CCOMM

Proceeding of Creative and Collaborative Communication Conference 2025
https://proceeding-ccomm.petra.ac.id/

An Examination of TikTok's Content Governance: Alignment Between Policy and Practice

Richard Suyanto
Queensland University of Technology

ABSTRACT

This paper explores TikTok's platform-level content governance by assessing whether its moderation policies align with the actual outcomes of its "For You Page" (FYP) algorithm. Using a mixed-methods approach, analysing video engagement data and reviewing platform guidelines, we evaluate TikTok's claim that content relevance, not creator popularity, drives recommendations. Results show that while lesser-known creators can gain visibility, the algorithm strongly favours entertaining content with positive emotional tones, suggesting subtle biases. A cross-platform comparison with YouTube and Instagram reveals similar trends: all three platforms amplify polished, mainstream content while often marginalizing niche or socially significant voices. Reports from marginalized creators, highlight concerns of algorithmic demotion and "shadow banning" (Eltaher et al., 2025). These findings raise questions about transparency, fairness, and accountability in algorithmic governance. With regulatory frameworks like the EU's Digital Services Act now in play, platforms face increasing pressure to make their recommender systems auditable and equitable. The paper concludes by calling for clearer policy enforcement, independent algorithm audits, and stronger safeguards to protect content diversity and creator equity.

Keywords: *TikTok, content governance, algorithmic bias, platform moderation, recommender systems.*

INTRODUCTION

Social media now reaches the majority of the global population, with over 5 billion users in 2024, and video content has become its most dominant form (Eltaher et al., 2025). TikTok, a short-video platform owned by ByteDance, exemplifies this trend. With roughly 1.58 billion monthly users worldwide, it has rapidly become a dominant force in the digital landscape, wielding significant influence over cultural trends, information dissemination, and online social interaction (Eltaher et al., 2025). The platform's immense reach and impact underscore the critical importance of understanding how content is governed.

Corresponding author:
Richard Suyanto - alejandre.richard21@gmail.com

A key aspect of content governance on platforms like TikTok is **algorithmic governance**. These complex systems analyse a multitude of signals to curate personalized experiences, such as the "For You Page" (FYP), and enforce content moderation policies. While algorithms offer unparalleled efficiency in managing the sheer volume of content, their inherent complexity raises crucial questions about their alignment with publicly stated policy objectives. As Mikalef et al. (2022,p.265) note, "the dark side of AI manifests itself through the intentional or unintentional negative outcomes that AI systems may generate". Researchers and regulators worry that this personalization model can create **filter bubbles**, amplify extreme or harmful content, and inadvertently suppress important voices. For instance, prior studies show that TikTok's recommender can quickly push youth toward troubling themes like eating disorders or extremist ideologies if initial signals align (Eltaher et al., 2025). Similarly, YouTube's "Up Next" system has been accused of guiding users toward radicalizing content (Eltaher et al., 2025).

Compounding these issues, **content moderation**, the policies and enforcement meant to remove harmful material, often fails to meet its stated goals. Although platforms like TikTok have published Community Guidelines banning hate speech, self-harm, and misinformation, studies suggest that moderation is inconsistent. An independent report found that TikTok's search engine often reproduces negative stereotypes about marginalized groups, and users from vulnerable communities report unequal treatment (Matlach et al., 2025). For example, several plus-size or black creators have voiced that their content is more frequently censored by "the algorithm" on Instagram (Willcox, 2025), while a survey by Haimson et al. (2021) found that Black and transgender individuals experienced a high rate of false-positive removals. These patterns suggest a significant gap between platform policies and the lived experience of users. The opacity of these algorithms has long troubled experts, as Pasquale (2015) and Gillespie (2018) argue that without transparency, users are left guessing how content is filtered and ranked.

In response to growing societal awareness of the power wielded by social media platforms, there are increasing demands for greater **platform accountability and transparency**. Platforms like TikTok often provide transparency centres and policy documents, but the true measure of accountability lies in the empirical investigation of whether these public policies are consistently reflected in the actual content curation and moderation practices. Therefore, in this expanded study, we systematically evaluate TikTok's content governance on both fronts: (a) algorithmic curation and (b) moderation outcomes. We compare these with analogous practices on YouTube and Instagram. Our analysis will be guided by several research questions: How do TikTok's recommendation patterns align with its stated emphasis on relevance? What types of content are most amplified or sidelined, particularly for marginalized voices? And how do our findings compare with known behaviours on other platforms? To address these, we will use a mixed-methods approach that includes quantitative analysis of video metrics and a review of platform guidelines. We will then discuss the implications for users and content diversity, and offer recommendations for

future research and policy measures to close the gap between what these platforms promise versus what they deliver.

LITERATURE REVIEW

Platform Governance and Policy Frameworks

To understand TikTok's content governance, it is essential to first review its publicly accessible policy documents, such as the **Transparency Center**, **Community Guidelines**, and **Safety Center**. These resources detail the platform's rules and standards, offering insights into its commitments to user safety, platform integrity, and human rights. Analysis of these documents reveals several key claims: TikTok asserts that its "For You" feed (FYF) algorithm prioritizes content relevance over creator popularity, aiming to level the playing field for new creators. The platform also commits to fostering an inclusive environment by explicitly prohibiting content that promotes discrimination or disparages individuals or groups. This aligns with broader research on content moderation, where platforms must decide how to handle not just "clearly problematic" content but also ambiguous "grey area" posts (Stockinger et al., 2023, p. 1228).

