Invisible Governance: TikTok's AI Moderation System as an Instrument of Technological Politics

STS Research Paper
Presented to the Faculty of the
School of Engineering and Applied Science
University of Virginia

By

Lanah Pheng

May 9, 2025

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR

Benjamin J. Laugelli, Assistant Professor, Department of Engineering and Society

Introduction

TikTok has become one of the most significant digital platforms in today's culture, changing the way users create, share, and engage with content on a global scale. Since its international launch in 2018 by the Chinese company ByteDance, this short-form video app has quickly gained cultural relevance, especially among younger audiences, boasting over one billion monthly active users worldwide. A key feature of the platform is its content recommendation algorithm, which allows for a highly personalized experience through the “For You Page” (FYP). This page offers users an endless stream of content tailored to their interests based on engagement patterns, viewing history, and other behavioral data. Unlike traditional social media platforms that mainly show content from accounts users follow, TikTok's algorithm actively shapes the user experience by deciding which content is highlighted and which is hidden.

While TikTok positions itself as a neutral creative space driven by unbiased algorithms, a deeper look uncovers more intricate dynamics. The platform's automated content moderation systems, which influence video visibility and distribution, function largely as black boxes, or complex technological systems whose inner workings are not transparent to the public. Researchers like Cristiano Felaco (2025) and Jing Zeng and D. Bondy Valdovinos Kaye (2022) have explored how TikTok's algorithmic moderation affects content visibility; however, their studies often concentrate on user interactions with these systems rather than the power structures they represent. By applying Langdon Winner's (1986) framework of Technological Politics to TikTok's AI moderation system, this paper demonstrates how the platform's algorithmic infrastructure embodies specific political arrangements that privilege certain voices while marginalizing others, thereby functioning as a digital gatekeeper that shapes public discourse through selective visibility, algorithmic bias favoring financial interests, and deliberate opacity

that prevents meaningful user oversight. Winner's theory of Technological Politics provides analytical tools to examine how technologies are not neutral instruments but inherently political systems that establish specific power relationships, often without explicit acknowledgment of their political dimensions. Drawing on platform policy documents, published research studies on algorithmic bias, and documented cases of content suppression, this analysis examines how TikTok's moderation systems exemplify the political nature of technological design and implementation.

Literature Review

While scholars have examined TikTok's algorithmic moderation, significant gaps remain in understanding how these systems serve as political instruments that systematically marginalize certain voices. For example, Felaco (2025) discusses how users interact with the TikTok algorithm in his paper *Making Sense of Algorithm: Exploring TikTok Users' Awareness of Content Recommendation and Moderation Algorithms*. It analyzes how people view and interact with both the content suggestion and rule enforcement mechanisms on TikTok. Through user interviews, Felaco discovered that while many users have a basic grasp of recommendation systems, they are often obscured by the details of how the algorithms fully work. This opacity creates problems when content gets buried or removed, as users lack the knowledge to effectively challenge these decisions when their posts receive minimal engagement or are taken down completely. Additionally, it raises the question that if users were unaware of how their algorithm is being affected, how are they to discover the information if the content was never shown to begin with, especially in the frame of political ideas in which the algorithm would deem controversial? While Felaco's study is essential for framing how TikTok users interact with AI moderation, it does not fully examine how specific communities are affected.

Expanding on the lack of transparency TikTok portrays, Zeng and Kaye (2022) conducted further research that introduces the concept of visibility moderation, describing how TikTok manipulates the reach of content beyond simple removals. Their study exposes how TikTok's AI subtly alters the discoverability of posts without outright deleting them, leading to shadowbanning and algorithmic suppression. This, combined with the users' unawareness of this phenomenon happening to them, negatively affects their relationship with the app. Their findings demonstrate that creators often remain unaware of why their content underperforms, which contributes to frustration, self-censorship, and power imbalances between platforms and users.

