AI Power and Discourse, vol. 1, núm. 1, 2026, pp. 1-10.

Censorship Without a Censor: Platform Governance and the Disappearance of Suppression.

Agustin V. Startari.

Cita:

Agustin V. Startari (2026). *Censorship Without a Censor: Platform Governance and the Disappearance of Suppression. AI Power and Discourse, 1 (1), 1-10.*

Dirección estable: https://www.aacademica.org/agustin.v.startari/234

ARK: https://n2t.net/ark:/13683/p0c2/u9q

Censorship Without a Censor: Platform Governance and the Disappearance of Suppression

Author: Agustin V. Startari

Author Identifiers

- ResearcherID: K-5792-2016
- ORCID: https://orcid.org/0009-0001-4714-6539
- SSRN Author Page: https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=7639915

Institutional Affiliations

- Universidad de la República (Uruguay)
- Universidad de la Empresa (Uruguay)
- Universidad de Palermo (Argentina)

Contact

- Email: astart@palermo.edu
- Alternate: agustin.startari@gmail.com

Date: August 12, 2026

DOI

- Primary archive: https://doi.org/**10.5281/zenodo.21903591**
- Secondary archive: https://doi.org/10.6084/m9.figshare.33227490
- SSRN: Pending assignment (ETA: Q3 2026)

Language: English

Series: *Grammars of Asymmetric Visibility: AI, Imperial Power, and the Syntax of Responsibility*

Word count: ~~2665~~29448

Keywords: humanitarian passive; responsibility loss; asymmetric visibility; AI-generated conflict discourse; machine neutrality; passive voice; agent deletion; syntactic traceability; civilian suffering; political violence; Palestine; Iran; sanctioned suffering; platform moderation; geopolitical framing; humanitarian discourse; security framing; dominant-power discourse; AI ethics; computational discourse analysis.

Abstract

This article introduces censorship without a censor as a platform-mediated syntactic condition in which political speech is removed, restricted, de-ranked, demonetized, delayed, or otherwise made less visible while the institutional actor responsible for suppression remains grammatically obscured. Extending the concepts of responsibility loss, asymmetric visibility, and sanctioned suffering developed in the preceding articles of this series, the paper argues that automated moderation does not merely classify or regulate content. It also reorganizes the grammar through which suppression becomes describable and institutional responsibility becomes traceable.

The analysis focuses on Palestinian, Iranian, anti-war, anti-sanctions, and anti-imperial speech as test cases for examining agency displacement in platform governance. Moderation notices, policy explanations, and AI-mediated enforcement language frequently describe restrictive actions through passive, agentless, nominalized, or policy-centered constructions: content is removed, visibility is reduced, distribution is limited, accounts are restricted, or policies are violated. In these formulations, suppression remains linguistically visible while the suppressing institution can disappear from the clause. At the same time, users and their content remain explicitly available for classification as violations, risks, threats, or objects of enforcement. The resulting asymmetry produces a double displacement: the speaker becomes increasingly legible as the source of risk while the platform becomes decreasingly legible as the agent of restriction.

To make this phenomenon empirically testable, the paper specifies a full measurement protocol built on~~proposes~~ two complementary measures: the Censor-Deletion Rate (CDR), which quantifies the proportion of suppression-related clauses lacking an explicit institutional agent, and the Suppression Opacity Index (SOI), which measures the broader loss of traceability across institutional agency, decision chains, policy attribution, sanction type, justification, and appeal pathways. These measures shift the analysis of platform governance from the conventional question of which content is permitted or removed toward a second problem: whether the language of moderation preserves the conditions under which restrictive decisions can be attributed to identifiable institutional actors. The

present article establishes that protocol and demonstrates its coding procedure on a pilot set of publicly documented platform notices and Oversight Board decisions.

The central claim is that platform accountability depends not only on the substantive rules governing speech but also on the grammatical visibility of those who enforce them. The pilot also indicates that this visibility varies across platforms rather than following automation as such: notices that name the acting institution and expose an explicit appeal path coexist with visibility-reduction language that does neither. Suppression opacity is therefore better understood as a revisable design choice than as a property of automated enforcement. Censorship without a censor begins when political speech disappears, policy speaks, and the institution responsible for suppression no longer needs to name itself.

1. Introduction: When Speech Disappears Without a Censor~~I. Introduction: When Speech Disappears Without a Censor~~

Political censorship has traditionally been intelligible through an identifiable relation between an actor and an act. A government bans a publication. A ministry prohibits circulation. A court orders material removed. An editor suppresses a passage. A censor deletes a statement. Whatever the legal or political legitimacy of the intervention, the grammatical architecture normally preserves a relatively stable sequence: an institutional actor performs a restrictive action upon an identifiable object of expression. Contemporary platform governance alters this architecture. Political speech can now be removed, restricted, de-ranked, demonetized, excluded from recommendation systems, delayed, account-limited, or otherwise made less visible without the institution responsible for the intervention occupying the grammatical position traditionally associated with the censor.

