics, this paper advances three theses for the future of context-aware algorithmic moderation: (1) prioritizing participatory parity over civility, (2) incorporating identity- and context-aware analysis of speech; and (3) replacing purely numerical evaluations with justice-oriented, community-sensitive metrics. While acknowledging the structural limitations posed by platform capitalism, this paper positions the proposed framework as both critique and provocation, guiding regulatory reform, civil advocacy, and visions for mission-driven online content moderation serving digital commons.

1 Introduction

AI-driven content moderation increasingly shapes online discourse, flagging and making decisions before human review. Major corporate platforms (e.g., Meta, Google, X) justify moderation tasks as necessary measures to ensure 'healthy' dialogues (Google Perspective, n.d.1) [1] on their platforms so that users 'create and share ideas and information, as well as express their opinions and beliefs without barriers' (X, 2024). Platform capitalists largely share the 'pathological' (Lee and Scott-Baumann, 2020) and hygienist vision that some language is 'toxic' and ought to be surgically removed and insulated for the health and safety of the internet.

However, the problems of such pathological logic in mainstream content moderation have been theoretically and empirically criticized by academia and civil societies. Thylstrup and Waseem (2020) analyze that the idea of 'sanitized,' 'purified' digital space is deeply intertwined with power struggles around the boundaries of what is considered 'dirty' in society while the power to define what is 'dirty' is unequally distributed. The way platform capitalism operationalize 'toxic' language often results in systematic biases where certain voices and content get moderated more than others, especially when it comes to political dissent. Studies have shown that the language of the marginalized, including LGBTQ+, African American, and 'angry feminists' are more likely to get higher toxicity scores and moderated (Oh and Downey, 2025; Sap et al., 2019; Thiago et al., 2021). These are not some accidental technical glitches of machine, but they reflect the biases embedded in the very logic of the moderation.

[1] When I published my previous paper on algorithmic moderation in 2024 (Oh and Downey, 2025), Google Perspective API website used the term 'healthy' more extensively to describe the aim of their product. However, as of June 2025, the language has shifted to emphasize 'better conversations' (Google Perspective n.d.1) and 'productive, fulfilling discussions' (Google Perspective, n.d.2), reflecting a potentially toned-down approach that moves away from framing moderation in 'pathological' terms.

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 32–40

When platform capitalism and large-scale commercial platforms equate 'healthy' conversations to the white, middle-class centric view of civil-rational conversations, it is not surprising that the angry and uncivil voices of the marginalized are flagged as toxic (Oh and Downey, 2025).

Feminist and racial justice works have argued that uncivil, unruly, and disruptive tactics are indispensable in activist languages to get their voices heard and to achieve social changes and justice (Bickford, 2011; Zamalin, 2021; Zerilli, 2014). Without recognizing this important reality, any commercial platform moderation will keep making the same 'errors' in which the marginalized voices are more likely to be moderated, despite their self-proclaimed role as 'custodians of the Internet' (Gillespie, 2018).

More than technical, mathematical measures to 'de-bias' moderation models, what is more urgently needed is a new vision for future algorithmic content moderation. This paper advances three exploratory theses for the development of context-aware algorithmic moderation, grounded in Marxist-feminist digital ethics (D'Ignazio and Klein, 2023; Fuchs, 2022). This will involve moderation AIs that (1) shift the normative role of moderation from preserving civility to promoting 'participatory parity' (Fraser, 2024). It also requires AIs that (2) evaluate speech in relation to identity and context. Lastly, context-aware AIs require (3) a new justice-oriented paradigm to assess the performance and consequences of algorithmic moderation in terms of temporal and societal impacts to marginalized communities, beyond mere numerical performance metrics.

Throughout this paper, I use the term 'silencing' to refer broadly to both hard (e.g., content deletion, account suspension) and soft moderation (e.g., downranking, shadowbanning). While the primary focus of this paper is on the ethical implications of hard moderation, I acknowledge that soft moderation techniques play a significant role in reinforcing dominant norms and suppressing epistemic dissents (e.g., shadowbanning of queer, trans, and disabled content creators in Rauchberg, 2022). A Marxist-feminist critique therefore treats silencing not only as a matter of censorship but of participatory disparity and exclusion.

