International Journal of Communication 20(2026), Forum 2168–2175 1932–8036/2026FRM0002

# Scaling Down to Scale Up?
# When Less Moderation Claims to Protect More

KYOOEUN JANG
University of Southern California, USA

In January 2025, Meta announced sweeping changes to its content moderation practices: replacing fact-checkers with community oversight, simplifying speech-related policies, and refocusing resources on "high-severity" violations. Drawing on Anna Tsing's (2012) nonscalability theory, this article examines how these changes attempt to transform the inherently relational nature of content moderation processes into discrete, manageable, and ultimately scalable units. It identifies three contradictions in Meta's scalability project: how community-based moderation and increased political content create tensions between truth and engagement; how categorizing content into "high" and "low" severity disregards interconnected harm ecosystems; and how geographic relocation and regulatory resistance introduce governance volatility that undermines consistent application of safety standards. This analysis shows how platform governance projects pursuing scale through standardization can paradoxically exacerbate the harms they claim to address, fragmenting rather than sustaining coherent safety frameworks.

*Keywords: content moderation, platform governance, nonscalability, participatory legitimacy, epistemic authority*

In January 2025, Mark Zuckerberg announced sweeping changes to Meta's content moderation strategy, describing it as a return to the company's "roots" in free expression (Kaplan, 2025). Zuckerberg claimed the company's current content moderation approach resulted too often in "censorship" and outlined six major changes: replacing fact-checkers with community notes, simplifying content policies, changing enforcement mechanisms, reintroducing civic content, relocating Trust and Safety teams, and pushing back against global regulation. By framing these changes as a return to Meta's "roots," Zuckerberg's narrative echoes what Anna Tsing (2012) calls a "scalability project"—an attempt to expand systems indefinitely without meaningful adaptation. This rhetorical strategy helps present standardized approaches as natural while obscuring the enormous effort required to construct and maintain them.

Meta's content moderation pivot can be understood as such a project, attempting to transform complex, relational elements of online discourse into what Tsing (2012) calls "nonsoels," the "nonsocial landscape elements" that can be managed as discrete, interchangeable units (p. 508). By separating "high-severity" from "lower-severity" harms, replacing expert judgment with community oversight, and relocating operations, Meta's proposed updates treat social phenomena as self-contained units. The new approach

Copyright © 2026 (Kyooeun Jang, kyooeunj@usc.edu). Licensed under the Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd). Available at https://ijoc.org.

positions content moderation as functioning without the transformative relationships that construct harm, delineate context, and complicate enforcement.

In this article, I ask three interrelated questions: How do Meta's shifts toward community-based moderation and increased political content reveal contradictions in scalable platform governance? How does bifurcating content into "high-severity" and "lower-severity" harms address or obscure the nature of interconnected online harms? What do spatial relocation and regulatory resistance reveal about the stability and sustainability of platform safety standards over time?

## Community-Based Moderation and Political Content: Participation Versus Epistemic Authority

Meta's announcement to replace fact-checking with crowdsourced decision making reveals the company's understandings about scale: that crowd size ensures representativeness, that faster decisions improve operations, and that open processes yield accountability. Some Trust and Safety experts support this view, finding traditional fact-checking to be too slow, resource-intensive, and jurisdictionally complex (Yee Carter, as cited in Soeth, 2025). This framing of "biased experts" versus "democratic community," however, constructs dichotomies between epistemic authority and participatory legitimacy. Research on X's Community Notes, for example, shows that displayed notes are driven more by consensus than accuracy, with less than 30% of fact-checkable tweets rated as helpful during the 2024 elections (Mahadevan & Mantzarlis, 2024).

This shift risks transforming fact from a matter of verifiability into one of popularity, potentially undermining trust in knowledge-producing institutions. The consequences extend beyond the United States. Many fact-checking organizations in countries such as Argentina, Brazil, and Pakistan, were integrated into Meta's third-party fact-checking program and relied on its support (Bhattacharya, 2025). In much of the Global South, social media platforms serve as primary news sources (Newman et al., 2024). The program's rollback has left fact-checking outlets facing operational uncertainty, jeopardizing their ability to hold public officials accountable (Huo & Bond, 2025). The risks are particularly heightened in contexts like Myanmar, where Facebook-amplified hate speech fueled violence against the Rohingya, and the Philippines, where citizens depend heavily on Meta's platforms for news (Arya & Kanozia, 2025).

Meta's revival of political content further complicates this project. In 2021, the company claimed users wanted less political material, yet it is now reintroducing it (Frenkel & Isaac, 2024). This shift reveals tensions between engagement-driven visibility and civic discernment, potentially reshaping users' right to be informed into exposure to low-quality, politicized material. Just as algorithmic recommendation systems that actively structure information environments cannot be regarded as neutral channels of distribution, engagement metrics cannot be treated as neutral indicators of relevance or public interest.

