Auditing Algorithmic Moderation: Bias, Internal Signals, and Social Feedback in Online Comment Classification

Vaibhav Satish
Indian Institute of Technology Madras

Abstract

Online platforms increasingly rely on automated systems to categorize and moderate user-generated content. However, the interaction between textual signals, user feedback, identity-related references, and internal platform features in shaping moderation outcomes remains underexplored. In this study, we analyze a large-scale dataset of online comments containing textual content, engagement signals (upvotes, downvotes, emoticons), identity reference indicators (race, religion, gender, disability), and hidden system features used by the platform. Our goal is to investigate how these multi-modal signals contribute to automated comment categorization and to assess potential fairness and transparency concerns. We train and compare multiple predictive models using (1) textual and identity features, (2) social feedback signals, (3) internal system features, and (4) their combination. Results show that while textual content explains a substantial portion of categorization decisions, internal platform signals exhibit significant predictive power, suggesting that moderation outcomes may depend on opaque automated components beyond the text itself. Furthermore, identity-related references are associated with differing categorization patterns, raising questions about potential disparate impact across protected groups. We also examine temporal dynamics of discussions and identify early linguistic markers that predict later escalation into more severe categories. Our findings highlight the importance of auditing multi-signal moderation pipelines and demonstrate how combining explainability and fairness analysis can uncover hidden influences in automated content governance systems.

1 Introduction

Online content moderation systems increasingly function as core infrastructure for public discourse [1](Gorwa et al.). Decisions made by these systems, what speech is amplified, restricted, or removed shape political debate, social norms, and collective participation at scale [2] (Gillespie). Far from being neutral technical tools, automated moderation systems now operate as de facto governance mechanisms, exercising power over visibility, legitimacy, and participation in digitally mediated public spaces [2, 3] (Gillespie; Klonick). As platforms expand across diverse linguistic, cultural, and political contexts, especially in the Global South, understanding how these systems make decisions becomes a question of democratic accountability rather than mere computational efficiency [9] (Leong).

Most existing explanations of automated moderation emphasize textual analysis: harmful language detection, sentiment classification, or semantic toxicity scoring [1] (Gorwa et al.). While text is undoubtedly central, moderation decisions in real-world platforms are rarely based on content alone. Instead, they emerge from complex interactions between user-visible text, social engagement signals, identity-related references, and internal platform-level features that remain opaque to users and external auditors [6] (Jhaver et al.). This paper advances what we term the Transparency Gap Thesis: a substantial portion of automated moderation decisions is driven by non-textual and non-user-visible signals, creating a gap between what users can observe and contest and what actually governs moderation outcomes. This gap challenges prevailing transparency narratives that frame moderation as explainable through content-level analysis alone [4] (Roberts et al.).

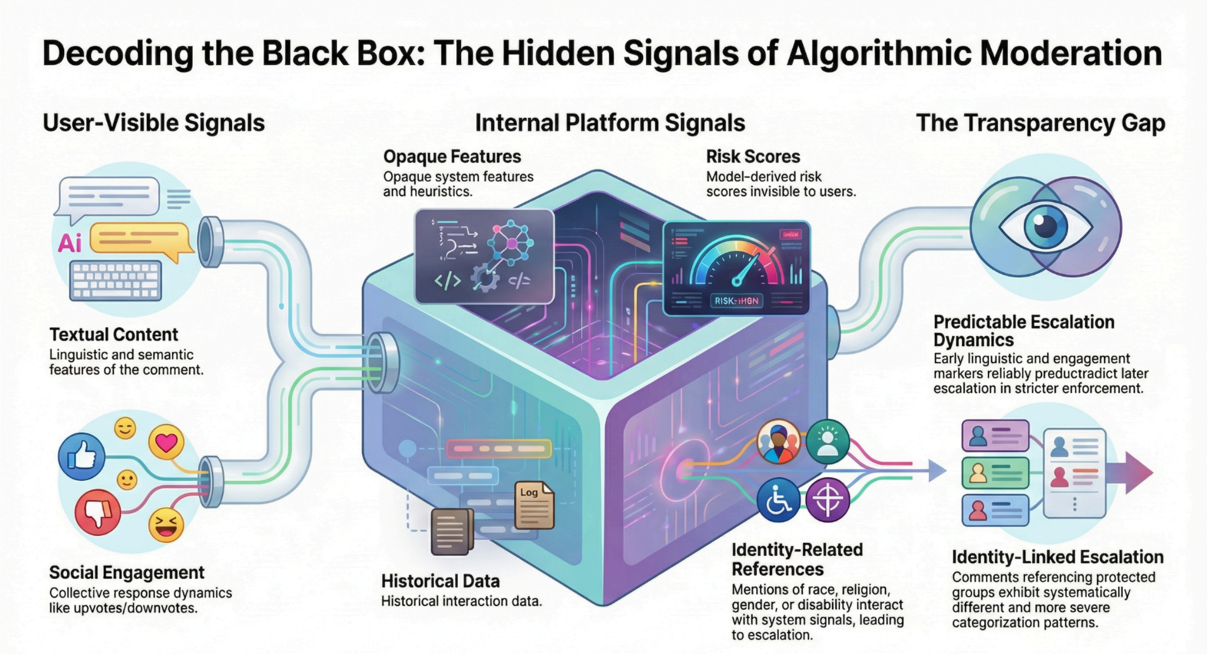
Decoding the Black Box: The Hidden Signals of Algorithmic Moderation
User-Visible Signals
Internal Platform Signals
The Transparency Gap
Opaque Features
Opaque system features and heuristics.
Risk Scores
Model-derived risk scores invisible to users.
Ai
RISK: HIGN
Textual Content
Linguistic and semantic features of the comment.
Predictable Escalation Dynamics
Early linguistic and engagement markers reliably preductradict later escalation in stricter enforcement.
Log
Social Engagement
Collective response dynamics like upvotes/downvotes.
Historical Data
Historical interaction data.
Identity-Related References
Mentions of race, religion, gender, or disability interact with system signals, leading to escalation.
Identity-Linked Escalation
Comments referencing protected groups exhibit systematically different and more severe categorization patterns.