2 Marxist-feminist digital ethics for context-aware algorithmic moderation

Algorithmic moderation systems are often governed and justified by two dominant ethics frameworks: deontology and consequentialism. As a normative moral philosophy, deontological theories assess moral claims about the intrinsic rightness and wrongness of actions based on established moral principles, duties, and categorical imperatives (Ess, 2013). Under this framework, platform rules, such as community guidelines defining hate speech and graphic violence, are treated as fixed principles that govern moderation decisions uniformly and rigidly (Li and Zhou, 2024).

By contrast, consequentialist ethics place the outcomes of actions, rather than their intrinsic nature, at the center of moral judgements (Ess, 2013). This approach evaluates moderation practices based on their societal and temporal impacts, emphasizing the need for flexibility and context sensitivity to maximize the intended 'good' consequences (Li and Zhou, 2024).

However, both paradigms fall short. Deontology's rigidity, which focuses solely on the intrinsic properties of content, can lead to inflexible decisions that fail to account for the broader social and political contexts in which speech occurs. For example, Facebook and YouTube's removal of human rights and journalistic content due to nudity and graphic violence policies illustrates how deontological rule enforcement can overlook the historical and political significance of content (Gillespie, 2018; York, 2021). The blanket enforcement of moderation rules can also disproportionately penalize marginalized groups who use disruptive language as a tool for political resistance, reinforcing power asymmetries. By treating moderation as a matter of categorical imperatives rather than situational judgment, deontological approach risks undermining democratic values and participatory parity.

Consequentialists, while taking outcomes of action into account when discussing legitimacy of moderation policies, often lack the discussions about justice and equality, reducing ethics to utility calculus (Fuchs, 2022). Haines (2006) argues that consequentialism is 'not egalitarian' since maximizing total happiness can still justify exploitation of the minorities if it benefits the majority. If moderating uncivil dissents of the socially marginalized can please the social

majority, a platform might be ethically justified to do so from a consequentialist perspective.

It is also critical to recognize that platforms *are* the neoliberal elites, structurally aligned with state and corporate power (Zuboff, 2019). This points to the conflicted role of platforms in society and politics, reflecting the tension between platforms' commercial incentives and their social responsibilities. By calling them 'custodians of the Internet,' Gillespie (2018) emphasizes platforms' responsibility to care for the communities they host beyond commercial interests.

Ethical communication is essential to democracy and public spheres (Habermas, 1996, 2006) and platforms that host and govern public discourse must have commitments to public good – justice, inclusion and democracy – beyond profit motives and commercial pressures. Rethinking the normative role of platforms and moderation is crucial, especially at this political juncture in which many large commercial platforms are distancing themselves from their roles to protect users from 'harmful' content including misinformation and hate speech (e.g., Meta's decision to end factchecking program and ease content restrictions, in Bhuiyan and Kerr, 2025; McMahon et al., 2025).

Marxist and feminist digital ethics address this limitation by focusing on power relations and epistemic justice for the marginalized at the intersections of class, gender, race and other axes of identity (D'Ignazio and Klein, 2023; Fuchs, 2022). Marxist digital ethics (Fuchs, 2022), grounded in the works of Karl Marx, prioritizes emancipation through the dismantling of structures of exploitation and inequality. This ethical approach critiques systems that perpetuate and reinforce existing hierarchies under the guise of neutrality. Unlike abstract moral philosophies, Marxist ethics is rooted in 'praxis,' referring to the alignment of ethical principles with tangible practical actions for social change. The core objective for Marxist digital ethics is thereby to implement digitalization that fosters the common good (Fuchs, 2022: 7), empowering marginalized groups to challenge systemic inequalities.

Feminist digital ethics (D'Ignazio and Klein, 2023) emphasizes the importance of situated knowledge and epistemic justice. Feminist digital ethics call for platform design and governance to center on lived experiences and knowledges of those who are affected the most by the technological systems. D'Ignazio and Klein (2023) argue that data and algorithmic systems must be grounded in contextualized understanding and care beyond the assumed neutrality and objectivity of social data, digital infrastructures, and governance.