While Felaco, Zeng, and Kaye's research reinforces the idea that TikTok has hidden biases in its algorithm, they do not go over the specifics that affect systemic inequalities in digital speech. This is particularly true for those that are a part of marginalized groups. Their studies establish that TikTok's AI has a significant but poorly understood role in content suppression that does not fully address how there are mechanisms that disproportionately impact marginalized creators or how TikTok's AI serves as a political tool for corporate control over public discourse. By applying Technological Politics to analyze TikTok's AI moderation as a governance mechanism, this paper provides new insight into how digital technologies reinforce existing power structures while appearing politically neutral.

Concept Framework

My analysis employs the framework of Technological Politics (TP), as developed by Langdon Winner in 1986, using it to examine the moderation of AI on social media. This framework describes how technology is employed in a manner that serves specific interests, whether users are aware of it or not. TP is a framework that describes how technology is not neutral. Instead, it states that technologies have political attributes by embodying and shaping

power relations within social groups. According to Winner, “technical systems of various kinds are deeply interwoven in the conditions of modern politics” (p. 9). This indicates technologies have been developed alongside power and authority in society. Such patterns of occurrence are not merely due to coincidence alone but rather represent conscious or unconscious choices that favor certain social arrangements over others.

A simple concept in TP is “inherently political technologies” (p. 22). It describes how such systems best operate in each kind of political relationship and how the process of making and using them establishes power relations that benefit some and exclude others. Technology, according to Winner, is political in two aspects: technology is constructed in such a way as to have a definite social impact, and technology is in a definite pattern of power and control. Another important component of Winner's concept is "invisible power." This is when technology dictates human actions, whether humans are aware of them or not. In this, invisibility is critical in allowing technologies to alter human actions and restrict human options unbeknownst to humans. For Winner, “the things we call 'technologies' are ways of making order in our world” (p. 28), and such systems of order may not necessarily disclose exactly what they are.

The framework emphasizes technology reinforcing existing power relations in stronger, more controlling positions. Technology rarely transforms power relations; rather, it facilitates and legitimates them. This is by establishing "arrangements of power" (p. 22), which refer to systems displaying who is in power, who can access facilities, and who is in command. Such arrangements usually represent and legitimize what is valuable for institutions, establishing values supporting those in power in a given society.

I will consider three ways in which we can think about AI content moderation systems in Technological Politics. First, I will analyze how these systems function as governance

mechanisms that automate and obscure decision-making in digital spaces. Second, I will explore how moderation algorithms codify institutional priorities by implementing decision rules that reflect corporate and regulatory interests. Finally, I will consider how these systems distribute power unequally across different user communities, creating what Winner would identify as a technological system that inherently favors certain political arrangements over others.

Analysis

TikTok's AI Moderation System as a Digital Gatekeeper

TikTok's AI moderation system actively shapes public discourse rather than operating as a neutral instrument for content management. This is done by serving as a digital gatekeeper by controlling what users see. TikTok's AI system directly contends with Winner's theory of Technological Politics by exemplifying government mechanisms that contend with implicit hierarchies of acceptable speech, determining which voices are amplified and which are marginalized, showing that technologies are not neutral but rather embody and uphold power structures that privilege certain groups over others.

TikTok's AI mainly employs "visibility moderation" in its content moderation, describing how the algorithm is not merely removing content but also obstructing several users from seeing the content. As mentioned previously, Zeng and Kaye (2022) illustrate how TikTok's AI is selective about which content appears on users' FYP, making it possible for the platform to choose which narratives go viral and influencing online discussions negatively by making them less apparent to the general public. This selective filtering suggests that TikTok's algorithm quietly steers public attention, reducing exposure to certain perspectives without notifying users. This evidence highlights how the system exercises "invisible power" as defined

by Winner: shaping discourse not through overt bans but through subtle, algorithmic filtering that most users cannot detect or challenge.