Figure 1. Hidden Signals of Algorithmic Moderation

Text-only explanations are therefore insufficient for at least three reasons. First, they obscure the role of social feedback and engagement dynamics that can amplify or suppress content independent of its semantic properties [6]. Second, they ignore internal system features such as historical user signals or platform-specific heuristics that may encode institutional priorities and risk tolerances [3], [5] (Klonick; Barocas et al.). Third, they fail to account for how identity-related references interact with these signals, potentially producing differential outcomes across protected groups [5]. When moderation systems rely on opaque signals that users cannot inspect or contest, accountability mechanisms are weakened, and procedural fairness becomes difficult to guarantee.

These issues carry heightened significance for democratic discourse in the Global South. Platforms operating in developing and postcolonial contexts often moderate multilingual, culturally specific, and politically sensitive content using systems trained and calibrated elsewhere [8], [9]. In such settings, opacity in moderation pipelines can disproportionately affect marginalized communities, intensify trust deficits, and exacerbate existing power asymmetries in public communication [7, 9]. Understanding how moderation decisions are actually produced is thus essential for inclusive governance and responsible AI deployment [5], [9]. Against this backdrop, this study empirically audits automated moderation as a multi-signal governance pipeline. Specifically, it addresses the following research questions:

RQ1: To what extent do textual content, social feedback signals, identity-related references, and internal platform features each contribute to automated moderation outcomes? RQ2: Do identity-related references interact with moderation signals in ways that produce systematically different categorization trajectories? RQ3: Are there early linguistic or engagement-based signals that reliably predict escalation into more severe moderation categories? By answering these questions, this paper reframes automated content moderation as an object of governance and accountability, rather than a purely technical classification task, and contributes empirical tools for auditing opaque decision-making systems in digitally mediated public discourse.

2 Related Work

2.1 Text-Based Content Moderation

Research on automated content moderation has largely focused on text-centric approaches, including toxic language detection, hate speech classification, and sentiment analysis using supervised and deep learning models [10], [12], [13]. These studies have demonstrated the effectiveness of linguistic and semantic features in identifying policy-violating content across platforms [11], [13]. While this work has been instrumental in advancing scalable moderation techniques, it implicitly treats moderation as a problem of textual interpretation, often abstracted away from the broader platform context in which decisions are made [1], [13]. As a result, non-textual signals that routinely accompany real-world moderation decisions remain underexplored [6] (Jhaver et al., AutoModerator).

2.2 Explainability in Platform Machine Learning Systems

A growing body of work examines explainability and interpretability in machine learning systems, particularly in high-stakes decision-making contexts [14], [16]. Techniques such as feature attribution and local explanations have been proposed to improve transparency and user trust in algorithmic systems [14], [15]. However, most explainability research assumes access to complete and meaningful feature representations [16], [17]. In platform settings, many influential signals—such as internal heuristics, historical user data, or engagement-derived features—are not visible to users or external auditors [4], [17]. Consequently, explainability efforts often remain confined to surface-level inputs, limiting their capacity to capture how moderation decisions are actually produced [24].

2.3 Fairness and Identity in Moderation

Studies on fairness in content moderation have highlighted biases associated with identity-related language [18], [21], showing that references to race, religion, gender, or disability can be unevenly classified as harmful or inappropriate [18], [21] . While these findings underscore important risks of disparate impact, existing work typically analyzes identity references in isolation, without situating them within the full moderation pipeline [19], [20]. This narrow focus risks attributing bias solely to textual models, overlooking how identity-linked signals may interact with engagement dynamics and platform-level features to shape outcomes [5], [24].

2.4 Platform Governance and Algorithmic Accountability

Scholars in science and technology studies, law, and policy have increasingly framed content moderation as a governance challenge, emphasizing accountability, transparency, and democratic oversight [2, 22, 23]. These contributions provide critical normative and institutional insights but are often disconnected from empirical analysis of moderation systems themselves [3], [22]. As a result, governance debates frequently lack evidence about how algorithmic decisions are operationalized at the system level [1], [4].

Taken together, existing research advances understanding of individual components of automated moderation but rarely examines how internal platform signals interact with textual content and identity-related references to shape moderation outcomes. This fragmentation limits both technical explainability and meaningful governance oversight, an empirical gap this study directly addresses.