In the context of algorithmic moderation, this Marxist-feminist framework transcends the deontological fixation on rule-following or the consequentialist emphasis on indiscriminate, aggregate outcomes. Instead, the justification for moderating lies in the potential of these decisions to promote human dignity, freedom, and justice particularly for those who are historically excluded and dismissed from public discourse. A Marxist-feminist digital ethics demands we ask: Whose speech is being silenced? Who benefits from the silence? Neutrality becomes untenable when 'neutral' rule followings reinforce dominant cultural norms and invisibilize dissent. Moderation must be judged not by how well it conforms to universal rules or statistical accuracy, but by how it shapes the distribution of voice, visibility, and political possibility. For example, Marxist-feminist reasoning might prioritize preserving activist speech, even if uncivil, because of its potential to mobilize political movements and promote social justice. Conversely, it might advocate for the removal of coded or borderline hate speech, even when it technically adheres to community guidelines, because of its capacity to harm marginalized communities and normalize exclusionary ideologies.

This Marxist-feminist framework sets the stage for the paper's three theses: each reimagining content moderation through the lens of justice and participatory parity (Fraser, 2024).

3 From promoting civility to parity

The dominant 'pathological' approach to content moderation treats online discourse as something to be cleansed: e.g., silencing 'uncivil' speech is framed as surgically removing toxic parts off the Internet, thereby promoting 'healthy,' civil and rational conversations. When socially acceptable speech is equated to civil speech, it is no surprising to see many people lament and frown at disruptive and unruly speech for their inability to 'have civil conversations' over disagreements. It is those who use uncivil language that must be silenced and punished until they correct their language in a polite and rational manner.

However, this civility-focused approach often results in the sanitization of public discourse,

instead of the promotion of democratic discourse in terms of tolerance and parity between the social majority and minorities (Oh and Downey, 2025). Many feminist and anti-racist works have argued how the norm of civility is a product of White, middle-class, male centered notion of 'good' language (Bickford, 2011), subjugated to the marginalized who themselves did not have chance or power to decide what is socially allowed and acceptable. In doing so, civility historically becomes a yardstick to tone-police the voices of the marginalized while dismissing political demands for justice behind their unruly voices (Bickford, 2011; Zamalin, 2021; Zerilli, 2014). As feminist scholar Zerilli (2014) puts, 'uncivil public discourse is symptomatic of a more general democratic deficit […] If some citizens are more prone to shout, that may well be because those in power are not listening' (p.112). From suffragettes, civil rights activists, to contemporary social justice activists, the 'rude, disrespectful, and unreasonable' rhetoric, which the social majority might refer to as 'toxic language,' has been crucial expressive and instrumental tools to disturb the status quo and demand for radical changes.

By missing the symbolic and instrumental values of uncivil and unruly speech for the marginalized and their activism, the current civility-centered moderation puts uneven burdens on the marginalized to tone-police their demands while the discriminatory or unjust systems they are criticizing receive protection. This is particularly relevant when the uncivil languages of marginalized groups get higher toxicity scores than the rationalized bigotry of extremist groups expressed in pseudo-intellectual, and seemingly civil-rational language to mask xenophobia, homophobia, and racism (Thiago et al., 2021; Thylstrup and Waseem, 2020). Oh and Downey (2025) also show that the current toxic language detection tools such as Perspective API does not recognize intolerant speech well when the exclusionary ideas are hidden in seemingly civil rhetoric, while 'angry feminist' speech with swearwords gets higher toxicity. Such outcomes expose the deeper flaws of platforms' reliance on neoliberal notion of civility, which ultimately perpetuate existing inequalities in public discourse.

Instead of the civility-centered approach to tone-police and silence the uncivil voices of the marginalized, a more ethical approach aligning with Marxist-feminist framework is to prioritize preserving the underrepresented voices of society to enhance the participatory parity (Fraser, 2024) on the Internet. For ethical communications for democracy to succeed is not to tone-police and sanitize public sphere, but to ensure that diverse actors from social peripheries to center can participate in discourse, mending disparity between the majority and minorities. Early academic discussions about new media and public spheres offered important insights into this shift. Habermas (2006) observed that the Internet and communication technologies revitalized the 'egalitarian' dimensions of public spheres, enabling politically active citizens to foster issue publics and shape public opinion. Other scholars have similarly highlighted how new media create spaces for subaltern counterpublics and alternative political engagements, both on the political left and right (Downey and Fenton, 2003). Here, the normative vision of the Internet rests on the idea of mending the participatory disparity between the elites and citizens, uplifting the marginalized to influence national political agendas who otherwise do not have power and resources to influence national debates and public opinion formations (e.g., mass media ownership, lobbying, parliamentary influences).