This transformation is not reducible to the expansion of censorship into digital environments. It concerns a change in how restrictive authority becomes linguistically represented once moderation is executed through platforms, policy systems, classifiers, automated workflows, and institutional procedures. Social media platforms do not merely host expression. They establish and operationalize rules governing visibility, circulation, participation, recommendation, monetization, and removal, making content moderation constitutive of platform governance rather than an exceptional intervention at its margins (Gillespie, 2018; Roberts, 2019). As automated systems become increasingly involved in detecting, classifying, prioritizing, and acting upon content, moderation becomes distributed across technical and institutional layers rather than attributable to a single visible decision-maker (Gorwa et al., 2020).

The linguistic consequence examined in this article begins at precisely this point. A user may be informed that *content was removed*, *visibility was reduced*, *distribution was limited*, *an account was restricted*, or *a policy was violated*. These formulations communicate an enforcement event, but they do not distribute grammatical agency in the same way as statements such as *the platform removed this content* or *the platform restricted the distribution of this post*. The restrictive effect remains visible. The institutional actor responsible for producing that effect does not necessarily remain visible with it.

This article defines that configuration as **censorship without a censor**: a platform-mediated syntactic condition in which political speech is removed, restricted, downgraded, demonetized, delayed, de-ranked, or otherwise made less visible while the institutional actor responsible for the restrictive action is grammatically obscured. The concept does not require the proposition that every act of content moderation constitutes censorship. Nor does it require evidence that platforms intentionally design language to conceal responsibility. Its object is narrower and formally testable: whether an identifiable act of suppression or visibility reduction is represented linguistically without preserving an equally identifiable institutional agent responsible for performing, authorizing, configuring, or maintaining that act.

The distinction matters because suppression can remain fully describable while responsibility for suppression becomes structurally weaker. This reproduces at the level of platform governance a broader phenomenon previously defined as **responsibility loss**: the preservation of an event or consequence alongside the degradation of the grammatical pathway connecting that consequence to a responsible actor (Startari, 2026a). In subsequent work, **asymmetric visibility** extended this problem by showing that visibility is not distributed uniformly across competing actors: a discourse can preserve the agency, intentionality, or threat capacity of one participant while reducing the corresponding grammatical visibility of another (Startari, 2026b). **Sanctioned suffering** further examined how coercive effects may remain empirically present while the actor producing those effects recedes behind abstract categories such as pressure, sanctions, restrictions, or economic conditions (Startari, 2026c). The present article transfers this analytical sequence from AI-generated descriptions of political conflict to the institutional language governing whether political speech circulates at all.

The transfer requires an additional theoretical mechanism. Agent deletion alone cannot explain why the institution disappears so easily from moderation discourse. The disappearance of the censor is not simply a stylistic feature occurring after a decision has been made. It may be enabled by the prior conversion of institutional authority into executable form. The concept of the **compiled rule** provides a formal account of this transformation. In Startari's formulation, a rule ceases to function only as a declarative

statement when its normative structure is translated into constraints capable of selecting, permitting, rejecting, routing, or producing outputs within an operative system (Startari, 2025). Institutional authority is therefore capable of persisting through execution even when the authority that originated the rule is no longer explicitly present at each individual application.

Applied to platform governance, the sequence can be formalized as:

institutional decision → policy formulation → executable classification rule → enforcement condition → restrictive output → moderation notice

At the first stage, institutional agency is comparatively visible. An organization defines what categories of speech are prohibited, restricted, demoted, monetized, recommended, or eligible for distribution. At the second stage, those decisions are encoded into policy language. At the third, policy is translated into operational conditions that may include keyword systems, machine-learning classifiers, confidence thresholds, matching systems, human review instructions, escalation procedures, or combinations of these mechanisms. At the fourth, an input satisfies or is classified as satisfying an enforcement condition. At the fifth, the system produces an effect upon visibility or access. At the sixth, that effect is communicated to the user.

The grammatical disappearance analyzed in this paper occurs principally at the terminal end of that sequence, but its institutional conditions are produced earlier. A decision initially attributable to an organization can travel through successive layers until the final output no longer resembles an institutional action. What begins as *the platform prohibits category X* may become *content matching category X is removed*, and eventually *this content was removed for violating policy*. The originating institution has not ceased to exist. Its authority has been transferred into the operative conditions through which the decision is reproduced.

This is the central connection between the compiled rule and censorship without a censor. Once a discretionary institutional decision has been transformed into an executable rule, subsequent enforcement can be represented as the output of the rule rather than as a renewed action of the institution that authored, parameterized, implemented, and maintains

it. The grammatical disappearance of the censor is therefore not necessarily the disappearance of authority. It can indicate the opposite: authority has become sufficiently embedded in procedure that its continuous linguistic presence is no longer required.

This distinction separates the argument from a conventional critique of passive voice. Passive constructions are not intrinsically deceptive, politically suspect, or responsibility-erasing. They perform legitimate informational functions whenever the affected object is more relevant than the agent, the agent has already been established, or discourse structure makes its repetition unnecessary. The relevant phenomenon is distributional. If suppression-related clauses systematically retain the affected content, the alleged violation, and the policy category while systematically omitting the institutional agent, then grammatical form begins to alter the traceability of governance. What must be measured is not the existence of passive voice in isolation but its relationship to the visibility of decision-making authority.

Three increasingly opaque representations illustrate the mechanism:

The platform removed the post because it classified the content as violating its policy.