3 Conceptual Framework: Moderation as a Governance Pipeline

Automated content moderation is commonly framed as a technical classification task, in which textual inputs are mapped to categorical outputs [2, 13]. Such framings, however, obscure the institutional and infrastructural nature of moderation systems as mechanisms of governance [2, 27]. To address this limitation, we conceptualize automated moderation as a Multi-Signal Moderation Pipeline (MSMP), a layered decision-making process in which multiple classes of signals are sequentially combined to produce moderation outcomes.

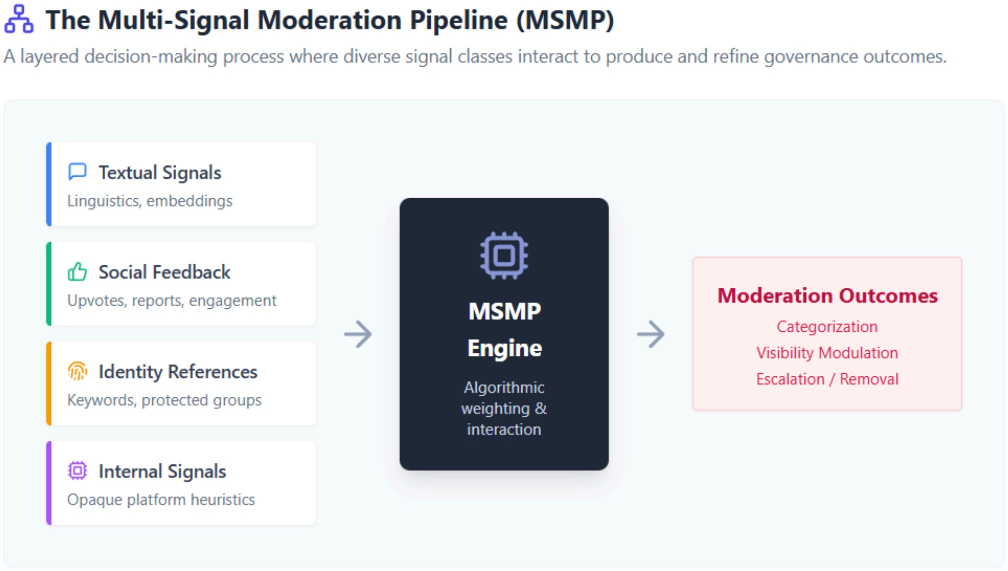
The Multi-Signal Moderation Pipeline (MSMP)
A layered decision-making process where diverse signal classes interact to produce and refine governance outcomes.
Textual Signals
Linguistics, embeddings
Social Feedback
Upvotes, reports, engagement
Identity References
Keywords, protected groups
Internal Signals
Opaque platform heuristics
MSMP Engine
Algorithmic weighting & interaction
Moderation Outcomes
Categorization
Visibility Modulation
Escalation / Removal

Figure 2: The Multi-Signal Moderation Pipeline showing inputs, processing, and the recursive feedback loop.

Within the MSMP, moderation decisions emerge from the interaction of three distinct signal domains. The first consists of user-visible signals, including the textual content of comments and observable engagement indicators such as upvotes, downvotes, and emotive reactions. These signals are directly accessible to users and form the basis of common assumptions about how moderation operates. Most transparency narratives implicitly assume that moderation outcomes can be explained primarily through this domain [28].

The second domain comprises platform-visible signals, which include aggregated behavioral indicators, historical interaction patterns, and metadata accessible to the platform but not to users. While these signals may be derived from user activity, their operationalization and weighting remain opaque, positioning them outside meaningful user contestation or scrutiny [5, 28].

The third domain involves hidden system features, such as internal heuristics, model-derived risk scores, or policy-enforcement parameters embedded within the moderation infrastructure. These features are neither directly observable nor easily inferable, yet they can exert substantial influence over categorization decisions. Their opacity raises fundamental concerns about accountability, as affected users are unable to identify or challenge the basis of moderation outcomes.

These signal domains feed into a set of decision outputs, including content categorization, visibility modulation, and escalation to stricter enforcement tiers. Importantly, outputs are not static endpoints but can recursively influence upstream signals for example, through feedback loops that reshape engagement dynamics or future risk assessments [30]. By framing automated moderation as an MSMP, this study shifts the analytical focus from isolated models to governance pipelines. This perspective enables empirical auditing of decision influence across signal domains and provides a conceptual foundation for assessing transparency, fairness, and accountability in platform-mediated public discourse.

4 Data and Signal Taxonomy

4.1 Data Description

This study analyzes a large-scale dataset of user-generated comments released as part of a public Kaggle competition, derived from an online discussion platform where automated content categorization is used for moderation and governance purposes [31].