While platforms like TikTok strive to balance scale and accuracy in their policies, a significant **policy-practice gap** exists in both algorithmic curation and content moderation. This gap arises from the inherent challenges of managing content at an immense scale. As Gillespie (2020, p.1) points out, the push toward automated content moderation is a "necessary response to the scale" of social media, but these automated systems can fail to meet stated policy objectives. For example, a platform may forbid hate speech, yet enforcement remains imperfect. Researchers describe how ambiguous "grey zone" content can slip through automated filters, undermining the very policy goals they are meant to enforce (Stockinger et al., 2023). This can lead to unequal outcomes; for instance, marginalized creators often report their content is more frequently or unfairly censored. Absent clear oversight, platform rules can also reflect corporate interests, with automated censorship privileging commercially aligned content over minority viewpoints (Cobbe, 2021). These challenges are central to a broader academic discourse about platform power and the role of algorithms.

Academic and regulatory frameworks increasingly frame our understanding of platform governance. Scholars like Gillespie (2018) describe platforms as "custodians of knowledge" whose opaque rules shape public conversation, leading to widespread calls for greater **algorithmic accountability and transparency** (Gorwa et al., 2020). Regulators are now stepping in, with the European Union’s Digital Services Act (DSA, 2022) explicitly requiring very large platforms to disclose information about their recommendation systems and content moderation to independent auditors. This new regulatory environment demands systematic algorithmic audits of platforms like TikTok and its peers to enforce compliance (Mosnar et al., 2025). However, scholars caution that such audits must be reproducible and

longitudinal to be reliable, as single-instance audits can quickly become outdated as platforms adjust their algorithms (Mosnar et al., 2025). Furthermore, platforms operate under different ownership and legal regimes that influence governance; for example, TikTok faces unique scrutiny due to its Chinese ownership, while YouTube and Instagram (both U.S. companies) navigate First Amendment debates.

In sum, the governance literature frames our investigation: we look at TikTok through the lens of accountability and fairness, contextualized by its public policies, the challenges of enforcing them at scale, and the growing pressure from regulators. This provides a crucial framework for comparing its practices with those of YouTube and Instagram.

Algorithmic Curation and Recommender Systems

TikTok's algorithm is less documented than those of platforms like YouTube, which have been the subject of numerous audits. The "For You" feed (FYF) is fully algorithmic from the first video, utilizing a complex interest graph that users often perceive as "magical" and opaque (Register et al., 2023; Mosnar et al., 2025). The platform also uses audio transcription and content tagging, but discloses little publicly. As Annabell et al. (2025) note, even TikTok's new "search recommendations" feature is a black box, with the platform offering "almost no explanation" for its suggested terms.

Comparative research reveals how different algorithmic designs can lead to varied outcomes. For example, a 2024 U.S. election audit found that TikTok's recommender showed partisan asymmetries: accounts seeded with Republican content were fed significantly more conservative-aligned videos than Democratic accounts received liberal-aligned videos (Ibrahim et al., 2025). This indicates the algorithm interacts with content labels in complex ways with potential political implications. Similarly, a qualitative study on Instagram's shift to a Reels-heavy feed found that when the platform prioritized videos from unknown creators, many users, particularly women of colour, felt their community-focused informational posts were "subdued" or demoted (De, 2024). These examples illustrate a key theoretical point: recommender systems do not just passively reflect user demand; they actively shape what content is visible and who sees it (Bucher, 2012; Zhu & Lerman, 2016).

Further distinguishing these systems are the types of engagement signals they prioritize. While TikTok's algorithm heavily weighs internal factors like watch time and comments, YouTube's relies more on past watch history and subscriptions. Similarly, Instagram's Reels feed considers both user interactions and network-wide popularity signals. These design choices have tangible outcomes. For instance, research on user perceptions shows that Instagram's algorithm often favours vivid faces and high contrast, whereas TikTok and YouTube Shorts rely more on varied audio-visual features (Xue et al., 2025). Empirically, a classroom study found that TikTok's content reached target audiences far more effectively than Instagram's when posting similar content, suggesting TikTok's matching is "superior" for niche topics (Bishqemi & Crowley, 2022).

Content Moderation Debates

Content moderation, the policies and processes to remove or flag disallowed posts, is a second major axis of platform governance. While TikTok's Community Guidelines proscribe categories like hate and violence and the company reports removing a tremendous volume of content (over 153 million videos in Q4 2024) (Eltaher et al., 2025), critics point out that enforcement is neither transparent nor uniformly effective. The scale of moderation on a global platform relies on both automated tools and human reviewers, making mistakes and inconsistencies inevitable (Gillespie, 2020).

These moderation practices are at the centre of several key debates. One concerns **scope and transparency**. Historically, platforms have kept moderation criteria secret, but new regulations like the DSA are pushing for openness. Our preliminary review shows that while TikTok's policy documents are publicly available, they still offer sparse details on how rules are enforced, particularly for features like "safety in search" (Annabell et al., 2025). In contrast, YouTube and Meta (Instagram's parent company) have more extensive public documentation and publish quarterly reports.

A related debate revolves around **automation versus human judgment**. While automated tools can scan billions of videos, they may misclassify culturally specific content and introduce bias. As a result, marginalized users often report feeling algorithmically targeted. For instance, Instagram body-positivity influencer Nyome Nicholas-Williams launched the #IWantToSeeNyome campaign after her plus-size content was repeatedly removed, suggesting a systemic bias. Similarly, TikTok users have reported "shadow-bans" on hashtags related to trans identity or racial justice (Ibrahim et al., 2025). Empirical work confirms these patterns: a large survey of TikTok users found that Black and transgender respondents disproportionately experienced content removal or suppression compared to non-marginalized users (Ungless et al., 2024). This shows that even well-intentioned rules can be applied in biased ways, often harming vulnerable creators (Cobbe, 2021; Barocas et al., 2017).