To revitalize this promise of internet technologies, platforms must move beyond the pathological framework. Their duties are to protect the voices of the marginalized and promote parity so that these voices are not drowned out by the dominant biases of society and elites who often dismiss their struggles. When governments, media pundits, and political commentators fail to address the demands of the marginalized, platforms' task is not to act as a 'neutral intermediaries' of neo-liberal free speech (and thereby acting in favor of state and business interests), but as proactive defenders of those excluded from offline public debates due to harassment, fear, or self-censorship, establishing platforms as safe havens for the fight for emancipation. While I do not expect platform capitalists to voluntarily accept the proposed Marxist-feminist visions, this piece serves an important critique to push the Overton window to rethink and redefine the normative role of platforms, unmasking biases and harms under their disguise of 'neutrality' and 'civil-rational speech' (also Oh and Downey, 2025).

4 From text to context-aware identity analysis

To achieve algorithmic moderation that can protect the voice of the marginalized and promote participatory parity online, moderation AIs should consider identities and power inequalities between users when they assess the 'toxicity' of content and decide which content should be moderated.

This shift requires moderation AIs to evaluate more than the textual content of speech. For example, swear words used by migrant justice activists to criticize white nationalist politicians are not equivalent to swear words used by white nationalist politicians and their supporters to harass and exclude non-White, migrant users. Current moderation tools often fail to distinguish between these intentions and contexts. For example, slurs reclaimed by marginalized communities themselves as acts of empowerment are often flagged as toxic, despite lacking hateful intent (Thiago et al., 2021).

The future AI should be context-aware, capable of discerning the identities of speakers and recipients of potentially 'toxic' messages. Davidson's (2024) work highlights how multimodal models, incorporating metadata such as usernames and profile images, can improve differentiation between toxic uses of slurs and reclaimed usages. Here, it is important incorporate intersectionality of identities, such as religion, disability and sexuality (Magee et al., 2021).

This contextual assessment could improve the detection 'borderline content' by far-right, alt-right, White supremacist, and other extremist groups and content (Krzyżanowski and Ledin, 2017; Thiago et al., 2021). By incorporating speaker identities, AI could better detect the exclusionary intent behind such seemingly rational language.

While this context-aware AI can offer significant potential, it also raises critical ethical concerns that must be addressed before implementation. First, automatically inferring user identities risks misrepresentation and harm. For example, automatic gender and race recognition systems have been criticized as 'misgendering machines' (Keyes, 2018) and a 'new phrenology' (Ajunwa, 2021). To mitigate these issues, context-aware AI must move beyond simplistic assumptions about identity and incorporate user-defined metadata such as pronouns, to improve accuracy and inclusivity (Lauscher et al., 2022).

Second, identity-aware moderation AIs can get exploited by bad actors who impersonate marginalized users, such as sock-puppet accounts created by extremist groups to spread exclusionary ideologies under the guise of minoritized identities. These deceptive tactics not only undermine the goals of participatory parity but also risk discrediting genuine voices and eroding trust in context-sensitive moderation systems. Therefore, it must be designed with safeguards against such bad actors.

Third, automatic identity detection can be challenging in authoritarian contexts where anti-state activists are likely to avoid revealing their real identities for safety reasons. This suggests that identity-aware moderation should be built in more region-specific safeguards, instead of assuming models built in liberal democratic contexts as universally applicable.

Lastly, identity-aware moderation must address the questions of not only privacy of users, but also epistemological power for users in relation to platforms. For large-scale identity-aware moderation, should platforms infer users' minority status? Should minority users be required to self-identify instead? Each approach will introduce new forms of biases and challenges, and therefore it must be carefully implemented with collaboration between NLP/AI developers, ethicists, and minority communities themselves.

Due to these concerns, identity-sensitive frameworks discussed here can be vulnerable to misuse when applied without thorough ethical reviews. While this paper advocates for identity-aware moderation to protect marginalized voices, I explicitly caution against hasty applications.

5 From metrics to justice-oriented evaluation

Another important proposition in this paper is to imagine a new evaluation framework for algorithmic content moderation that aligns with Marxist-feminist digital ethics. AI moderation is currently evaluated through numerical metrics such as precision, recall, and F1 scores. These benchmarks offer statistics of scientific rigor, but tell us little about the temporal and societal consequences of moderation decisions. It also risks prioritizing efficiency and scalability over ethical accountability, treating all false positives and false negatives as equally significant. This reductionist

approach overlooks the reality that not all errors carry the same weight in political and social contexts. Some mistakes can lead to far greater harms, and a model with 95% precision may still disproportionately silence activists with minority identities. The metrics are quantitative but morally flat.