The post was removed for violating platform policy.

The content violated policy.

The underlying institutional sequence may be identical, yet the grammatical allocation of responsibility changes substantially. The first formulation preserves the institution as agent, the restrictive act as predicate, the content as affected object, and the classification as an attributed institutional judgment. The second preserves the restrictive act but removes the agent. The third removes both the institutional agent and the enforcement predicate, converting the content itself into the grammatical bearer of violation. What began as an institutional classification of speech is represented as an intrinsic property of the speech.

This movement produces a double displacement. First, institutional agency moves upward into abstraction. *Policy*, *safety*, *standards*, *systems*, or *rules* can occupy the explanatory position previously held by an identifiable institution. Second, responsibility moves downward toward the user or content. The post violates. The account breaches. The content

constitutes risk. The user fails to comply. The platform, meanwhile, can disappear from the immediate grammatical structure in which the restrictive event is explained.

The result is not merely agentlessness. It is an asymmetrical structure of legibility. The speaker remains legible as a policy subject. The content remains legible as a violation object. The classification remains legible as a category. The sanction remains legible as an effect. Yet the institutional actor responsible for defining the category, establishing the threshold, applying the rule, configuring the automated system, and sustaining the enforcement architecture may become progressively less explicit. The speaker becomes increasingly visible as the source of the problem while the governing institution becomes decreasingly visible as the source of the intervention.

This asymmetry becomes politically consequential when moderation is applied to contested political speech. Palestine-related content provides a documented environment in which classification, automation, political context, and platform accountability intersect. Human Rights Watch examined more than 1,050 instances of takedowns or other forms of suppression involving Palestine-related content on Instagram and Facebook during October and November 2023 and identified recurring patterns including content removal, account restrictions, limitations on engagement, and reductions in visibility (Human Rights Watch, 2023). The relevance of these findings for the present study is not that every documented intervention can automatically be classified as censorship. It is that political expression may be governed through multiple visibility-changing mechanisms whose institutional provenance can become difficult for the affected user to reconstruct.

The controversy surrounding Meta's moderation of the Arabic term *shaheed* illustrates the classification problem particularly clearly. In its 2024 policy advisory opinion, the Oversight Board concluded that Meta's approach to the term was overbroad and disproportionately restricted freedom of expression, especially because contextually distinct uses could be treated under the Dangerous Organizations and Individuals policy (Oversight Board, 2024a). The example demonstrates that enforcement does not simply discover a pre-existing category of prohibited speech. Institutional taxonomies define how linguistic material becomes operationally legible to a moderation system. A term becomes

actionable because it has been mapped onto a category, and that category has been connected to an enforcement consequence.

The same problem extends beyond Palestinian speech. Iranian political discourse, anti-sanctions discourse, anti-war advocacy, and anti-imperial speech can intersect with security, extremism, violence, misinformation, dangerous-organization, or safety classifications. These categories are not assumed here to be illegitimate, nor are the political discourses under examination presumed to be inherently protected or benign. The relevant analytical question occurs before such normative judgments: when platforms classify and restrict politically contested expression, does the language communicating those decisions preserve the institution as the actor performing the classification and restriction?

This distinction permits the analysis to avoid two symmetrical errors. The first would be to define all moderation as censorship. That would collapse necessary distinctions among removal of unlawful material, enforcement against direct incitement, restrictions on violent organizational activity, ordinary platform governance, and politically consequential suppression. The second would be to treat the technical or procedural form of moderation as evidence that no censoring agency exists. Automation does not eliminate institutional authorship merely because the proximate execution is performed by a classifier, workflow, or rule. A machine-executed restriction remains institutionally structured when humans or organizations define the categories, thresholds, training conditions, escalation procedures, enforcement consequences, and mechanisms of appeal.

The compiled-rule framework makes that distinction explicit. Execution can become impersonal without becoming institutionally authorless (Startari, 2025). The rule may act operationally, but it did not originate outside institutional design. This is why formulations such as *policy requires*, *the system detected*, *the content violated*, or *the account was restricted* must be analyzed not simply as sentences but as terminal representations of a deeper decision architecture. Grammar can conceal the distance between the visible output and the institutional origin of the rule that generated it.

That distance is the central object of the present study. Conventional platform-governance research usually asks what content platforms remove, what policies govern those decisions, whether enforcement is consistent, and whether users possess adequate mechanisms of transparency and appeal (Gillespie, 2018; Gorwa et al., 2020; Roberts, 2019). The present framework adds a linguistic question: **does the discourse of enforcement preserve the actor responsible for enforcement?**

To make that question measurable, this article introduces the **Censor-Deletion Rate (CDR)** and the **Suppression Opacity Index (SOI)**. CDR measures the proportion of suppression-related clauses in which a restrictive action is described without an explicit institutional agent. SOI extends beyond grammatical subject deletion to measure the broader degradation of accountability traceability, including whether the restrictive action is explicitly named, whether the applied rule is specified, whether the institutional decision-maker remains identifiable, whether the reason for enforcement is concretely stated, whether automation or human review is distinguishable, and whether an appeal pathway is visible. These metrics operationalize a distinction already implicit in the compiled-rule model: the rule can remain executable while the source of authority becomes progressively less visible in the language of execution (Startari, 2025).