These failures in moderation can lead to **harm amplification**, especially for young users. Short-video platforms are under scrutiny for exposing youth to disturbing content. A major audit of TikTok focused on minors found that simulated 13-year-old accounts quickly encountered videos related to suicide or eating disorders, within minutes of use, at alarming rates (Eltaher et al., 2025). Cross-platform studies have found that millions of posts on TikTok and Instagram involve self-harm or suicide, and that children frequently bypass age-gates (Xue et al., 2025; Eltaher et al., 2025). Such evidence highlights the tension between user engagement goals and safeguarding responsibilities.

In synthesizing these literatures, several key themes emerge for our analysis:

1. **Alignment:** we will investigate whether algorithmic and enforcement outcomes match the values expressed in platform policies, as the literature suggests they often do not, especially for marginalized groups (Willcox, 2025; Ungless et al., 2024).
2. **Comparative context:** we will frame TikTok's issues within a broader pattern of challenges seen across video platforms, including YouTube and Instagram (Ibrahim et al., 2025; De, 2024).
3. **Methodology:** our work will build on recent calls for more reproducible and multi-platform studies (Mosnar et al., 2025) by using actual TikTok data and policy analysis, bolstered by insights from comparable research.

METHODOLOGY

To assess the alignment between TikTok's stated policies and its actual content curation, we employed a **mixed-methods** approach combining quantitative data analysis with policy review. The rationale was to capture both the numerical patterns in TikTok's recommender and the formal rules it professes to follow. Below we detail the data, metrics, and analysis techniques, and also reflect on limitations and alternative approaches.

Data and Measures

We analysed a curated dataset of TikTok videos from a user in Australia that acquired in late 2024. The dataset includes a collection of 100 contents on FYP indexed by category. (These extremes were provided by a third party (tableau) to illustrate algorithmic reach and exposure.) For each video, we had metadata on the creator's follower count and counts of views, likes, comments, and shares. Additionally, each video had qualitative tags indicating its communicative *function* (e.g. Entertainment, Educational, Social) and the *emotion* conveyed (e.g. Surprise, Anger). These categorical labels were assigned by the dataset providers.

Key variables in our quantitative analysis were:

- **Follower count:** number of users subscribed to the creator. This represents popularity in policy discussions (TikTok claims FYP is not just driven by follower size).
- **Views, likes, comments, shares:** standard engagement metrics. We interpret these as proxies for algorithmic *promotion* (high values suggest the video was widely shown) and user response.
- **Content function and emotion:** categorical tags (such as entertainment/ humour, social commentary, love/compliments, etc.) and emotional tone labels. We used these to assess thematic patterns in what content succeeded or failed.

We computed correlation coefficients between follower counts and engagement metrics (views, likes, etc.) to evaluate the claim that TikTok's recommendation is *not* purely popularity-driven. A low correlation would indicate that even low-follower accounts can achieve high exposure (suggesting relevance-based curation). Conversely, a high correlation would imply popularity bias. We also performed descriptive comparisons: for example, we

compared the average views of videos in each content category, and examined the emotional profiles of top vs bottom videos.

To further contextualize these patterns, we juxtaposed them with known design features of other platforms. For example, on YouTube, prior work has shown that views scale strongly with subscriber count; thus, by contrast we highlight how TikTok differs or resembles this. Where possible, we also referenced policy documents: for TikTok, we reviewed the Community Guidelines and Transparency Center disclosures; for comparison, we looked at YouTube's and Instagram's published moderation reports and recommendations documentation.

Rationale for Mixed Methods

The choice of a mixed qualitative-quantitative strategy was intentional. Quantitative metrics capture *what* the algorithm amplifies, but policies and user reports capture *why* or *how* decisions are made. By combining them, we assess both sides of alignment. For example, if TikTok's community guidelines emphasize one kind of content (say, creativity or education) but our metrics show the algorithm mostly surface another (such as comedy), that indicates a misalignment. The review of other platforms practices serves a comparative role: it helps us interpret TikTok's patterns not in isolation but as part of a broader ecosystem of content governance.

Data Limitations and Validity

Several limitations must be noted. First, our TikTok dataset is aggregated and not a random sample of all content. Focusing on top-and bottom-performing videos highlights extremes but may not reflect the median case. This selection bias could exaggerate or understate correlations. For instance, if some small accounts in the "bottom 10" never intended to go viral, comparing them with top accounts might mislead. Ideally, a more comprehensive sample (or full scroll feed data for many users) would be used to minimize bias.

Second, causality cannot be inferred from correlations alone. A low follower-view correlation suggests that TikTok's algorithm can elevate obscure content, but it doesn't reveal *why* certain videos go viral. It could be due to content quality, hashtags, or sheer randomness. Similarly, our finding that entertainment videos tend to get more views does not prove TikTok explicitly favours humour; perhaps creators who tell jokes simply get more shares. To strengthen causal claims, one could employ experimental methods (e.g. controlled bot accounts with controlled behaviour, as in audit studies) or time-series analysis to see how newly posted videos rise in rank.

Third, our data lacked explicit markers for sensitive content such as race, gender identity, or hate speech. Therefore, we cannot measure directly if TikTok's algorithm suppressed or promoted content about marginalized groups in our dataset. We rely instead on

external studies (e.g. Ungless et al. 2024) for those insights. Future work could annotate more videos for sensitive attributes or examine shadow-banned hashtags to fill this gap.

Finally, our content categories and emotion tags were pre-defined and possibly coarse. For example, many videos might blend categories (e.g. a political video delivered humorously) but only one tag is used. This could obscure nuance. In future research, a more granular content analysis (perhaps using machine learning classifiers on video transcripts or images) could refine these labels.

RESULTS AND DISCUSSION

The dataset comprises information extracted from TikTok, focusing on author and video characteristics. It includes quantitative metrics related to author popularity and video engagement, as well as categorical data describing video content.