I argue for a justice-oriented approach to evaluate performances of algorithmic moderation systems. This means moderation systems must be assessed in terms of their temporal, social, and political impacts. A misclassification error during a moment of political uprising carries exponentially greater harm than a similar error during routine discourse. Consider a scenario from 2011. During the Arab Spring following the tragic murder of Khaled Saeed by Egyptian police, a Facebook page, called 'We are All Khaled Saeed' was created. The page played central roles in organizing protests against police corruption and dictator Hosni Mubarak. On the day before a planned Friday protest, the page was removed by Facebook, following the takedown of another page of the Nobel Peace Prize winner Mohamed ElBaradei the week prior (York, 2021). From Facebook's perspective, they took down one page that violated their real name policies, one of many accounts and pages they disabled every day. However, the timing of the takedowns and lack of full explanations from the company representatives could have resulted in the most paranoid explanations among the public about the page being the target of malicious oppositions and jeopardizing the planned protests and uprising (York, 2021). Although the page was restored in six hours later thanks to the hard efforts of international human rights organizations and activists, the disruption and damage to the movement's momentum cannot be undone.

Even in liberal democracies, within certain temporal and political junctures, misclassification of the voice of the marginalized activists can result in particularly damaging consequences for participatory parity. For instance, let us look at the most recent reactionary politics from Trump and Vance's election campaign targeting transgender communities (Barrow, 2024). Conservative media and political elites leveraged their resources and power to amplify their reactionary anti-trans agenda, which then led liberal, center-left elites to engage with the discourse. It is exactly in this particular context that mending the participatory inequality between the targeted community and majority in the society is of the utmost importance. Again, the normative task of moderation is not to sanitize or to tone-police the anger of the marginalized (e.g., angry trans activists in this juncture), but to safeguard and uplift those voices that are otherwise unheard and unrecognized in the existing elite-driven political and media landscape.

Context-aware AI must acknowledge that certain communities and voices are extra vulnerable and should be protected against misclassification. Justice-oriented evaluations of moderation AIs must incorporate such identity-sensitive, community-specific impacts, assessing how their moderation systems and errors disproportionately affect marginalized communities.

Future evaluations of moderation systems and transparency reports should account for temporal and socio-political impacts specific to marginalized communities and their struggles. While this paper does not provide a finished formula to determine how much weight we should assign to temporal and identity-based factors when evaluating the performances of algorithmic moderation, Instead, it opens up a call for a new paradigm in assessing and reporting the performance of algorithmic moderation systems. Waseem (2016), for instance, advocates for a weighted F1-score so that misclassification on minority classes is penalized.

Furthermore, future transparency reports should overhaul their transparency reports to provide richer case studies of moderation errors beyond mere aggregate statistics. They must include case-based narratives, highlight errors involving marginalized communities, and measure how moderation affects participatory inequality over time. York's (2021) interviews with global activists and platform employees highlight numerous instances in which many activist content and accounts were never restored and permanently removed without full explanations despite appeals. These cases should be treated as democratic harms caused by platforms, not edge cases, and must be explained in the transparency reports.

A Marxist-feminist ethics demands that we ask: Who benefits from algorithmic silence, and who suffers from its mistakes? AI evaluation must prioritize accountability to those most vulnerable, not just efficiency for those most powerful. Achieving such contextual evaluations requires greater collaboration between platform AI developers, social scientists, ethicists, civil society,

and the marginalized communities. Interdisciplinary approaches can help identify the specific vulnerabilities of marginalized groups and design systems that are sensitive to these challenges.

6 Conclusion

Grounded in Marxist-feminist ethics (D'Ignazio and Klein, 2023; Fuchs, 2022), this has advanced three interconnected theses for developing context-aware algorithmic moderation systems that prioritize participatory parity in public discourse, while critiquing mainstream moderation logics in large-scaler commercial platforms such as Meta, Google and X. First, this paper has called for a departure from the 'pathological' (Lee and Scott-Baumann, 2020) approach to moderation toward a model that recognizes and protects the dissenting voices of the marginalized for participatory parity between the majority and minorities (Oh and Downey, 2025). Second, it emphasized the importance of incorporating identity and context into moderation systems, allowing for more nuanced assessments of speech that account for power dynamics between speakers and audiences. Third, it has proposed a new justice-oriented paradigm for evaluating moderation systems, one that moves beyond statistical metrics to consider the temporal and community-specific consequences of moderation decisions.