These contradictions stem from an assumption that moderation can be scaled through technological standardization, one that does not fully consider the inherently diverse and contextual nature of harmful content. After years of investment in professional moderation, the company now positions community notes as both a cost-saving measure and a more democratic alternative involving user participation. In practice,

https://doi.org/10.65476/0zpmnf83

however, this approach relies on the rhetoric of participation while simultaneously withdrawing the resources required for meaningful oversight.

Effective platform governance acknowledges that certain elements of content moderation, such as contextual judgment, domain expertise, and cultural understanding, resist standardization. Research shows that community participation can detect misinformation patterns missed by centralized teams (Micallef et al., 2022) but only when supported by demographic representation and algorithmic transparency. The most effective models integrate community participation with professional expertise rather than substituting one for the other (The International Fact-Checking Network, 2025). For example, community notes can be used to address routine misinformation while professional fact-checking can be reserved for high-risk, politically contested content (Arya & Kanozia, 2025). Community notes are trusted across political affiliations, and their effectiveness derives less from crowdsourcing itself than from the explanatory context they provide alongside expert judgment (Drolsbach et al., 2024).

**Legibility Versus Nuance: The Failure to Recognize the Interconnected Nature of Harm**

Meta's recent updates, particularly the loosening of policy restrictions while redirecting automation toward "high-severity" violations, reflect a belief that online harms are discrete and separable. By "scaling-down" restrictions and "scaling-up" detection, Meta's approach equates safety with legibility at scale (Scott, 1998), treating harms as manageable only when simplified into standardized categories. This framing risks compartmentalizing interconnected ecosystems of harm. For instance, Meta's revised hate speech policy now permits posts likening women to "household objects" or LGBTQ+ people to the mentally ill (Huo & Bond, 2025). Such adjustments have immediate consequences for marginalized groups, reproducing biases that frame dangerous stereotypes as merely offensive (Noble, 2018). Standardization over contextualization also obscures coded racial harassment, particularly against Black and Indigenous users (Matamoros-Fernández & Farkas, 2021).

This approach creates a false trade-off between nuance and legibility. Projects of legibility simplify complex practices for administrative ease rather than reflect lived realities (Scott, 1998). While Meta portrays simplification as expanding expression, this framing misrepresents how these values function in practice, where the meaning and safety of speech depend on contextual nuance. Content Meta classifies as "lower-severity" (hate speech, political misinformation, and targeted harassment) requires context-sensitive intervention, since harmfulness depends on situation, target, and cumulative effects. Without contextual judgment, "freer" expression can silence voices and distort democratic processes. Artificially separating these approaches rather than integrating them may privilege operational efficiency over the protection for marginalized users, revealing a limited account of the relational nature of online communication.

Concentrating resources on automated detection of "high-severity" harms like terrorism and child abuse instead of "lower-severity" harms also underestimates how serious harms often emerge cumulatively from smaller violations. Research has shown that cyberbullying erodes mental health and can contribute to depression and heightened suicide risk, especially among youth (Sumner et al., 2021). Cyberbullying appears nearly twice as often in the digital activity of youth who later attempt suicide, and risk escalates sharply when combined with violent or depression-related content (Sumner et al., 2021). Paradoxically, this

process undermines the very populations Meta claims to prioritize under its "high-severity" framework. In reality, these harms operate as an interconnected system where minor violations can compound and amplify one another. Meta's compartmentalized approach risks misconstruing this reality, treating posts as isolated rather than part of broader dynamics. Effective moderation requires recognizing context and cumulative dynamics and that nuance and expression must be understood as interdependent rather than opposed.

This approach reveals the tension between ethical responsibility and procedural fairness. While ethical responsibility prioritizes preventing harm to vulnerable communities, procedural fairness prioritizes consistent rule application regardless of context. Although Meta's classification system often categorizes precursor communications as "lower-severity," celebrity suicides in East Asia from persistent hateful comments exemplify how "lower-severity" speech harms can lead to lethal consequences (Murakami & Park, 2020; Yi & Cha, 2019). Similarly, research on the January 6th Capital riots traced how "Stop the Steal" rhetoric moved from occasional (small-reach) posts questioning election integrity to coordinated mobilization culminating in real-world violence (Freelon et al., 2022). Platforms prioritize procedural fairness, however, because of its ability to scale more efficiently and provide a defensible posture under public and regulatory scrutiny.