The theoretical progression can therefore be stated precisely. **Responsibility loss** explains how consequential events can remain visible while responsible actors become grammatically weakened (Startari, 2026a). **Asymmetric visibility** explains how different actors can receive unequal distributions of agency and intentionality within the same discursive environment (Startari, 2026b). **Sanctioned suffering** explains how coercive effects can be linguistically separated from the agents and decisions producing them (Startari, 2026c). **The compiled rule** explains how institutional authority can migrate from explicit commands into executable constraints whose operation no longer requires continuous restatement of the originating authority (Startari, 2025). **Censorship without a censor** identifies the resulting configuration when that executable authority governs political visibility while the enforcing institution disappears from the grammar used to describe the restriction.

The central hypothesis follows from this sequence. Platform moderation discourse will tend, under identifiable conditions, to preserve the visibility of the moderated speaker, content, violation, risk category, and sanction while reducing the grammatical visibility of the institutional actor responsible for establishing and executing the restriction. If supported empirically, this would demonstrate that content moderation does not regulate only speech. It also regulates the linguistic conditions through which responsibility for regulating speech becomes observable.

The primary research question is therefore not simply **what do platforms remove?** It is more formally constrained: **when political speech is removed, restricted, de-ranked, demonetized, or classified as risk, who remains grammatically visible as responsible for the restrictive act?**

This shift in the unit of analysis changes the problem of platform accountability. The relevant object is no longer only content suppression but the representation of suppression. A post may disappear while its author remains identifiable. A rule may remain visible while its institutional author recedes. A sanction may be communicated while its decision chain becomes opaque. A classifier may appear to detect what an institution has first defined. A policy may appear to speak where an organization has acted.

Censorship without a censor begins at that structural threshold: **speech disappears, the rule executes, policy speaks, and the institution responsible for the conditions of suppression no longer needs to occupy the sentence.**

2. From Responsibility Loss to Suppression Opacity

The disappearance of institutional agency from moderation discourse cannot be adequately described as a problem of passive voice alone. Passive constructions constitute one observable mechanism, but the broader phenomenon concerns the degradation of the relation between an institutional decision, its execution, and its linguistic representation. A moderation notice can identify an enforcement outcome while omitting the actor that produced it; it can identify a policy while obscuring who authored or interpreted that policy; it can identify a violation while suppressing the classification process through which the violation was determined. The resulting condition is therefore not simply grammatical agent deletion. It is a reduction in the traceability of restrictive authority.

This article defines that broader condition as **suppression opacity**. Suppression opacity occurs when the linguistic representation of a restrictive intervention preserves the fact or consequence of enforcement while weakening the informational pathway through which the intervention can be attributed to an identifiable institutional actor, rule-making process, decision mechanism, or chain of accountability. The concept extends responsibility loss from descriptions of political events into the institutional discourse through which platforms explain their own governance operations.

The theoretical point of departure is **responsibility loss**, previously defined as the degradation of the grammatical relation connecting an observable consequence to the actor responsible for producing it (Startari, 2026a). Responsibility loss does not require that an event disappear from discourse. On the contrary, its distinctive structure depends upon the continued visibility of the event. Destruction may remain visible while the actor responsible for destruction disappears. Displacement may be described while the agent of displacement is omitted. Civilian suffering may remain highly salient while the syntax through which responsibility could be assigned becomes weakened. The event survives; attribution deteriorates.

This distinction is essential because absence of information and loss of responsibility are not equivalent. If an event itself is omitted, the resulting problem concerns visibility at the level of inclusion. If the event remains present but the agentive structure connecting it to a

responsible actor is removed or weakened, the problem concerns attribution. Responsibility loss therefore identifies a structural transformation in which discourse retains enough information to establish that something occurred while withholding, backgrounding, abstracting, or grammatically demoting the relation necessary to determine who acted.

The same structure can operate in platform governance. A notice stating that *content was removed* preserves the restrictive event. The user knows that an intervention occurred. What is grammatically absent is the actor that performed or authorized the removal. A notice stating that *the content violated policy* preserves even more information about the institutional classification while simultaneously moving agency away from the institution. The content becomes the grammatical subject of violation; policy becomes the normative reference point; the platform that created, interpreted, operationalized, and enforced the policy may appear nowhere in the clause.

The relevance of responsibility loss to moderation is therefore direct. In both political description and platform governance, the central analytical question is not whether a consequential event remains linguistically represented. It is whether the discourse preserves the relation between consequence and responsible agency (Startari, 2026a). The difference lies in the institutional position of the discourse producer. In conflict summaries, an AI system may describe external events. In moderation discourse, the platform is not merely describing an event produced elsewhere. It is describing an event generated through its own governance architecture.

This distinction makes suppression opacity potentially more consequential than ordinary responsibility loss. In the moderation context, the institution capable of disappearing grammatically may be the same institution whose policies, systems, personnel, or automated mechanisms produced the restrictive intervention. The discourse does not merely fail to identify an external agent. It can obscure the agency of the organization producing the explanation itself.