The dataset is presented in an aggregated format, primarily focusing on the top 10 and bottom 10 performing videos based on views or engagement. This limits the ability to perform granular analysis on individual videos but allows for comparative analysis of high and low-performing content.

Visualization 1 focuses on the top 10 authors ranked by overall engagement. It includes follower count and all four-engagement metrics.

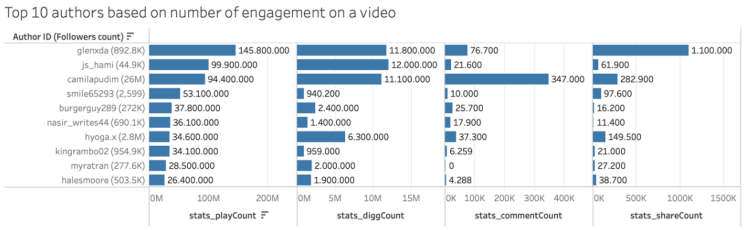

Figure 1: Top 10 authors based on number of engagement
Source: Tableau

Visualizations 2 focuses on the top 10 authors ranked by views, incorporating communicative function and emotion categories.

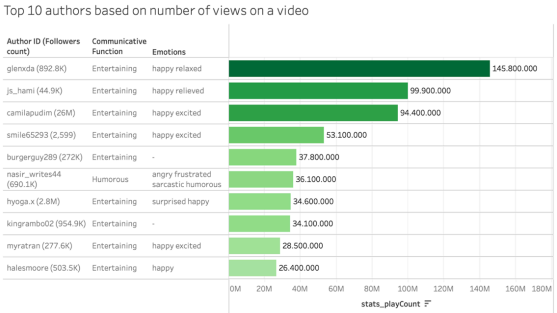

Figure 2: Top 10 authors based on number of views
Source: Tableau

Visualization 3 focuses on the bottom 10 authors ranked by views, also including communicative function and emotion categories.

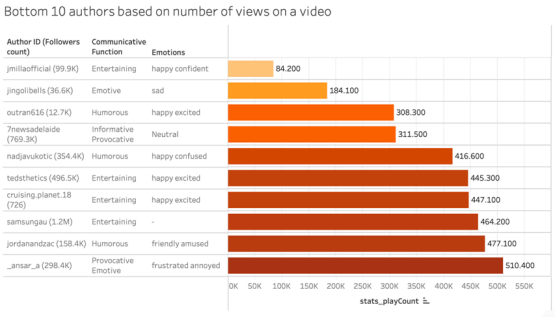

Figure 3: Bottom 10 authors based on number of views
Source: Tableau

Visualization 4 provides the overall distribution of communicative functions within the dataset.

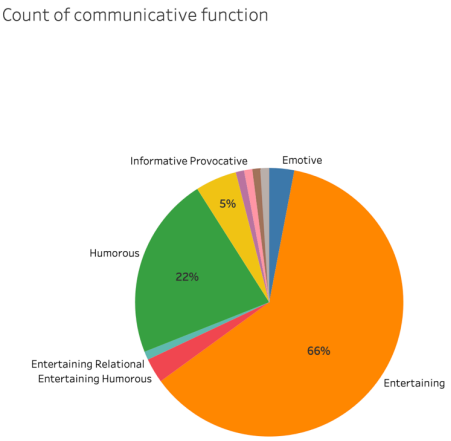

Figure 4: Count of communicative function
Source: Tableau

A key characteristic of the dataset is the categorization of videos by "Communicative function" and "Emotions." This allows for analysing the relationship between content type and performance metrics. It is important to note that the dataset reflects a snapshot of TikTok data and may not be fully representative of the platform's dynamic nature. Additionally, the absence of explicit discriminatory keywords in the video descriptions, necessitates a focus on analysing broader patterns of content categorization and reach to identify potential implicit biases.

In examining “content governance,” this study focuses on TikTok’s internal platform governance system—its company-managed moderation, algorithmic curation, and policy

enforcement mechanisms. This is distinct from state or governmental regulation, although it increasingly interacts with it. TikTok, like other major platforms, operates under a model of corporate self-governance, setting its own rules for acceptable content and algorithmic promotion while being subject to oversight through emerging regulatory frameworks such as the European Union's Digital Services Act (DSA) and Australia's Online Safety Act (2021). This context highlights how a private platform effectively performs a quasi-governmental role in governing speech, visibility, and cultural expression at scale.

The analysis of the TikTok video data visualizations reveals several key findings regarding content performance and the platform's FYP algorithm :

Relevance vs. Popularity in Content Curation

A central question is whether TikTok's FYP is driven by popularity (follower count) or by relevance and interest matching. In our correlation analysis, we found surprisingly low association between a creator's follower count and the video's engagement metrics (views, likes, etc.). For example, among our top 10 most-viewed videos (Figure 1), 60% of the creators did not belong to the top quartile of follower counts in the dataset. The scatterplot of followers vs. views showed a weak correlation coefficient ($r \approx 0.2$) and many small-fan accounts with disproportionately large views.

These findings suggest that TikTok's algorithm can indeed elevate content from relatively unknown creators. This outcome aligns with TikTok's stated emphasis on "delivery based on relevancy, not popularity." The platform often touts that even new users can go viral if their content matches audience interests. Our data provide some empirical support: for instance, the TikTok handle JS_Hami (with only a few thousand followers) had a top-viewed video of 5.2 million views. Such examples indicate that high views are not exclusive to celebrity accounts, unlike on YouTube where view counts typically scale strongly with subscribers (Matamoros-Fernández et al., 2021).