As plantation agriculture creates monocultures that suppress ecological diversity (Tsing, 2013), Meta's moderation strategy imposes standardization in ways that fail to capture the interconnected nature of online harms. Just as Matsutake mushrooms resist cultivation because they depend on complex forest relationships, online harms cannot be separated from context without losing essential meaning (Tsing, 2012). By stripping away these "transformative relationships" between content, context, and impact, Meta makes moderation technically manageable but contextually limited, producing scalable rules that struggle to address how harms emerge within social ecosystems (Tsing, 2012, p. 507).

Meta's approach has economic justifications, as it reduces the complexity of content moderation to standardized processes applied across billions of users. However, as platforms scale and contexts become more diverse, sustaining such comprehensive moderation becomes increasingly untenable. Research on automated moderation systems shows that while platforms frame algorithmic solutions as essential for economic viability, these systems consistently struggle to address complexity (Gorwa et al., 2020). The sheer volume of content has made automation appear inevitable, yet these technologies remain opaque and publicly unaccountable, producing a governance framework that prioritizes efficiency over effectiveness.

## Platform Sovereignty Versus Democratic Accountability: Volatile Governance That Undermines Consistency

Meta's geographic relocation of Trust and Safety operations from California to Texas, alongside Zuckerberg's support for the Trump administration's resistance to foreign regulations, illustrates how platforms attempt to leverage governance volatility to manage their online ecosystems. By framing political volatility as justification for reshaping moderation, the platform creates dichotomies between "biased" versus "unbiased" territories, substituting geographic symbolism for substantive reform.

https://doi.org/10.65476/0zpmnf83

This geographic strategy reveals tensions in Meta's neutrality claims, as substituting one set of political associations for another can avoid contextual judgments necessary to address how different groups experience harm online. For instance, Meta's recent rollback of its *Hateful Conduct* policy now permits "allegations of mental illness or abnormality when based on gender or sexual orientation," justified as protecting "political and religious discourse about transgenderism and homosexuality" (Meta, 2025, para. 5). The company's resistance to foreign regulation further exposes how "neutrality" functions as a political position rather than an absence of politics, one that prioritizes operational flexibility over consistent safety standards. As Michael McConnell, cochair of Meta's Oversight Board, noted, the timing makes these changes appear as "buckling to political pressure" rather than reflecting principled reform (Huo & Bond, 2025, para. 7).

Meta's relocation also applies to select forms of labor, affecting primarily full-time employees based in Silicon Valley rather than the thousands of outsourced content moderators working in low-wage regions who perform the invisible frontline labor of content moderation. This selective spatial strategy reveals a broader tension in commercial content moderation, in which major platforms outsource the morally demanding labor of moderation to low-wage workers in the Global South (Roberts, 2019) while practicing geopolitical pragmatism at the corporate level, relocating operations to jurisdictions like Texas, Dublin, or Manila to minimize political friction, corporate tax burdens, and labor costs.

Meta's resistance to non-U.S. regulatory frameworks extends this precariousness globally. The EU's Digital Services Act emphasizes transparency, user rights, and external audits, treating platforms as public infrastructures requiring democratic oversight. In contrast, Meta's announcement positions the company as seeking to retain discretion over moderation standards, enforcement logics, and system transparency, a posture that emphasizes agility and proprietary control while treating external regulation as a compliance burden rather than a democratic necessity. This legitimacy derived from market dominance and global reach, however, comes at a cost to institutional accountability, as it treats platforms as quasi-sovereign digital entities entitled to define their own norms with minimal constraint. By scaling protection through proprietary rules, the platform aligns with U.S. deregulatory preferences while diverging from the EU's regulatory model, fragmenting governance across jurisdictions, and destabilizing democratic accountability globally. Through geographic relocation and regulatory resistance, Meta may expose its governance structures to increased volatility, privileging corporate discretion over public oversight and framing content moderation as a relocatable technical issue rather than a governance challenge requiring robust public institutions.

## Conclusion

Meta's January 2025 content moderation pivot reveals fundamental contradictions in platform scalability projects. First, community-based moderation and the expansion of political content create tensions between epistemic authority and democratic participation, shifting facts from matters of verification to matters of engagement, while algorithmic curation continues to determine what content is eligible for evaluation in the first place. Second, the bifurcation of speech into "low-severity" and "high-severity" violations treats interconnected harms as discrete categories, producing trade-offs between nuance and legibility that privilege operational efficiency over contextual understanding and overlook how harms

accumulate and evolve. Third, by geographically relocating moderation labor and resisting regulatory oversight, platform governance becomes volatile, privileging corporate discretion over democratic accountability and reinforcing U.S. exceptionalism at the expense of other jurisdictions' frameworks.