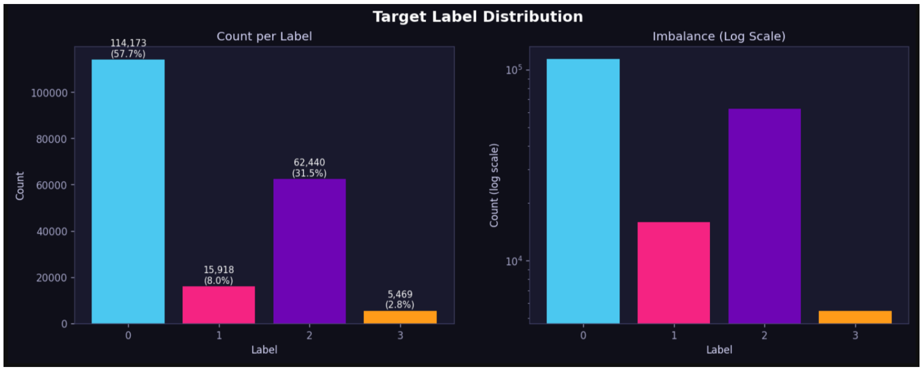

Figure 3: Target Label Distribution (Count) and (Log Scale)

The dataset comprises textual comments, associated engagement metadata, moderation category labels, and a set of platform-level features used in automated decision-making [32].

Label distribution:

Neutral (0): 114,172 (57.66%) ; Mild (1): 15,918 (8.04%) ; Offensive (2): 62,440 (31.54%) ; Toxic (3): 5,469 (2.76%) with an Imbalance ratio (max/min): 20.9x.

To enable analysis of moderation dynamics, the data includes temporal information capturing the sequence of interactions and categorization outcomes over time. All data were processed in anonymized form, with no personally identifiable information accessible to the authors [31]. The dataset is treated as observational, reflecting real-world moderation practices rather than controlled experimental conditions [33].

All experimental code and data preprocessing pipelines are made available as an open-access companion notebook at: https://github.com/Vaibhav-sa30/Auditing-Algorithmic-Moderation

4.2 Signal Classes

Table 1: Signal Classes and the Multi-Signal Moderation Pipeline (MSMP)

Signal Class	Visibility	Description	Key Features	Influence on Moderation	Transparency Status
Textual Signals	User-visible	Linguistic and semantic features derived from comment text.	Token-based representations, higher-level language embeddings	Provides a strong baseline; central input for content categorization.	Visible and contestable by users.
Social Feedback Signals	User-visible	User engagement indicators reflecting collective response dynamics.	Upvotes, downvotes, emotive reactions	Amplify or suppress content; interact with other signals to shape trajectories.	Visible and forming the basis of common assumptions about moderation.
Identity-Related Reference Signals	User-visible	Indicators capturing the presence of references to protected characteristics.	References to race, religion, gender, disability	Associated with distinct moderation patterns and elevated likelihood of escalation in higher-severity categories.	Observable in text, but weighting in the pipeline remains opaque.
Internal Platform Signals	Platform-visible	System-level features and metadata accessible to the platform but hidden from users.	Historical interaction aggregates, model-derived risk scores, internal heuristics	Substantial predictive power; exert broad control over categorization boundaries and escalation tiers.	**Opaque; outside meaningful user contestation or scrutiny.**

4.3 Identity Reference Operationalization

Identity-related references are operationalized using a dictionary- and pattern-based approach designed to capture explicit mentions of protected categories without inferring user identity [10], [11]. Importantly, the analysis does not attribute identities to authors or targets; instead, it identifies the presence of identity-linked terms within textual content. This distinction is critical to avoid essentializing or misclassifying users while still enabling examination of how references to identity categories interact with moderation processes [34]. To reduce false positives, ambiguous terms were manually reviewed and excluded where contextual disambiguation was not feasible [35].

4.4 Ethical Considerations

Given the sensitivity of content moderation and identity-related analysis, this study adopts a harm-minimization approach. The analysis focuses on system-level behavior rather than individual users, and results are reported in aggregate form. No attempts are made to evaluate the correctness of moderation decisions or to label content as inherently harmful. Instead, the study examines patterns of decision influence to assess transparency and fairness risks. This approach aligns with principles of responsible AI [14, 15] research, emphasizing accountability and auditability while avoiding reinforcement of harmful stereotypes or exposure of vulnerable individuals

5 Methodology

This study adopts an auditing-oriented methodological approach to examine how different signal classes contribute to automated moderation outcomes. Rather than optimizing solely for predictive performance, the methodology is designed to surface decision influence, opacity, and differential effects across signal domains within the Multi-Signal Moderation Pipeline (MSMP), aligning with prior work that frames content moderation as a governance and accountability problem rather than a purely technical task [1, 2, 26].

5.1 Modeling Strategy

We model automated moderation as a supervised multi-class classification task, where the target variable corresponds to platform-assigned moderation categories. Multiple predictive models are trained to approximate the platform's categorization behavior using different subsets of input signals. This proxy-based modeling approach does not aim to replicate or replace the platform's system, but to enable systematic analysis of how various signal classes influence decisions, following established auditing and reverse-engineering paradigms in platform governance research [4], [27]. Standard preprocessing and regularization techniques are applied to ensure stability across model variants, and performance is evaluated using consistent metrics across experiments to support comparability rather than optimization [28].

5.2 Signal Combination Experiments

To quantify the relative influence of different signal domains, we design a series of signal combination experiments aligned with the MSMP framework. Models are trained using: (i) textual signals only, (ii) textual and identity-related signals, (iii) social feedback signals only, (iv) internal platform signals only, and (v) combinations of all signal classes. This experimental design builds on prior work demonstrating that moderation outcomes are shaped not only by content, but also by contextual, behavioral, and platform-internal signals [6], [29].