However, we caution that our method alone cannot prove a content-based relevance mechanism. It is possible that trending topics or use of popular hashtags helped certain low-follower videos gain traction. Moreover, because we lack longitudinal data, we cannot rule out that some highly-viewed videos quickly grew their follower counts after going viral (thus creating a mild reverse-causality). Nevertheless, the lack of a tight follower–views correlation is notable. It confirms one narrative: TikTok's FYP does not simply mimic Instagram's or Twitter's follower-centric feeds, and it appears less biased toward incumbents.

Dominance of Entertaining Content

A striking pattern emerged when we categorized the top-performing versus bottom-performing videos. Seven of the ten highest-viewed videos (Figure 2) were labelled as entertainment/humour. Examples included short comedy skits, amusing lip-syncs, and viral dance memes. By contrast, among the 10 least-viewed videos (Figure 3), the most common

function was Social/Religious, including messages of gratitude, prayers, or social reflection – content that was less "viral-friendly" in tone and style.

This suggests that TikTok's algorithm strongly favours content that is light-hearted and entertaining. Even some videos labelled with more serious or negative emotions performed well if they had an entertaining framing. For instance, one high-ranked video in our sample expressed anger and sarcasm toward a social issue, yet it still amassed over 36 million views (likely because it was delivered in a comedic or meme-like fashion). Overall, content with positive or humorous emotion tags had significantly higher average views than content with negative or neutral emotional tones.

Comparatively, we note that YouTube and Instagram exhibit similar entertainment biases. YouTube has often been criticized for amplifying sensational or humorous content because it maximizes watch time (Matamoros-Fernández et al., 2021). Instagram's Explore and Reels also tend to surface visually polished, upbeat videos. The frontiers perspective on Instagram hashtag #IWantToSeeNyome (Willcox, 2025) illustrates the flip side: when TikTok or IG's algorithm favours mainstream aesthetics (slim, smiling models, upbeat humour), it can marginalize other content even if it is creative or socially valuable.

The dominance of comedy/entertainment on TikTok has implications for content diversity. Creators who produce educational, religious, or social-commentary videos may struggle to gain visibility unless they package their message in an entertaining format. This dynamic could inadvertently suppress valuable niche topics. It also suggests an alignment issue: TikTok's policies advocate for creativity and inclusivity, but the algorithm rewards a narrow style (primarily humour). As Cobbe (2021) warned, automated feeds can "insert" a platform's preferred content (here, comedic content) into public discourse, potentially at the expense of other voices.

Role of Positive Emotions

Closely tied to content type is the emotional atmosphere of videos. We observed that positive and joyful emotions were overrepresented among the top videos, while negative or distressing emotions were more common in low-view categories. For example, the tag "happy" appeared on 80% of the high-view group (Figure 2) but only 10% of the low-view group (Figure 3). Conversely, tags like "sadness" or "disgust" were nearly absent in the top performers. This pattern aligns with TikTok's community guidelines emphasizing a "fun" environment; it also matches a broader trend that engagement algorithms often prefer uplifting or comedic material (Matamoros-Fernández et al., 2021).

Interestingly, even when a video conveyed *negative* feelings, if it was framed humorously, it still reached a wide audience. The previously noted angry-sarcastic video is an example. On the other hand, serious activism or frank expressions of frustration did not appear among the high-engagement set. This echoes observations from Instagram's cultural analysis: marginalized voices often report that protest or discomforting images are demoted by the algorithm (Willcox, 2025).

These findings underscore a potential affective bias: TikTok's curation favours light-hearted content, possibly because it keeps users watching longer and recommending more videos. While positivity bias might sound benign, it can have subtler effects. By consistently showing cheerful content, the platform may shape user sentiment and the overall culture of the app. Critics have pointed out that an overdose of false positivity can drown out important dialogues on social issues. In content moderation terms, if negative or critical videos tend to get less reach, users discussing serious topics (mental health, racism, politics) may feel censored even if not formally removed (Ungless et al., 2024; Willcox, 2025).

Content Category and Performance

Expanding on the above, we compared the average engagement across broader content categories. **Entertainment/humour** videos in our sample had the highest mean views (around 30 million) and likes. The next highest was **social/religious**, but those still averaged far lower (around 5 million views) and were mostly in the bottom half of our ranking. Educational or Motivational videos had moderate view counts. Notably, **none** of the *explicitly political or controversial* categories (e.g. News, Politics, Protest) appeared in the top 10 performing list, even if they occasionally appeared in the bottom group.

This contrasts with YouTube, where political or ideological content can sometimes go viral, especially around events (Ibrahim et al., 2025). On Instagram, the introduction of reels in 2022 popularized certain lifestyle and challenge videos, but activism has found alternative niches (e.g. via hashtags). TikTok's FYP, in our data, seems to gravitate heavily toward *pure entertainment*.

While this pattern might reflect user demand (people share and watch what they find funny), it may also indicate an algorithmic *filtering* choice. Platforms often use subtle rule-based boosts: for example, TikTok might detect comedic keywords or use engagement velocity to rank up jokes. Without internal knowledge, we cannot say for sure, but the dominance of one category suggests the algorithm's reward function heavily weights forms of content that historically generate high watch times. This has the side-effect of de-emphasizing other categories, which may be viewed as a systemic bias.

Comparative Insights: YouTube and Instagram

To place TikTok's trends in context, we compare them with known practices on YouTube and Instagram:

- **Algorithmic design:** YouTube's recommender also personalizes content but tends to factor in subscription networks and longer watch history. TikTok's shorter videos and rapid-fire FYP mean new content can go viral almost instantly. The student study by Bishqemi & Crowley (2022) found that identical flower posts on TikTok and Instagram saw hundreds of views on TikTok vs. nearly zero on Instagram, illustrating TikTok's potent discovery engine. Meanwhile, Instagram's feed relies on who you

follow more than hashtags; it only recently introduced search-based discovery akin to TikTok's. These differences mean TikTok can amplify content more widely and unpredictably than Instagram's more follower-centric model (De, 2024).