These contradictions are not incidental implementation failures but structural consequences of the mismatch between the standardized, discrete units demanded by scalability and the inherently relational, contextual nature of online harms. Governance strategies that scale through simplification paradoxically generate tensions that exacerbate the very harms they claim to mitigate, fragmenting safety frameworks into disparate and often conflicting interpretations rather than sustaining coherent and defensible standards.

## References

Arya, R., & Kanozia, R. (2025). Meta's misguided path: Global consequences of abandoning third-party fact-checking. *Media Asia*, 1–25. Advance online publication. https://doi.org/10.1080/01296612.2025.2467893

Bhattacharya, A. (2025, January 10). *Global fact-checkers were disappointed, not surprised, Meta ended its program.* Rest of World. https://restofworld.org/2025/meta-drops-fact-checking-partnerships-global-watchdogs-scramble/

Drolsbach, C. P., Solovev, K., & Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media. *PNAS Nexus, 3*(7), 217. https://doi.org/10.1093/pnasnexus/pgae217

Freelon, D., Marwick, A., & Kreiss, D. (2022). False equivalencies: Online activism from left to right. *Science, 369*(6508), 1197–1201. https://doi.org/10.1126/science.abb2428

Frenkel, S., & Issac, M. (2024, September 24). How Meta distanced itself from politics. *The New York Times.* https://www.nytimes.com/2024/09/24/technology/meta-election-politics.html

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), 1–15. https://doi.org/10.1177/2053951719897945

Huo, J., & Bond, S. (2025, January 10). *Meta built a global fact-checking operation. Will it survive?* NPR. https://www.npr.org/2025/01/10/nx-s1-5252738/meta-fact-checking-international

The International Fact-Checking Network. (2025, January 28). *An open letter to Mark Zuckerberg from the world's fact-checkers, nine years later*. Poynter. https://www.poynter.org/ifcn/2025/an-open-letter-to-mark-zuckerberg-from-the-worlds-fact-checkers-nine-years-later/

https://doi.org/10.65476/0zpmnf83

Kaplan, J. (2025, January 7). *More speech and fewer mistakes*. https://www.facebook.com/zuck/videos/its-time-to-get-back-to-our-roots-around-free-expression-were-replacing-fact-che/1525382954801931/

Mahadevan, A., & Mantzarlis, A. (2024, November 13). *X's community notes had a marginal, if any, impact on the election*. Poynter. https://www.poynter.org/commentary/2024/x-community-notes-role-2024-presidential-election/

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media, 22*(2), 205–224. https://doi.org/10.1177/1527476420982230

Meta. (2025, January 7). *Hateful conduct*. Transparency Center. https://transparency.meta.com/policies/community-standards/hateful-conduct/

Micallef, N., Armacost, V., Memon, N., & Patil, S. (2022). True or false: Studying the work practices of professional fact-checkers. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW1), Article 127. https://doi.org/10.1145/3512974

Murakami, S., & Park, J.-M. (2020, June 3). *Death of reality TV show star in Japan spotlights cyber bullying*. Reuters. https://www.reuters.com/article/lifestyle/death-of-reality-tv-show-star-in-japan-spotlights-cyber-bullying-idUSKBN23A0SM

Newman, N., Fletcher, R., Robertson, C. T., Ross Arguedas, A., & Nielsen, R. K. (2024). *Reuters Institute digital news report 2024.* Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024

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

Roberts, S. T. (2019). *Behind the screen.* New Haven, CT: Yale University Press.

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven, CT: Yale University Press.

Soeth, M. (Host). (2025, March 12). *Navigating fact checking* (Feat. Yee Carter) [Audio podcast episode]. In *Safety is sexy*. https://www.matthewsoeth.com/safety-is-sexy-podcast/fact-checking-co

Sumner, S. A., Ferguson, B., Bason, B., Dink, J., Yard, E., Hertz, M., Hilkert, B., Holland, K., Mercado-Crespo, M., Tang, S., & Jones, C. M. (2021). Association of online risk factors with subsequent youth suicide-related behaviors in the US. *JAMA Network Open, 4*(9), e2125860. https://doi.org/10.1001/jamanetworkopen.2021.25860

Tsing, A. L. (2012). On nonscalability: The living world is not amenable to precision-nested scales. *Common Knowledge, 18*(3), 505–524. https://doi.org/10.1215/0961754X-1630424

Yi, H. Y., & Cha, S. (2019, November 28). *Cyber bullying, star suicides: The dark side of South Korea's K-pop world*. Reuters. https://www.reuters.com/article/world/cyber-bullying-star-suicides-the-dark-side-of-south-koreas-k-pop-world-idUSKBN1Y20U7

https://doi.org/10.65476/0zpmnf83