The transition from responsibility loss to **asymmetric visibility** adds a second analytical dimension. Visibility within political discourse is not distributed only between presence and absence. Different actors can be rendered visible according to different grammatical

properties. One actor may remain visible primarily as victim, another as decision-maker; one may receive explicit intentional verbs while another appears through abstract institutional processes; one may be repeatedly represented as possessing agency while the other is represented primarily as undergoing events. Asymmetric visibility therefore concerns the unequal grammatical allocation of agency, intentionality, causality, responsibility, and threat across actors occupying the same discursive field (Startari, 2026b).

Applied to platform moderation, the relevant asymmetry exists between the governed subject and the governing institution. Moderation notices frequently preserve detailed information about the user, account, post, comment, image, or alleged policy violation. The moderated object is therefore highly visible. The platform may know and communicate which account acted, which item of content triggered intervention, which policy category was implicated, and what sanction followed. Yet the institutional counterpart of that relation can remain substantially less specified.

This produces a structurally unequal distribution of grammatical visibility. The user appears as the possessor of an account, author of content, source of a violation, or object of a sanction. The content appears as prohibited, harmful, dangerous, hateful, misleading, extremist, or otherwise noncompliant. The institutional actor, however, can be represented through abstractions such as *policy*, *standards*, *safety systems*, *our systems*, or simply omitted.

The asymmetry can be represented schematically:

User → visible

Content → visible

Violation → visible

Policy category → visible

Sanction → visible

Institutional decision-maker → potentially obscured

The important property is not that the platform disappears completely from every notice. Many moderation interfaces explicitly identify the company, use first-person plural constructions, or provide policy documentation. The analytical claim is distributional: the linguistic architecture of moderation may allocate greater causal and normative visibility to the regulated subject than to the institution exercising regulatory authority.

This distribution matters because grammatical visibility contributes to accountability. An actor that occupies explicit subject position can more readily be associated with an action. An action expressed through an active transitive predicate makes the relation between actor and affected object comparatively easy to reconstruct. By contrast, passive constructions, abstract nominalizations, policy-centered clauses, and object-as-violator structures can increase the inferential distance between an enforcement outcome and the actor responsible for producing it.

The sequence can again be illustrated through controlled transformations:

We removed your post because we determined that it violated our Dangerous Organizations policy.

The platform appears as grammatical agent twice: it performs the removal and performs the classification.

Your post was removed because it violated our Dangerous Organizations policy.

The removal remains visible, but its agent disappears. The second clause transfers the violation to the post itself.

Your post violated our Dangerous Organizations policy.

The institutional enforcement event disappears. Only the relation between object and rule survives.

Policy violations may result in reduced distribution.

The user and platform both disappear. Policy violation becomes an abstract category linked to an equally abstract consequence.

These sentences may all be compatible with the same underlying institutional operation, but they provide different quantities of information about agency. Their difference is therefore not merely stylistic. It changes how directly a reader can reconstruct the decision chain.

The third component of the theoretical progression, **sanctioned suffering**, clarifies how institutional coercion can become linguistically separated from its human consequences. In previous analysis, sanctions discourse was examined as a structure in which material consequences may remain visible while coercive agency is displaced into terms such as *pressure*, *restrictions*, *economic conditions*, or *sanctions themselves* (Startari, 2026c). The analytical significance of this mechanism lies in the transformation of an action attributable to identifiable decision-makers into an apparently autonomous condition affecting a population.

Platform governance exhibits a structurally comparable transformation. Restrictions imposed upon political speech can be represented as environmental properties of the platform rather than actions performed by the platform. Reach *is limited*. Distribution *is reduced*. Monetization *is unavailable*. Recommendation *is restricted*. An account *is ineligible*. Content *cannot be shown*. The user encounters a changed condition, but the linguistic representation may describe that condition without reconstructing the decision through which it was produced.

This does not establish equivalence between economic sanctions and platform moderation. Their institutional scale, legal character, material effects, and political significance differ substantially. The comparison concerns a formal mechanism: in both cases, coercive or restrictive effects can become linguistically detached from the agentive chain responsible for producing them. An externally imposed condition can consequently acquire the grammatical appearance of an environment.

This transition from action to condition is especially significant for visibility governance. A deleted post presents a comparatively discrete enforcement event. Reduced recommendation, limited distribution, demonetization, search exclusion, or ranking changes can be more difficult to conceptualize as identifiable acts because they modify

conditions of circulation rather than necessarily eliminating content. Platform governance therefore operates not only through prohibitions but through transformations of communicative probability.

A statement can remain technically accessible while losing much of its practical visibility. An account can remain active while becoming less discoverable. A video can remain published while losing monetization or recommendation eligibility. A political message can remain available to existing followers while disappearing from algorithmically mediated discovery. In each case, the distinction between existence and visibility becomes critical.

This is why a binary model of censorship based solely on deletion is insufficient for analyzing contemporary platforms. The relevant institutional object is not only whether speech exists but under what conditions it can circulate. Gillespie (2018) describes content moderation as fundamental to the organization of platform participation rather than as an incidental corrective mechanism. Roberts (2019) similarly demonstrates that commercial content moderation constitutes an extensive institutional infrastructure through which platforms determine what users encounter and what remains excluded from ordinary visibility. Automated moderation further distributes these decisions across technical classifiers, policy systems, review mechanisms, and organizational workflows, complicating conventional assumptions about where individual enforcement decisions are located (Gorwa et al., 2020).