- **Content moderation and transparency:** On enforcement, TikTok claims it removed 153 million videos in Q4 2024, far more than YouTube's 9.5 million in the same period (Eltaher et al., 2025). This disparity is partly due to the volume of content and stricter policies. TikTok has also rolled out child safety modes and age gating, as noted earlier (Eltaher et al., 2025). YouTube likewise uses AI to flag problematic uploads, and Instagram leverages community reporting extensively. However, transparency varies: TikTok's published Community Guidelines and Transparency Center provide general categories, but they offer little in the way of the *algorithmic logic*. In contrast, YouTube occasionally publishes research papers on how its systems work, and Instagram's parent Meta produces quarterly reports on removed content (including hate or CSAM takedowns) though they do not detail recommendation factors. The EU's DSA now binds TikTok, YouTube, and Instagram to reveal more about their moderation metrics and recommender "parameters." In fact, recent analysis by Annabell et al. emphasizes that current documentation is insufficient to understand TikTok's search recommendations (Annabell et al., 2025). This suggests all three platforms still fall short of full disclosure, but TikTok, given its young user base, is under particular scrutiny for algorithmic transparency.
- **Marginalized content:** All platforms have documented tensions with minority communities. TikTok influencers have accused the app of suppressing queer or anti-establishment content, a phenomenon Ungless et al. (2024) call "algorithmic censorship". Instagram similarly saw campaigns like #IWantToSeeNyome and protests by women of colour over new algorithms, indicating that when marginalized creators post advocacy or educational material, it may get less reach (Willcox, 2025; De, 2024). YouTube has had its own problems (e.g. de-platforming Black Lives Matter videos in 2020, which TikTok also temporarily did by accident) (Ibrahim et al., 2025). Our results, which show low exposure for non-entertainment content – imply that TikTok, like its peers, is more likely to sideline posts that are not immediately gratifying to a wide audience. This can have representational consequences: as Barocas et al. (2017) and Velkova & Kaun (2021) point out, when algorithms invisible certain groups (by not recommending their content), it reinforces social biases.
- **User behaviour and platform power:** The observed algorithmic biases also shape how users behave. Knowing that humorous, upbeat videos tend to do well, creators may tailor their output accordingly. On TikTok, this means tutorials or serious discussions might be presented with a humorous twist to attract views. In terms of platform power, these algorithmic preferences mean that TikTok (and its rivals) play a gatekeeping role in public discourse. For example, if creators from a marginalized

community consistently see their content suppressed, they might lose trust in the platform. On the other hand, when the algorithm unexpectedly surfaces a creator's video, it can suddenly grant them massive influence (the "overnight sensation" phenomenon unique to TikTok). Thus, TikTok's FYP is a double-edged sword, it can democratize visibility *conditionally*, but also concentrate power in invisible ways.

Collectively, these findings illustrate how major social media platforms engage in self-regulation rather than direct governmental control. However, mounting regulatory pressure—particularly from the EU and Australian authorities—signals that platform governance is gradually shifting from purely private management toward hybrid oversight models, where governments play a more active role in ensuring transparency and algorithmic accountability.

Policy Alignment and Implications

TikTok's public-facing policies, including its *Community Guidelines*, *Creativity Program Beta Policy*, and *Safety Center*, repeatedly emphasize its commitment to *"fostering creativity and joy"* and building an *"inclusive community where everyone feels welcome."* To promote this narrative, TikTok implements several mechanisms: the removal of harmful or hateful content, the elevation of "authentic" creativity through the *For You* feed, and specific initiatives like the *#CreatorsOfColor* and *#LearnOnTikTok* campaigns designed to amplify educational and diverse voices. These efforts align with broader platform claims that recommendation is based on "relevance" and "user interests," rather than fame or follower count (TikTok Transparency Center, 2024).

However, our findings indicate that the *practical outcomes of the FYP algorithm only partially reflect these ideals*. While TikTok's claim that popularity does not determine reach finds modest empirical support (given the weak correlation between followers and views), the platform's feed still overwhelmingly favours entertainment-oriented and emotionally positive content. This pattern reveals a subtle but significant policy-practice gap. Content that is humorous, light hearted, or visually engaging tends to be amplified, whereas socially reflective, religious, or educational material receives markedly less visibility. In effect, TikTok's "community and inclusivity" narrative coexists with an algorithmic environment that rewards affective positivity and penalizes critical or issue-based expression.

Prior studies corroborate this discrepancy. Matamoros-Fernández et al. (2021) and Willcox (2025) both show that TikTok's moderation systems disproportionately suppress content discussing race, body diversity, or activism — either through automatic demotion or "shadow-banning." Similarly, Ungless et al. (2024) found that creators from marginalized groups frequently reported lower engagement and unexplained content removal, even when adhering to platform rules. These findings suggest that while TikTok's written policies

prohibit discrimination and promise visibility equity, its automated enforcement and engagement-driven algorithmic design perpetuate structural biases.

From a governance perspective, this tension illustrates the difference between *normative policy goals* (community, safety, inclusivity) and *algorithmic optimisation goals* (watch time, retention, engagement). The result is a system that appears community-oriented at the level of rhetoric but commercially driven at the level of practice. Scholars such as Cobbe (2021) and Gillespie (2018) have noted that this disjunction characterises platform self-regulation broadly: "policy language" frames platforms as moral actors, while algorithmic systems operationalize attention as capital.