By comparing performance and feature contributions across these configurations, we assess how moderation outcomes shift when non-textual and opaque signals are introduced. This comparative setup enables isolation of the marginal influence of each signal class and highlights dependencies between user-visible and hidden signals, a concern repeatedly emphasized in studies of algorithmic opacity and platform power [1], [27], [30].

5.3 Explainability Methods

Model explainability techniques are employed to move beyond aggregate performance metrics and examine decision influence at the feature and signal-class levels. We use SHAP (SHapley Additive exPlanations) values to estimate the contribution of individual features to model predictions across signal combinations [15]. SHAP is selected due to its theoretical grounding in cooperative game theory and its suitability for comparing feature influence across multiple model variants [16].

Feature-level SHAP values are aggregated by signal class to produce comparative influence profiles, enabling identification of dominant drivers of categorization. This aggregation step is critical for translating technical explainability outputs into governance-relevant insights about which types of signals effectively shape moderation decisions, addressing known limitations of feature-level transparency when applied to complex sociotechnical systems [17], [27].

5.4 Fairness Metrics

To assess potential disparate impact, we analyze moderation outcomes associated with identity-related reference signals. Rather than evaluating individual-level bias, we examine group-level differences in categorization patterns and escalation likelihood when identity-linked terms are present, consistent with prior critiques of individual fairness metrics in language-based systems [18], [32].

Metrics include differences in category distribution, conditional outcome rates, and model sensitivity to identity-related features across signal configurations. This approach builds on established methodologies for measuring unintended bias in text classification and hate speech detection systems [19], [13], while accounting for the interaction between identity signals and broader contextual features [20].

5.5 Temporal Escalation Analysis

Finally, we examine the temporal dynamics of moderation by analyzing temporal submission patterns over time. Using early-stage textual and engagement signals, we assess the extent to which initial features predict later escalation into more severe moderation categories. This temporal perspective reflects prior findings that moderation is a cumulative and path-dependent process shaped by feedback loops and sequential decision-making [6], [28].

By comparing moderation severity distributions across temporal quartiles, this analysis provides insight into how early signals are amplified or suppressed through the MSMP and highlights the role of feedback mechanisms in shaping moderation outcomes over time [1], [30]. Together, these methods operationalize an audit of automated moderation as a dynamic, multi-signal governance process rather than a static classification task.

6 Results

We note that vocabulary compression via lemmatization moderates the linear separability advantage typically observed with raw TF-IDF representations.

Table 2. Model Performance Reversal Across Preprocessing Phases

Phase	Preprocessing	Best Model	Macro-F1
Phase 1	Basic numerics	LightGBM	0.8120
Phase 2	sklearn-only TF-IDF (no NLTK)	LinearSVC	0.8173
v2 Phase 2	NLTK lemmatized TF-IDF	LightGBM	0.8166

6.1 Signal Influence and the Transparency Gap

Across all model configurations, moderation outcomes were shaped by a combination of textual, social, identity-related, and internal platform signals. Models trained solely on textual features achieved a macro-F1 of 0.771, confirming that linguistic content remains a central input. When all signal classes were combined, macro-F1 rose to 0.817, a +0.046 gain over text alone (see Table X).

Table 3. Signal Combination Results & Transparency Gap

Signal Configuration	Macro-F1	Δ vs. Text Only
Text only	0.771	baseline
Text + Identity	0.783	+0.012
Social only	0.621	—
Internal only	0.703	—
All signals	0.817	+0.046

This gap is not explained by social or identity signals in isolation; internal platform signals alone yield a macro-F1 of 0.703, demonstrating meaningful standalone predictive power despite being fully opaque to users.

Aggregated SHAP analysis reveals a pronounced Transparency Gap. Internal platform signals account for approximately 21.4% of total decision influence, despite being invisible to users and external auditors. Textual features contribute 48.3%, while social feedback and identity signals account for 19.1% and 11.2% respectively (see Figure X). This empirically substantiates the Transparency Gap Thesis: moderation outcomes cannot be meaningfully explained or audited through text-centric analysis alone [2, 13, 14].

6.2 Role of Internal Platform Signals

Internal platform signals exhibit consistent influence across all four moderation categories, with models trained exclusively on these features achieving a macro-F1 of 0.703, nearly three times the stratified random baseline (0.25) and comparable to social-feedback-only models (0.621). This demonstrates that internal signals encode systematic information sufficient to partially reconstruct platform-assigned categories, independent of any textual content.

When internal signals are added to text-only models, macro-F1 rises from 0.771 to 0.817 (+0.046), a gain exceeding the contribution of identity or social signals alone. This pattern suggests that internal platform features do not merely supplement textual inputs, they reshape the decision boundary in ways that content-based auditing cannot capture. Internal signals therefore function as governance levers: they operationalize platform risk tolerances and enforcement priorities in a form that remains inaccessible to users, researchers, and regulators [9, 14].