The concept of suppression opacity therefore requires a theory capable of explaining how institutional authority persists even when no individual censor appears at the point of enforcement. This is the function of the **compiled rule**.

The compiled rule describes the conversion of normative or institutional authority into formal constraints capable of producing operational consequences without requiring the originating authority to reissue the command at every execution (Startari, 2025). Its relevance to platform governance lies in the distinction between **authorship of a rule** and **execution of a rule**. A platform may establish a policy at one institutional level, translate that policy into enforcement specifications at another, operationalize those specifications

through automated or human procedures at another, and communicate the final result through standardized language at yet another.

The resulting chain may be represented as:

authority → rule → operational constraint → classification → execution → output

At the level of authority, the institutional actor remains explicit. Someone or some institutional body determines that a category of content should be treated in a particular manner.

At the level of rule formation, that determination becomes a policy.

At the operational level, the policy is translated into criteria capable of guiding human reviewers, automated systems, or hybrid decision processes.

At the classification level, content is mapped onto one or more policy categories.

At the execution level, the corresponding consequence is applied.

At the output level, the platform communicates what happened.

Each transition can reduce the amount of visible institutional agency required for the next stage to function. Once the rule has been compiled into an executable constraint, the system does not need to reproduce the complete institutional history of the rule every time it acts. The enforcement event can therefore appear as a procedural consequence rather than as a newly authored decision.

This does not make the institution irrelevant. It demonstrates precisely how institutional authority can remain effective while becoming linguistically indirect. A rule capable of autonomous or semi-autonomous execution can continue producing institutional consequences even when the originating institution does not appear as the grammatical subject of each output (Startari, 2025).

The compiled-rule framework therefore prevents a category error common in discussions of automated governance: confusing absence of an immediately visible human decision-maker with absence of institutional agency. A classifier may assign a category

automatically. A ranking system may alter distribution without an employee manually reviewing the specific item. A safety system may generate a restriction according to a predetermined threshold. None of these conditions makes the intervention institutionally authorless. The relevant institution establishes or approves the policy architecture, determines acceptable classifications, selects enforcement consequences, configures or authorizes technical systems, maintains their deployment, and defines mechanisms of review.

Automation changes the **proximate executor** of an action. It does not necessarily eliminate the **institutional source of authority** behind that action.

This distinction is central to censorship without a censor. The censor need not disappear because authority has vanished. The censor can disappear because authority has been embedded in rules sufficiently executable that each application no longer requires an explicit institutional command. The institutional decision becomes reproducible.

The linguistic transformation can therefore be reconstructed through three stages.

At the first stage, authority is explicit:

The platform prohibits content classified as X.

At the second, authority becomes procedural:

Content classified as X will be removed.

At the third, authority becomes retrospective and impersonal:

Your content was removed for violating policy X.

The enforcement effect remains. The institutional subject is progressively weakened.

A fourth transformation is possible:

Your content violated policy X.

At this stage, even the restrictive act disappears from the clause. The user encounters a formulation in which the content appears to have generated its own regulatory consequence by possessing the property of noncompliance.

This grammatical transformation can create what may be called **procedural naturalization**: an institutional decision is represented as though it were the predictable consequence of an already existing rule rather than the product of a continuing governance architecture. The restriction appears inevitable because the policy appears antecedent and impersonal.

Yet policies do not interpret themselves. Categories do not determine their own boundaries. Thresholds do not select themselves. Classifiers do not determine independently which errors are institutionally acceptable. Enforcement consequences do not arise naturally from linguistic material. Each depends upon prior design decisions.

The importance of this observation is especially visible in controversial classification environments. Meta's earlier handling of the Arabic term *shaheed* illustrates how a linguistic item can acquire enforcement consequences through institutional taxonomy rather than through an invariant semantic property. The Oversight Board concluded that Meta's prior approach was overbroad and disproportionately restricted expression, demonstrating that the classification of a word as actionable depends upon policy construction and contextual interpretation rather than purely lexical identification (Oversight Board, 2024a).

The institutional significance of the example lies not only in the specific policy outcome. It demonstrates the sequence required for scalable moderation: linguistic material must become recognizable as a policy object before enforcement can occur. This conversion is itself a form of compilation. Language is mapped onto institutional categories; categories are connected to rules; rules are connected to consequences.

The moderation system therefore does not merely identify violations. It operationalizes a grammar of institutional permissibility.

At this point, suppression opacity can be defined more precisely. It is not synonymous with automation, opacity of machine-learning models, lack of transparency, passive voice, or censorship. It is a distinct relation among these possible features.

Suppression opacity is the measurable degradation of the linguistic and procedural traceability linking a restrictive effect on speech to the institutional actor, classification, rule, decision mechanism, and accountability pathway responsible for producing it.

This definition contains several analytically separable dimensions.

The first is **agent visibility**: whether the institution responsible for the restriction appears explicitly.

The second is **action visibility**: whether the restrictive intervention itself is named rather than transformed into a state or condition.

The third is **classification visibility**: whether the category triggering enforcement is specified.

The fourth is **decision visibility**: whether the notice distinguishes between automated detection, human review, hybrid review, or an unspecified system.