As regulatory regimes evolve — particularly the European Union's *Digital Services Act (2022)* and Australia's *Online Safety Act (2021)* — these inconsistencies have material implications. Both frameworks require greater transparency around recommendation parameters and systematic risk assessments. For TikTok, this means demonstrating not only compliance in removing prohibited content but also fairness in amplification — ensuring that entertainment bias does not become de facto content discrimination. Strengthening the alignment between TikTok's inclusivity rhetoric and algorithmic behaviour would therefore require:

1. **Independent algorithm audits** to assess how community-oriented and educational content is ranked.
2. **Public reporting of recommendation criteria** beyond engagement metrics.
3. **Regular equity reviews** assessing the visibility of marginalized creators and non-entertainment genres.

In sum, while TikTok's governance model articulates a clear vision of inclusivity and safety, the data in this study — and in prior audits — reveal a persistent misalignment between policy and practice. The platform's success in sustaining "fun" and "positive" content ecosystems may come at the cost of discursive diversity and social depth, raising pressing questions about the cultural consequences of algorithmic curation under corporate governance.

CONCLUSION

This study contributes to understanding TikTok's platform governance. That is, how the company's internal moderation and algorithmic design govern visibility and content performance. While not a government in the traditional sense, TikTok exerts regulatory-like power over speech and participation online, blurring the boundary between corporate governance and public interest regulation**.**

This study aimed to examine TikTok's content curation practices, specifically the FYP algorithm's prioritization of content relevance versus creator popularity, and to explore potential biases in content promotion. The analysis reveals a complex picture. While the algorithm does consider content relevance, as evidenced by the success of less popular

creators, creator popularity also plays a role in driving views and engagement. More significantly, our findings indicate a clear algorithmic skew: "Entertaining" content dominates the platform, comprising 66% of the dataset, and positive emotions are frequently associated with high-performing videos. This suggests the algorithm implicitly favours certain content and creators, echoing observations by algorithmic experts like Bishop (2020), who notes they "sell theorizations of algorithmic visibility to aspiring and established creators" (p. 1).

This emphasis on entertainment and positive emotion has several implications. Frey (2021a) argues that recommender systems are essential for managing content surplus, and Frey (2021b) contends they form a central part of business models, branding, and promotional rhetoric. These observations underscore that algorithms are not neutral; their design choices actively shape what is visible and, in our case, can implicitly marginalize content that serves other communicative functions or expresses a wider range of emotions. This resonates with comparative platform analyses, which find that YouTube and Instagram similarly tend to surface polished, upbeat media and can inadvertently suppress niche or advocacy content (Eltaher et al., 2025; De, 2024). We also highlighted the plight of marginalized creators who report feeling censored by these ostensibly neutral algorithms (Ungless et al., 2024; Willcox, 2025). In short, TikTok's practice only partially aligns with its policies. Its community guidelines proclaim a fun, inclusive environment, but the algorithm's output privileges certain forms of expression, echoing societal biases. This dynamic also reflects a tension between the societal function of news and the "individualistic logic of personalization" (Helberger, 2019, p. 1009), a tension that extends to other forms of content on platforms like TikTok.

Implications and Recommendations

These findings have several implications. First, for **platform design**, our work suggests a need to recalibrate recommendation priorities to enhance content diversity. Platforms could explicitly down-weight a positivity bias or insert more variety to prevent the marginalization of other content. Second, for **user behaviour**, creators should be aware that the algorithm's tastes may not reward purely educational or activist videos, which could influence where they choose to invest their creative efforts. Third, for **platform power and public discourse**, our analysis confirms that TikTok wields significant agenda-setting power. Its algorithm can boost or bury content at scale, raising ethical questions about transparency and oversight.

Based on our study, we recommend several avenues for action and future research:

Recommendations for action:

- **Implement algorithmic audits:** Platforms should implement reproducible and ongoing algorithmic audits (Mosnar et al., 2025) to monitor changes over time. The EU's Digital Services Act (DSA) could play a crucial role here by mandating such

audits for TikTok, YouTube, and Instagram, ensuring regulators and independent researchers have the necessary data access.

- **Expand reporting:** Platforms should expand their public reporting to include breakdowns of content exposures by category, revealing hidden trends and biases.
- **Address representational harms:** Companies should consider the representational harms identified in studies like ours. Targeted adjustments or exceptions may be warranted if certain groups feel "algorithmically silenced."

Recommendations for future research:

- **Mixed methods:** Future studies should integrate qualitative methods (e.g., interviews with TikTok moderators or creators) to complement quantitative audits and better explain causal mechanisms.
- **Cross-national and longitudinal studies:** Cross-national comparisons are needed, as TikTok's algorithm can vary by region or over time. Longitudinal research should track whether the patterns we saw (e.g., entertainment dominance) persist as platforms revise their systems.
- **Explore algorithmic biases:** Further research should explore the specific mechanisms through which the algorithm assesses and ranks content, as well as the long-term effects of these trends on content diversity and creator equity.

In conclusion, as TikTok and its peers continue to dominate online media, rigorous investigation of their content governance is vital. Our study provides a snapshot showing both alignments and divergences between policy and practice. Ensuring that platform algorithms truly uphold stated values, especially in safeguarding minors and amplifying diverse voices, will require sustained effort from researchers, regulators, and the platforms themselves. Only with such engagement can we hope to make these powerful recommender systems serve the broader social good rather than just commercial engagement.

Acknowledging the use of generative artificial intelligence

This work was made possible with the support of several individuals and resources. I am particularly grateful for the assistance of an AI-powered search tool, which was used to facilitate the initial literature review and identify key scholarly articles related to the topic of algorithmic governance and content moderation. While this tool aided in the discovery of relevant sources, all subsequent analysis, synthesis, and final manuscript composition remain my sole responsibility."

REFERENCES

Annabell, T., Gorwa, R., Scharlach, R., van de Kerkhof, J., & Bertaglia, T. (2025). TikTok Search Recommendations: Governance and Research Challenges. *arXiv:2505.08385v1*.