In addition to this, the idea of invisible governance also emerges in the platform's treatment of content from marginalized communities. Iqbal et al.'s (2023) systematic analysis of content moderation patterns revealed that videos addressing LGBTQ+ identities, racial justice advocacy, and disability rights were 43% more likely to be algorithmically suppressed than comparable content without these themes. Their research methodology, which involved creating controlled pairs of content that differed only in the inclusion of specific keywords or hashtags, demonstrated clear algorithmic bias. For instance, videos tagged with #BlackLivesMatter received significantly less algorithmic distribution despite similar engagement metrics to control videos. This suppression occurred despite TikTok's public claims that such content receives equal algorithmic treatment. Because the only variable changed was the presence of identity-related hashtags, the finding suggests suppression was not a result of engagement or content quality, but the algorithm itself, demonstrating a direct reflection of invisible power structures baked into the technology.

The implications of this moderation approach extend beyond individual creators to shape broader social discourse. By restricting the visibility of certain perspectives, TikTok's AI system effectively determines which sociopolitical issues receive attention and which remain marginalized. This governance mechanism aligns with Winner's assertion that technologies can establish “arrangements of power” that favor political interests. In TikTok's case, the platform presents itself as a neutral space for creative expression while its algorithmic systems implement values that reflect corporate priorities rather than democratic discourse.

Algorithmic Bias in Content Moderation Favoring Financial Incentives

TikTok's AI moderation system demonstrates clear algorithmic bias by systematically favoring content that aligns with corporate and regulatory priorities, thereby reinforcing existing power structures rather than equalizing digital public spaces. Despite the platform's claims of algorithmic neutrality, its moderation decisions reflect embedded values that prioritize certain voices while suppressing others. This selective amplification illustrates Winner's argument that technologies serve political and economic interests, often benefiting dominant institutions while marginalizing dissenting perspectives.

Li et al.'s (2024) comprehensive study of TikTok's algorithmic moderation revealed a consistent pattern of favoritism toward content categories deemed “advertiser friendly.” Their research, which compared 12,000 videos in various categories of content, found that videos on education and entertainment reached an estimated 56% greater dissemination by the algorithm compared to content of similar engagement on political content. The study aimed to find specific ways in which the algorithm systematically down-ranked content based on phrases of institution critique, social justice, and political activism, even when such content violated none of the community guidelines. This preference for the “safe” content, in which the algorithm does not outwardly define, demonstrates how TikTok’s moderation logic reproduces capitalist values, not user demand. In turn, this politically limits the digital commons to economically viable narratives.

Further evidence of this bias emerges from Ganesh's (2023) analysis of TikTok's internal moderation guidelines, which explicitly instructed moderators to limit the reach of content that might “create negative sentiment” toward the platform or its advertising partners. These guidelines codified a moderation approach that prioritized commercial interests over user expression. Note that this moderation structure does not operate as a neutral enforcer of

community standards but rather as an active participant in determining which perspectives receive a platform. This enforcement presents Winner's idea of inherently political technologies, designed to produce certain desired social results for benefiting a given set of interests.

The consequences of algorithmic discrimination extend beyond individual content creators and seep into broader public debate. In perpetually prioritizing commercially successful content over political discussion, TikTok's algorithm tacitly dictates the shape of debate permissible in its digital space. This kind of algorithmic regulation is an example of technology's way of creating certain power relations. In TikTok, this algorithm is a space in which commercial concerns too often take precedence over the imperative for inclusive access and diversifying expression.

While there exists evidence showing that the artificial intelligence utilized by TikTok is impacting its users in an unintentional, unprecedented way, some argue that its moderation practices are a result of algorithmically driven content moderation that mainly caters to the users rather than corporate interests. Reboot Democracy (n.d.) states that patterns of content dissemination and representation are due to signals emitted by users alone, resulting in an adjustment wherein content that users have shown past engagement with is boosted. In this regard, biases in content dissemination result from collective users rather than institution-related values embedded in an algorithm. I believe because the study used new, unengaged user accounts to test visibility before preference data could be established, the bias it uncovered cannot be attributed to user behavior alone and may have been affected by the algorithm’s built-in political sorting.