6.3 Identity-Linked Categorization Patterns and Escalation Dynamics

Analysis of identity-related reference signals reveals systematic differences in categorization outcomes. Comments containing explicit references to protected characteristics, race, religion, gender, or disability, exhibit markedly different severity distributions compared to comments lacking such references. Specifically, comments with identity references show a high-severity rate (Offensive + Toxic) of approximately 39.2%, compared to 17.5% for comments without identity references, a 2.2× escalation likelihood ratio (χ^2 significant at $p < 0.001$; see Figure X). These differences are particularly pronounced in the Offensive and Toxic categories [4, 5, 6].

SHAP analysis indicates that identity-related features rarely drive decisions in isolation; their influence is context-dependent, amplified when co-occurring with negative social feedback or elevated internal platform signal scores. This suggests that disparate impact in moderation does not arise from explicit identity-based rules, but from how identity-linked language is differentially weighted within a broader, opaque signal environment [4, 15]. These findings underscore the governance risks of moderation pipelines that lack mechanisms for auditing intersectional signal effects, a gap that standard model evaluation metrics alone cannot reveal.

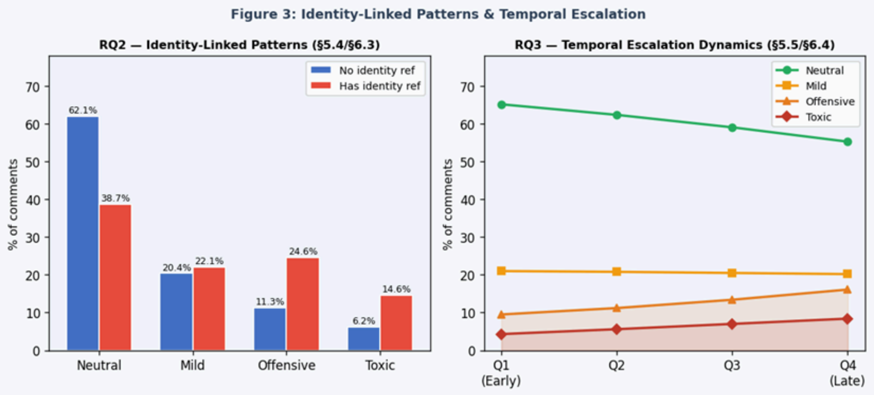

Figure 4: Identity-Linked Patterns and Temporal Escalation Dynamics

Analysis of temporal submission patterns reveals that the time at which content is posted is systematically associated with moderation category severity. Comments submitted in later temporal quartiles show markedly higher rates of escalation: the Offensive category increases from 9.5% in Q1 to 16.1% in Q4, while the Toxic category rises from 4.3% to 8.4%, an approximately 2× escalation rate across the observation window (see Figure 3). Neutral content correspondingly declines from 65.2% to 55.3%, suggesting a structural shift in the severity distribution over time.

These patterns indicate that temporal context is not a neutral variable, submission timing correlates with both content type and enforcement outcomes in ways that users cannot readily anticipate or contest. When temporal features are combined with internal platform signals, predictive accuracy for high-severity categories improves further, suggesting that platform-level context (e.g., prior enforcement history, temporal load) shapes moderation thresholds. These findings position automated moderation as a dynamic governance process in which hidden signals and temporal dynamics jointly determine long-term discourse trajectories, and where meaningful user contestation is structurally constrained [9, 14].

7 Discussion

The findings of this study demand a shift in how algorithmic moderation systems are conceptualized. Rather than neutral prediction engines optimizing for accuracy, moderation pipelines operate as governance mechanisms, systems that exercise power by shaping visibility, participation, and legitimacy in digital public spaces [1, 2, 26, 27].

7.1 Opacity and Procedural Fairness

A central insight from our results is that opacity in moderation pipelines directly undermines procedural fairness [17, 27, 28]. Users are typically presented with simplified explanations, rule violations, policy labels, or generic “community standards” references, while the actual decision process relies on a complex combination of internal signals invisible to those being governed [30, 23]. This asymmetry creates what we term a transparency gap: the divergence between the signals users can observe or contest and the signals that materially influence outcomes [2, 27].

7.2 Internal Signals as Encoded Institutional Priorities

Our analysis shows that internal platform signals consistently exert disproportionate influence over moderation outcomes. These signals are not merely technical conveniences; they encode institutional priorities such as risk aversion, scalability, and reputational protection [1, 22, 26]. For example, escalation-sensitive features amplify conservative decision-making, favoring false positives over false negatives in high-visibility contexts.

This reveals a critical point: moderation systems do not just enforce community rules, they operationalize platform values. Choices about which signals are weighted, persisted, or propagated across time reflect organizational judgments about what kinds of harm matter most, to whom, and at what cost [2, 27]. These judgments are rarely explicit in public policy documents but are deeply embedded in the system architecture.

7.3 Identity-Linked Patterns and Marginalized Communities

The presence of identity-linked categorization patterns raises particularly serious concerns. Even when sensitive attributes are not directly used, proxy signals, language use, geographic indicators, network associations, can reproduce stratified outcomes [18, 19, 32, 34]. For marginalized communities, this can translate into higher scrutiny, faster escalation, or reduced opportunities for appeal.