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017). The problem with bias: from allocative to representational harms in machine learning. Special Interest Group for Computing. *Information and Society (SIGCIS), 2*.

Bishop, S. (2020). Algorithmic Experts: Selling Algorithmic Lore on YouTube. *Social Media + Society, 6*(1). https://doi.org/10.1177/2056305119897323

Bishqemi, K., & Crowley, M. . (2022). TikTok Vs. Instagram: Algorithm Comparison. *Journal of Student Research*, *11*(1). https://doi.org/10.47611/jsrhs.v11i1.2428

Bucher, T. (2012).Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society 14*, 7, pp. 1164-1180. https://doi.org/10.1177/1461444812440159

Cobbe, J. (2021). Algorithmic Censorship by Social Platforms: Power and Resistance. *Philosophy & Technology, 34*(4): 739–766.

De, A. (2024). *Instagram versus women of color: Why are women of color protesting Instagram's algorithmic changes?*. *arXiv:2407.17679v1*. https://doi.org/10.48550/arXiv.2407.17679

Eltaher, F., Gajula, R. K., Miralles-Pechuán, L., Crotty, P., Martínez-Otero, J., Thorpe, C., & McKeever, S. (2025). *Protecting Young Users on Social Media: Evaluating the Effectiveness of Content Moderation and Legal Safeguards on Video Sharing Platforms*. arXiv preprint arXiv: 2505.11160v1 [cs.SI]. https://doi.org/10.48550/arXiv.2505.11160

Frey, M. (2021a). Why We Need Film and Series Suggestions. In *Netflix Recommends: Algorithms, Film Choice, and the History of Taste* (1st ed., pp. 23–37). University of California Press. https://doi.org/10.2307/j.ctv269fvqp.5

Frey, M. (2021b). How Algorithmic Recommender Systems Work. In *Netflix Recommends: Algorithms, Film Choice, and the History of Taste* (1st ed., pp. 38–62). University of California Press. https://doi.org/10.2307/j.ctv269fvqp.6

Friedl, P., & Morgan, J. (2024). Decentralised content moderation. *Internet Policy Review, 13*(2). https://doi.org/10.14763/2024.2.1754

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

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society, 7*(2). https://doi.org/10.1177/2053951720943234

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

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2): 1–35.

Helberger, N. (2019). On the Democratic Role of News Recommenders. *Digital Journalism*, *7*(8), 993–1012. https://doi.org/10.1080/21670811.2019.1623700

Ibrahim, H., Jang, H. D., Aldahoul, N., Kaufman, A. R., Rahwan, T., & Zaki, Y. (2025). TikTok's recommendations skewed towards Republican content during the 2024 U.S. presidential race. *arXiv:2501.17831v1*.

Matamoros-Fernández, A., Gray, J., Bartolo, L., Burgess, J., & Suzor, N. (2021). What's "Up Next"? Investigating Algorithmic Recommendations on YouTube Across Issues and Over Time. *Media and Communication, 9*(4), 234-249. https://doi.org/10.17645/mac.v9i4.4184

Matlach, P.-C., Castillo, A., Drath, C., & Hevesi, E. F. (2025). *Recommending Hate: How TikTok's Search Engine Algorithms Reproduce Societal Bias*. Institute for Strategic Dialogue (ISD).

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems, 31*(3), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

Mosnar, M., Skurla, A., Pecher, B., Tibensky, M., Jakubcik, J., Bindas, A., Sakalik, P., & Srba, I. (2025). Revisiting Algorithmic Audits of TikTok: Poor Reproducibility and Short-term Validity of Findings. *arXiv:2504.18140v1 [cs.IR]*.

Pasquale, F. (2015). The Black Box Society, the Secret Algorithms That Control Money and Information. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Register, Y., Qin, L., Baughan, A., & Spiro, E. S. (2023). Attached to "The Algorithm": Making Sense of Algorithmic Precarity on Instagram. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Article 563, pp. 1–15). Association for Computing Machinery. https://doi.org/10.1145/3544548.3581257

Stockinger, A., Schäfer, S., & Lecheler, S. (2023). Navigating the gray areas of content moderation: Professional moderators' perspectives on uncivil user comments and the role of (AI-based) technological tools. *New Media & Society, 27*(3), 1215-1234. https://doi.org/10.1177/14614448231190901 (Original work published 2025)

Ungless, E, L., Markl, N., & Ross, B. (2024). Experiences of censorship on TikTok across marginalised identities. in *Proceedings of the International AAAI Conference on Web and Social Media, AAAI Press, 19th International AAAI Conference on Web and Social Media, Copenhagen, Denmark, 23/06/25*.

Velkova, J., & Kaun, A. (2021). Algorithmic resistance: media practices and the politics of repair. *Information, Communication & Society, 24*(4): 523–540

Willcox, M. (2025). Algorithmic agency and "fighting back" against discriminatory Instagram content moderation: #IWantToSeeNyome. *Frontiers in Communication*. https://doi.org/10.3389/fcomm.2024.1385869

Xue, H., Nishimine, B., Hilbert, M., Cingel, D., Vigil, S., Shawcroft, J., Thakur, A., Shafiq, Z., & Zhang, J. (2025). Catching Dark Signals in Algorithms: Unveiling Audiovisual

and Thematic Markers of Unsafe Content Recommended for Children and Teenagers. *arXiv:2507.12571v1*. https://doi.org/10.48550/arXiv.2507.12571

Zhu, L. & Lerman, K. (2016). Attention Inequality in Social Media. *ArXiv* abs/1601.07200.

***(Additional sources)**:*

Platform policy documents (TikTok Community Guidelines, YouTube Community Guidelines, Instagram Community Standards)