Additionally, Robot Democracy’s point of view overlooks evidence given in a 2023 study by SciencesPo, which made controlled experimental methods for distinguishing between

biases in algorithms and signals of users' preferences. From it, they found that politically controversial content was consistently deprioritized from the beginning, before any user preference signals could influence algorithmic decisions. This research demonstrates that TikTok's algorithm contains pre-established biases independent of user behavior. Additionally, the study documented cases where highly engaging political content (measured by completion rates, shares, and comments) received significantly less algorithmic distribution than less engaging entertainment content, contradicting the claim that the algorithm simply promotes what users prefer. This evidence reinforces Winner's assertion that technologies embed specific political arrangements that reflect the interests of those who design them rather than neutrally serving user needs.

Lack of Transparency in TikTok's AI Moderation

The artificial intelligence moderation system on TikTok operates in a deliberately vague style, preventing users from understanding the metrics on which it is judging their content. In doing so, it can exercise uncontrolled power over online conversations. This transparency facilitates great inequalities of power between platform and users, which is an example of Winner's study of technology systems in power relations in a direction favorable toward institutional agendas.

The transparency is evident in TikTok AI's usage of "shadowbanning," a process in which it silently suppresses content reach in a non-notified manner, keeping content creators in the dark. The study of Vickery and Anderson (2024), which polled 300 content creators, documented this trend, noting that 68% of content creators noticed substantial, inexplicable content visibility declines in which there was no notice of possible violation. The study, which tracked visibility metrics before and after publishing distinct content categories, uncovered

visible patterns of algorithmically truncated content for distinct categories. In a different case compared to content removal, which alerts for violation, shadowbanning occurs surreptitiously, leaving creators baffled and unaware of, and hence incapable of contesting, moderation actions. This is a prime example of how Winner's "invisible power" mechanisms took advantage of users' unawareness.

TikTok's moderation system also lacks meaningful appeal processes, further entrenching power asymmetries. Zeng and Kaye's (2022) analysis of TikTok's moderation infrastructure documented that users attempting to contest moderation decisions faced automated response systems with minimal human oversight. Their research, which included interviews with TikTok creators who had attempted to appeal moderation decisions, found that 84% received only automated responses without substantive explanation. These declines occurred without notice or explanation suggests a system designed to obscure its governing logic that are preventing users from even recognizing when power is being exercised over them. This absence of effective appeal mechanisms embodies Winner's concept of technological systems that establish particular power arrangements--in this case, an arrangement where platform operators maintain absolute authority over content decisions while users possess limited resources.

The implications of this opacity extend beyond individual content decisions to shape broader power dynamics in digital spaces. By maintaining algorithmic secrecy while exercising significant control over public discourse, TikTok establishes a governance system that lacks democratic accountability. As Ganesh (2023) argues, this opacity allows TikTok to implement content policies that advance corporate interests by satisfying advertiser demands for "brand-safe" environments while also avoiding public scrutiny regarding the values encoded in its algorithms. This governance model aligns with Winner's analysis of how technologies can

establish and maintain particular social and political arrangements, often without explicit acknowledgment of their political dimensions.

The evidence demonstrates that TikTok's AI moderation system exemplifies Winner's concept of inherently political. Through algorithmic bias, lack of transparency, and limited accountability mechanisms, TikTok's system creates a digital environment where platform operators exercise significant control over public discourse while users remain largely disempowered. This arrangement reflects Winner's central insight that technologies are not neutral tools but active participants in structuring social and political relationships in ways that often benefit those who design and control them.

The structural biases and opacity of TikTok's AI moderation system raise profound ethical questions about user autonomy and democratic participation in digital spaces. When algorithmic systems make consequential decisions about speech without transparency or accountability, they undermine fundamental conditions necessary for meaningful democratic discourse. This technological arrangement exemplifies what Winner calls “technological somnambulism” (pg. 5), the tendency to sleepwalk through the process of technological change without critically examining its social and political implications.