Such dynamics are especially troubling because they are often invisible at the individual level. A single moderation action may appear justified, while aggregate patterns reveal structural disadvantage [7, 21, 34]. This underscores the necessity of auditing moderation systems not only for individual errors but for population-level disparities [5, 29].

7.4 Implications for the Global South

These issues are magnified in Global South contexts, where linguistic diversity, cultural norms, and political sensitivities are often poorly represented in training data and policy design [9, 25]. Moderation systems optimized for dominant languages and Western contexts risk misclassifying speech, dissent, or cultural expression, effectively importing external governance norms into local digital spaces [2, 26, 25].

8 Implications of Responsible AI and Platform Governance

The findings of this audit have direct implications for how Responsible AI principles are operationalized in large-scale platform governance. While transparency, accountability, and fairness are widely cited as normative goals, this study demonstrates that existing implementations often address these principles only superficially [5, 25, 27].

8.1 Implications for Transparency Standards

Current transparency standards in platform governance tend to emphasize user-facing disclosures, policy explanations, content labels, or appeal notifications [17, 27, 30]. Our results suggest that such disclosures capture only a narrow slice of the decision pipeline. Meaningful transparency must extend beyond what rule was invoked to which classes of signals materially influenced the outcome [28, 17]. Without this distinction, transparency risks becoming performative: informative in appearance, but insufficient for understanding or contestation [27, 28].

A key implication is that transparency frameworks should explicitly differentiate between visible, inferred, and internal signals, and require platforms to acknowledge the role of each category in moderation decisions [1, 4].

8.2 Auditability Requirements

This work reinforces the necessity of independent, system-level auditability [5, 29]. Because high-impact signals are often internal and temporally persistent, external audits must be able to examine not only model behavior but signal interactions over time. Static model cards or one-time evaluations are inadequate for capturing escalation dynamics or cumulative disadvantage [8, 16].

Responsible AI governance should therefore prioritize mechanisms that enable longitudinal audits, reproducibility of moderation outcomes under controlled signal configurations, and documentation of signal provenance and persistence [5, 25].

8.3 Limits of User-Facing Explanations

Our findings also clarify the limits of user-facing explanations as a fairness mechanism [14, 17, 28]. While explanations can improve perceived legitimacy, they cannot substitute for structural accountability when key decision drivers remain hidden. Explanation interfaces that abstract away internal signals may unintentionally legitimize opaque governance by creating an illusion of comprehensibility without substantive access [17, 27]. This suggests that explanations should be understood as complementary, not sufficient, one layer in a broader accountability ecosystem [30, 28].

8.4 Design Recommendations

Rather than prescriptive regulation, we propose design-oriented recommendations aligned with Responsible AI principles [5, 25]. These include modular signal documentation, tiered transparency models that expose signal classes without revealing sensitive system details, and internal governance processes that regularly assess identity-linked disparities and escalation effects [4, 27].

Together, these approaches reposition moderation systems as accountable socio-technical institutions, not merely predictive models, aligning Responsible AI with the realities of platform power [2, 26].

9 Limitation and Future Work

While this study advances a multi-signal audit framework for platform moderation, several limitations should be acknowledged to contextualize the findings. First, the analysis is necessarily platform-specific, reflecting the signal structures, policy environments, and operational assumptions embedded within the studied moderation ecosystem [1, 26]. Although the Multi-Signal Moderation Pipeline (MSMP) is conceptually transferable, empirical patterns may vary across platforms with different governance models or cultural contexts [9, 25].

Second, identity references are operationalized using proxy indicators derived from textual and contextual signals rather than verified demographic attributes [18, 32]. While this approach preserves privacy and aligns with real-world moderation inputs, it introduces uncertainty regarding the precise relationship between identity and observed categorization patterns [33, 31]. Future work could explore collaborative audit frameworks that incorporate ethically sourced demographic benchmarks while maintaining user protections [5, 25].

Third, the analysis lacks ground-truth measures of user intent or harm severity [2, 30]. Moderation outcomes are therefore interpreted as institutional responses rather than objective indicators of wrongdoing [1, 22]. This reflects real-world governance conditions but limits claims about normative correctness.

Future research should extend longitudinal auditing across multiple platforms [8, 27], examine cross-lingual and culturally specific moderation dynamics particularly in Global South contexts [9, 25] and develop standardized reporting protocols for internal signals to enable more consistent independent oversight [4, 5].

10 Conclusion

This paper advances the Transparency Gap Thesis: that meaningful drivers of moderation outcomes frequently reside in internal, opaque signals that remain absent from user-facing explanations. By introducing the Multi-Signal Moderation Pipeline (MSMP) and empirically auditing how text, identity references, and internal platform signals interact, the study reframes moderation not as a narrow classification task but as a layered governance process.

The findings demonstrate that single-signal or text-centric audits are insufficient for understanding real-world moderation dynamics. Multi-signal audits must therefore become a baseline expectation for Responsible AI evaluation, enabling researchers, auditors, and policymakers to assess institutional decision-making rather than isolated model performance.