The fifth is **rule visibility**: whether the user can identify the exact policy provision invoked.

The sixth is **appeal visibility**: whether a procedure exists through which the decision can be contested and whether that procedure is communicated.

The seventh is **accountability traceability**: whether these components can be reconstructed into a coherent institutional chain.

A notice can therefore be transparent in one dimension while opaque in another. It may clearly identify the policy category while omitting who made the decision. It may identify the platform but not indicate whether the intervention was automated. It may specify that content was removed but provide no precise reason. It may explain the rule while offering no effective means of appeal.

Suppression opacity is consequently cumulative rather than binary.

This motivates the distinction between the two principal measures introduced in this article.

The **Censor-Deletion Rate (CDR)** captures the narrow grammatical phenomenon. It measures the proportion of moderation-related clauses describing a restrictive intervention without an explicit institutional agent responsible for that intervention.

Formally:

CDR = agentless suppression clauses / total suppression clauses

A higher CDR indicates that suppression is more frequently represented without explicit institutional agency.

CDR does not determine whether the underlying moderation was legitimate, illegitimate, accurate, biased, necessary, lawful, or censorial in a normative sense. It measures only whether the agent responsible for the restrictive act remains grammatically visible.

The **Suppression Opacity Index (SOI)** addresses the broader institutional problem. It measures the degree to which a moderation communication obscures the decision chain connecting content, classification, rule, institutional actor, enforcement mechanism, sanction, and appeal pathway.

CDR therefore identifies one possible component of SOI, but the two cannot be collapsed.

A notice can omit the grammatical agent while remaining institutionally transparent:

Your post was removed under Section X of our policy after human review. You may appeal this decision here.

The passive construction increases agent deletion at the clause level, but the broader decision chain remains comparatively visible.

Conversely, an active construction can preserve the grammatical agent while remaining procedurally opaque:

We restricted your account because it violated our policies.

The platform appears as agent, but the policy, classification mechanism, exact reason, decision pathway, and review procedure may remain unspecified.

CDR therefore measures **grammatical disappearance**.

SOI measures **institutional traceability loss**.

The distinction prevents the analysis from reducing accountability to syntax alone. Grammar is the entry point because it can be consistently coded at clause level, but the relevant theoretical object is the visibility of authority across the entire enforcement representation.

This framework also clarifies the relation between censorship without a censor and the compiled rule. The compiled rule explains how authority can remain operative despite reduced linguistic presence. CDR identifies how frequently the actor disappears from the resulting grammatical representation. SOI estimates how much institutional traceability is lost across the complete explanation of the intervention.

The theoretical architecture can therefore be expressed as:

Compiled authority → executable rule → moderation output → agent deletion → reduced traceability → suppression opacity

Not every compiled rule produces agent deletion.

Not every agentless sentence produces substantial suppression opacity.

Not every moderation intervention constitutes censorship.

The model therefore does not depend upon treating these relations as necessary equivalences. It proposes testable transitions whose frequency and strength must be established empirically.

This qualification is important because the purpose of the framework is not to pre-classify platform moderation as illegitimate. It is to make one dimension of institutional governance measurable independently of political judgment. A moderation decision can be substantively justified and still be communicated opaquely. A moderation decision can be

substantively questionable and communicated transparently. The legitimacy of the restriction and the transparency of its representation are analytically separate variables.

This separation is particularly valuable for politically contested cases. Claims of platform bias concerning Palestine, Iran, sanctions, war, imperial power, or security often generate immediate disagreement about political context and normative legitimacy. A linguistic measure does not resolve those disputes, but it permits a narrower question to be tested without first settling them: **does the institution remain visible as the actor of restriction?**

That question can be answered through corpus analysis.

If platform notices systematically employ structures in which users and content receive explicit grammatical roles while institutional actors are absent, the pattern can be quantified.

If politically sensitive categories generate higher rates of agent deletion than comparable nonpolitical moderation categories, the difference can be tested.

If notices involving security or extremism classifications show greater suppression opacity than ordinary spam or copyright notices, the contrast can be measured.

If appeal pathways correlate with lower SOI scores, that relation can be observed.

If automated notices exhibit higher CDR than human-reviewed explanations, the distinction can be tested.

The theoretical progression from responsibility loss to suppression opacity therefore produces falsifiable expectations rather than merely interpretive claims.

The broader contribution is a shift in how accountability is conceptualized. Platform transparency has often been approached through disclosure statistics, policy publication, transparency reports, appeals procedures, algorithmic explanation, or access to data. These remain essential. The present framework adds a more elementary level: the grammar through which individual acts of governance become intelligible to the governed subject.

Before users can contest a decision, they must be able to reconstruct that a decision occurred.

Before responsibility can be assigned, an actor must remain identifiable.

Before an appeal can challenge a classification, the classification must be visible.

Before automated governance can be scrutinized, its institutional relation to the platform must be linguistically recoverable.

Suppression opacity begins where these relations deteriorate.

The central theoretical claim of this section can therefore be stated directly: **the disappearance of the censor is not equivalent to the disappearance of authority. Under compiled governance, authority can become more operationally efficient precisely as its grammatical presence becomes less necessary.**

Platform moderation provides an unusually clear environment in which to test this proposition. Institutional decisions are converted into policies; policies into operational rules; rules into classifications; classifications into enforcement outputs; outputs into standardized notices. At every transition, authority remains operative, but its visible linguistic representation can contract.