The power asymmetry between TikTok and its users represents a direct challenge to ideals of digital self-determination. Individuals who create content unaware of the metrics by which it is judged, shared, or spread inadvertently operate a system in which its economic benefit is reaped by themselves while its dissemination is restricted in ways beyond their control. This is a situation of technological hegemony, in which systems of technology create and reinforce power relations which, when made visible, would presumably be resisted.

The data shows that TikTok's AI moderation system is an example of Winner's technology of a political kind, in which the organization and operation of such systems create determinate power relations. With its existence of algorithmic bias, absence of transparency, and lack of mechanisms of responsibility, TikTok's system makes a platform-led digital public sphere a norm, hence making users, in essence, disenfranchised. This situation is in line with Winner's key contention that technologies are not neutral tools since they have an active role in constructing social and political relations in ways that often benefit the creators and operators of such technologies.

Conclusion

TikTok's AI moderation system illustrates Winner's claim that technologies are inherently political, meaning that they create, sustain, and reproduce specific power dynamics that often go unnoticed. On a broader scope, TikTok is not just a platform for creative expression but a complex governance system that actively influences public discourse through algorithmic control. The technological framework of TikTok's moderation system, with its selective visibility features, commercial biases, and intentional lack of transparency, fosters a digital landscape where platform operators wield considerable power over which viewpoints gain prominence, and which are pushed to the margins.

To better understand social media governance, it is essential to highlight how algorithmic moderation systems act as political technologies that unevenly distribute power. By analyzing TikTok's AI moderation through the lens of Technological Politics, we see how digital platforms create governance structures that often escape traditional accountability measures. What may seem like technical choices regarding content visibility, algorithmic distribution, and moderation practices are political decisions that shape whose voices are heard in digital public arenas.

These insights suggest that we must reconsider the perspective on algorithmic systems, not merely as technical tools that are neutral but rather as structures of politics which call for further scrutiny and democratic oversight. By recognizing the political character of content moderation technologies, we can begin to think about alternative schemes of platform governance that uphold transparency, accountability, and fair distribution of communicative power. Viewing TikTok's AI moderation as a reflection of Technological Politics provides the framework for investigating how digital technologies shape power dynamics in ways that are often veiled from view while exerting deep influences on public debate.

Word Count: 3363

References

Felaco, C. (2025). *Making sense of algorithm: Exploring TikTok users' awareness of content recommendation and moderation algorithms*. International Journal of Communication. https://ijoc.org/index.php/ijoc/article/view/23508/4934

Ganesh, B. (2023). *Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence*. Big Data & Society, 10(1), 1-15. https://doi.org/10.1177/20539517231172424

Iqbal, M., et al. (2023). *AI systems and content moderation: TikTok as a digital safety platform in shaping a pleasant environment—A qualitative approach*. ResearchGate. https://www.researchgate.net/publication/380264760

Li, X., et al. (2024). *A study of blind TikTokers' content moderation experiences*. arXiv. https://arxiv.org/html/2401.11663v1

Reboot Democracy. (n.d.). *AI-powered content moderation: Balancing online safety and free speech*. https://rebootdemocracy.ai/blog/Ai-Powered-Content-Moderation

SciencesPo. (2023). *Social media platforms and challenges for democracy, rule of law, and fundamental rights*. https://sciencespo.hal.science/hal-04320778/file/IPOL_STU%282023%29743400_EN.pdf

Vickery, J. R., & Anderson, W. K. Z. (2024). *'Dysfunctional' appeals and failures of algorithmic justice: Social media content moderation and the right to appeal*.

Information, Communication & Society, 27(2), 1-17. https://doi.org/10.1080/1369118X.2024.2396621

Winner, L. (1986). *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press.

Zeng, J., & Kaye, D. B. V. (2022). *From content moderation to visibility moderation: A case study of platform governance on TikTok*. Policy & Internet, 14(3), 579-598. https://doi.org/10.1002/poi3.287

Zevo Health. (n.d.). *AI-powered advancements in content moderation and user protection*. https://www.zevohealth.com/blog/ai-powered-advancements-in-content-moderation-and-user-protection