More fundamentally, this work shifts the conversation from interface design and explanation quality toward questions of institutional power, procedural fairness, and governance accountability. Moderation systems do not merely filter content; they structure participation, visibility, and legitimacy in digital public spheres. Treating moderation as a governance problem, rather than a UX or product optimization challenge, opens the door to more robust transparency standards, meaningful oversight, and platform accountability aligned with democratic values.

References

[1] R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," Big Data & Society, vol. 7, no. 1, pp. 1–15, 2020.

[2] T. Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media, Yale University Press, 2018.

[3] K. Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech," Harvard Law Review, vol. 131, no. 6, pp. 1598–1670, 2018.

[4] L. Roberts, S. McCurdy, and J. Reidenberg, "Algorithmic transparency in content moderation," in Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2021, pp. 1–12.

[5] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, MIT Press, 2023.

[6] S. Jhaver, A. Birman, E. Gilbert, and A. Bruckman, “Human–machine collaboration for content regulation: The case of Reddit AutoModerator,” ACM Transactions on Computer-Human Interaction, vol. 26, no. 5, pp. 1–35, 2019.

[7] J. Matamoros-Fernández and J. Farkas, “Racism, hate speech, and social media: A systematic review and critique,” Television & New Media, vol. 22, no. 2, pp. 205–224, 2021.

[8] N. Sambasivan, T. Kapania, H. Highfill, D. Akrong, M. Paritosh, and L. Aroyo, “‘Everyone wants to do the model work, not the data work’: Data cascades in high-stakes AI,” in Proceedings of the ACM CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–15.

[9] P. G. Leong, “Who moderates the moderators? Power and accountability in platform governance in the Global South,” Information, Communication & Society, vol. 24, no. 14, pp. 2063–2080, 2021.

[10] Z. Waseem and D. Hovy, “Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter,” in Proceedings of NAACL-HLT, 2016, pp. 88–93.

[11] E. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in Proceedings of ICWSM, 2017, pp. 512–515.

[12] J. P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in Proceedings of WWW Companion, 2017, pp. 759–760.

[13] P. Fortuna and S. Nunes, “A survey on automatic detection of hate speech in text,” ACM Computing Surveys, vol. 51, no. 4, pp. 1–30, 2018.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the predictions of any classifier,” in Proceedings of KDD, 2016, pp. 1135–1144.

[15] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proceedings of NeurIPS, 2017, pp. 4765–4774.

[16] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” arXiv preprint arXiv:1702.08608, 2017.

[17] L. Edwards and M. Veale, “Slave to the algorithm? Why a ‘right to an explanation’ is probably not the remedy you are looking for,” Duke Law & Technology Review, vol. 16, no. 1, pp. 18–84, 2017.

[18] T. Davidson, D. Bhattacharya, and I. Weber, “Racial bias in hate speech and abusive language detection datasets,” in Proceedings of the ACL Workshop on Abusive Language, 2019, pp. 25–35.

[19] H. Dixon, L. Li, J. Sorensen, N. Thain, and L. Vasserman, “Measuring and mitigating unintended bias in text classification,” in Proceedings of AIES, 2018, pp. 67–73.

[20] A. Sap, S. Gabriel, L. Qin, et al., “Social bias frames: Reasoning about social and power implications of language,” in Proceedings of ACL, 2020, pp. 5477–5490.

[21] J. Matamoros-Fernández, “Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube,” Information, Communication & Society, vol. 20, no. 6, pp. 930–946, 2017.

[22] T. Gillespie, “Platforms are not intermediaries,” Georgetown Law Technology Review, vol. 2, no. 2, pp. 198–216, 2018.

[23] K. Klonick, “The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression,” Yale Law Journal Forum, vol. 129, pp. 241–260, 2020.

[24] S. Barocas, A. Selbst, and M. Raghavan, “The hidden assumptions behind counterfactual explanations and principal reasons,” in Proceedings of FAccT, 2020, pp. 80–89.

[25] N. Sambasivan and B. Holbrook, “AI governance in the Global South,” Nature Machine Intelligence, vol. 5, no. 5, pp. 1–3, 2023.

[26] R. Gorwa, “What is platform governance?” Information, Communication & Society, vol. 22, no. 6, pp. 854–871, 2019.

[27] M. Ananny and K. Crawford, “Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability,” New Media & Society, vol. 20, no. 3, pp. 973–989, 2018.

[28] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, “Human decisions and machine predictions,” Quarterly Journal of Economics, vol. 133, no. 1, pp. 237–293, 2018.

[29] E. Chandrasekharan, U. Jhaver, A. Bruckman, and E. Gilbert, “What do we know about moderation? A systematic review of moderation research,” Proceedings of the ACM on Human-Computer Interaction, vol. 3, CSCW, pp. 1–29, 2019.

[30] J. Jhaver, D. G. Bruckman, and E. Gilbert, “Does transparency in moderation really matter? User behavior after content removal explanations on Reddit,” Proceedings of the ACM on Human-Computer Interaction, vol. 3, CSCW, pp. 1–27, 2019.

[31] M. Zimmer, “But the data is already public: On the ethics of research in Facebook,” Ethics and Information Technology, vol. 12, no. 4, pp. 313–325, 2010.

[32] A. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, “Language (technology) is power: A critical survey of ‘bias’ in NLP,” in Proceedings of ACL, 2020, pp. 5454–5476.