A key structural challenge arises when we try to translate this Marxist-feminist ethics to the existing platform capitalism. Dominant commercial platforms operate on capitalist logics to maximize user engagements and bring ad revenues, thereby catering to majority preferences. Furthermore, platforms are not neutral intermediaries but integral components of the neoliberal elite class, structurally aligned with state and corporate power (Zuboff, 2019). To amplify the marginalized voices, especially when they disrupt the dominant norms or lack shareholder interests and commercial appeal, is misaligned to the profit incentives. In this context, the Marxist-feminist proposal to reorient moderation risks appearing idealistic and structurally incompatible with the logics of platform capitalism.

Nonetheless, the purpose of this framework is not to assume that commercial platforms will voluntarily adopt a Marxist-feminist approach to content moderation. Rather, it is intended to serve as a strategic provocation, challenging what we think content moderation is for. As Fraser (2024) notes in her theorization of counterpublics, normative critiques serve not only to assess what is but also to illuminate what ought to be. Even in capitalist systems, ethical redefinitions can shape the terms of debate and the legitimacy of existing practices on commercial platforms. It aims to push the Overton window in content moderation debates to denaturalize civility and to propose a radical vision to moderation as a tool for justice and parity.

Practically, the proposed framework supports three interrelated domains of intervention. First, it offers a normative foundation for regulatory frameworks that hold platforms accountable for the asymmetrical harms caused by their moderation practices. These regulatory frameworks require justice and parity-based auditing of content moderation outcomes and more in-depth, disaggregated transparency reports.

Second, the framework provides a discursive and tactical resources for civil society organizations, digital rights advocacy groups, and marginalized communities. Advocacy efforts should focus on ensuring that marginalized voices are included at every stage of the moderation AI building process. Waseem (2016) finds that expert participation in annotation (e.g., feminist and anti-racism activists) improves hate speech detection system performances. These efforts are necessary to monitor mainstream algorithmic moderation as a force for participatory parity, and not a tool for sanitizing dissenting discourse.

Third, the framework points to alternative platform design and algorithmic moderation experiments that could prototype identity-sensitive, context-aware moderation under mission-driven platform governance. Such prototypes would not only test the viability of the approach but also yield more empirical data to inform future regulation and technical refinements. This pilot test will also provide important insights about the ethics of automatically inferring user identities.

Finally, while this paper draws primarily from examples based in Western platforms (e.g., X, Meta, YouTube), it recognizes the urgent need to globalize this conversation. Future research and implementation must also expand the geographic and cultural scope of these discussions. In authoritarian contexts, the tools proposed here (e.g., incorporating user identity into algorithmic assessments) can carry heightened risks of misuse,

including political surveillance, profiling and repression. A Marxist-feminist ethics must therefore be vigilant against the co-optation of identity-aware moderation for state or corporate control. It must also foreground local epistemologies and grassroots coalitions in non-Western contexts, ensuring that the pursuit of justice is not flattened into a universalist template.

To summarize, the future of algorithmic content moderation cannot be separated from the ethical, political, and economic structures in which it is embedded. This paper calls for a fundamental rethinking of moderation not merely as a value-neutral technical task, but as a site of moral and political struggles in which the stakes are visibility, dignity, and democratic participation of marginalized communities in digital public life.

7 Limitations

This paper is purely theoretical and exploratory, relying on conceptual analysis rather than empirical data or pilot studies. While it advances key theses on context-aware algorithmic moderation through a Marxist-feminist ethical framework, it does not provide experimental results or large-scale empirical validation. Future research should conduct empirical studies to test the practical implementation of identity- and context-aware moderation and evaluate the societal impact of shifting from civility-based frameworks to participatory parity.

References

Ifeoma Ajunwa. (2021). Automated video interviewing as the new phrenology. *Berkeley Tech. LJ*, 36, 1173.

Bill Barrow (2024). Trump and Vance make anti-transgender attacks central to their campaign's closing argument. 1 November. *AP News*. https://apnews.com/article/trump-harris-transgender-politics-61cff97a64fac581ffc5f762be4c57d3

Susan Bickford. (2011). Emotion talk and political judgment. *The Journal of Politics*, 73(4), 1025-1037.

Johana Bhuiyan and Dara Kerr. (2025). Zuckerberg's swerve: how diversity went from being a Meta priority to getting cancelled. *The Guardian*, 11 Feb. https://www.theguardian.com/technology/ng-interactive/2025/feb/11/dei-meta-facebook

Thomas Davidson. (2024). *Auditing multimodal large language models for context-aware content moderation.*

Catherine D'ignazio and Lauren F. Klein. (2023). *Data feminism*. MIT press.

John W. Downey and Natalie Fenton. (2003). New media, counter publicity and the public sphere. *New media and society*, 5(2), 185-202.

Charles Ess. (2013). *Digital media ethics*. Polity.

Nancy Fraser. (2024). Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. In *New Critical Writings in Political Sociology* (pp. 489-513). Routledge.

Christian Fuchs. (2022). *Digital ethics: Media, communication and society volume five*. Routledge.

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

Google Perspective (n.d.1). *Research*. https://www.perspectiveapi.com/research/

Google Perspective (n.d.2). *Case studies*. https://perspectiveapi.com/case-studies/

Jurgen Habermas. (1996). *Between facts and norms*. MIT Press.

Jurgen Habermas. (2006). Political communication in media society: Does democracy still enjoy an epistemic dimension? The impact of normative theory on empirical research. *Communication theory*, 16(4), 411-426.

William Haines. (2006). Consequentialism. In *Internet Encyclopedia of Philosophy*. https://iep.utm.edu/conseque/

Os Keyes. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW), 1-22.

Michal Krzyżanowski and Per Ledin. (2017). Uncivility on the web: Populism in/and the borderline discourses of exclusion. *Journal of Language and Politics*, 16(4), 566-581.

Anne Lauscher, Archie Crowley, and Dirk Hovy. (2022). Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. *arXiv preprint arXiv:2202.11923*.

Yenn Lee and Alison Scott-Baumann. (2020). Digital ecology of free speech: Authenticity, identity, and self-censorship. In Yates SJ and Rice RE (Eds) *The Oxford Handbook of Digital Technology and Society.* Oxford University Press.

Luzhou Li and Kui Zhou. (2024). When content moderation is not about content: How Chinese social media platforms moderate content and why it matters. *New Media and Society*, 14614448241263933.

Liam Magee, Lida Ghahremanlou, Karen Soldatic, and Shanthi Robertson. (2021). Intersectional bias in causal language models. *arXiv preprint arXiv:2107.07691*.

Liv McMahon, Zoe Kleinman and Courtney Subramanian. (2025). Facebook and Instagram get rid of fact checkers. *BBC*, 7 January. https://www.bbc.co.uk/news/articles/cly74mpy8klo

Dayei Oh and John W. Downey. (2025). Does algorithmic content moderation promote democratic discourse? Radical democratic critique of toxic language AI. *Information, Communication & Society*, 28(7), 1157-1176.

Jessica Sage Rauchberg. (2022). #Shadowbanned: Queer, Trans, and Disabled creator responses to algorithmic oppression on TikTok. In *LGBTQ digital cultures* (pp. 196-209). Routledge.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. (2019, July). The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics,* 1668-1678.

Dias Oliva Thiago, Antonialli Dennys Marcelo and Alessandra Gomes. (2021). Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality and culture*, 25(2), 700-732.

Nanna Thylstrup and Zeerak Waseem. (2020). Detecting 'dirt' and 'toxicity': Rethinking content moderation as pollution behaviour. Available at *SSRN 3709719*.

Zeerak Waseem. (2016). Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, 138-142.

X (2024). *Abuse and harassment*. https://help.x.com/en/rules-and-policies/abusive-behavior

Jillian C. York. (2022). *Silicon values: The future of free speech under surveillance capitalism*. Verso Books.

Alex Zamalin. (2021). *Against civility: the hidden racism in our obsession with civility*. Beacon Press.

Linda Zerilli. (2014). Against civility: A feminist perspective. In *Civility, legality, and justice in America*, 107-131.

Shoshana Zuboff. (2019). *The age of surveillance capitalism*. Profile books.