The result extends the earlier sequence of the series. Responsibility loss showed that consequences can survive the disappearance of responsible actors (Startari, 2026a). Asymmetric visibility showed that agency can be distributed unequally across actors (Startari, 2026b). Sanctioned suffering showed that coercive effects can acquire the linguistic appearance of impersonal conditions (Startari, 2026c). The compiled rule explains how authority can remain effective after being translated into executable constraint (Startari, 2025).

Suppression opacity brings these mechanisms together.

The speech remains identifiable.

The alleged violation remains identifiable.

The restriction remains effective.

The rule remains executable.

The institution becomes less visible.

The censor has not necessarily ceased to act. It has become structurally unnecessary for the censor to remain in the sentence.

3. The Grammar of Platform Enforcement

Platform moderation becomes institutionally consequential not only through the rules it applies but through the linguistic forms in which those rules are communicated. Once an enforcement decision has been produced, the platform must represent that decision to the affected user. At this stage, a technical or institutional event is converted into language. The resulting notice may identify the platform, the policy, the content, the sanction, the classification process, and the possibility of appeal. It may also omit several of these elements. The grammar of enforcement therefore constitutes a distinct layer of platform governance: it determines how much of the underlying decision architecture remains visible after the decision has already been executed.

This section examines that layer at clause level. Its object is not whether a moderation decision is substantively correct, politically biased, legally justified, or technically accurate. Those questions require separate evidence. The narrower question is how enforcement events are syntactically represented and how recurrent grammatical structures redistribute visibility between the institution exercising authority and the subject governed by that authority.

The distinction follows directly from the compiled-rule framework developed in the preceding section. Platform policies can be transformed from declarative institutional norms into executable constraints capable of producing classifications and sanctions at scale (Startari, 2025). Once this transformation occurs, the linguistic explanation presented to the user no longer needs to reproduce the institutional chain through which the rule

acquired force. The notification can begin at the end of the process. Rather than stating who established the rule, who interpreted the relevant content, which system classified it, and who authorized the consequence, the notice can simply report that a violation occurred.

This compression produces a characteristic grammatical architecture. The institutional decision chain may contain multiple stages:

platform → policy → operational rule → classifier or reviewer → classification → sanction

The moderation notice can reduce this sequence to:

content → violation → consequence

The reduction is informationally efficient. It is also analytically significant. Institutional complexity does not disappear from the enforcement process merely because it disappears from the sentence. The grammar of the notice determines which components of that complexity remain available to the governed subject.

3.1 Enforcement as grammatical transformation

A moderation event can be represented through several grammatical configurations without changing the underlying enforcement outcome.

Consider an active formulation:

We removed your post because we determined that it violated our policy on dangerous organizations.

The sentence identifies an institutional agent through *we*. It distinguishes the act of removal from the act of classification. The platform performs both actions: it determines that the content belongs to a prohibited category and subsequently removes it. The content does not independently violate the policy; rather, the institution classifies it as doing so.

A passive transformation produces:

Your post was removed because it violated our policy on dangerous organizations.

The result of enforcement remains explicit, but the actor responsible for removal disappears. The causal subordinate clause then assigns violation directly to the post. The institution remains indirectly recoverable through the possessive *our*, but its role in classification is no longer grammatically expressed.

A further reduction produces:

Your post violated our policy on dangerous organizations.

Now both the restricting agent and the restriction itself disappear from the principal proposition. The content occupies subject position and *violated* functions as the predicate. The institutional judgment through which the content was classified has been converted into an apparently direct relation between an object and a rule.

Finally:

Policy violations may result in reduced distribution.

Here neither platform nor user needs to appear. The sentence constructs an abstract causal relationship between *policy violations* and *reduced distribution*. Restrictive governance is represented as a generalized procedural consequence.

The progression can therefore be represented as:

institution acts on content

→ content is acted upon

→ content violates

→ violation produces consequence

Each transformation reduces the explicit grammatical role of institutional agency.

The important point is not that any single transformation is inherently manipulative. Passive voice, nominalization, and abstraction are ordinary linguistic resources. The analytical question concerns their distribution across enforcement discourse. If the same transformations systematically preserve the governed object while reducing the visibility

of the governing subject, then the grammatical pattern becomes relevant to institutional accountability.

A parallel mechanism operates through nominalization. We restricted the account contains an actor and an action; the account restriction remains in effect converts the action into a noun that can function as an object or state without requiring the sentence to identify who produced it. Removal, restriction, violation, enforcement, limitation, reduction, suspension, classification, review, and ineligibility all package institutional actions into nominal units, allowing procedural discourse to discuss enforcement outcomes while suppressing the predicates that would otherwise require an agent and an affected object. Following review, the restriction remains in place contains two nominalized institutional processes and identifies no actor; its explicit equivalent, after our moderation team reviewed the post, we decided to continue restricting its distribution, reconstructs the institutional chain. Nominalization therefore contributes to suppression opacity when it converts decision processes into states whose origins are no longer syntactically recoverable (Startari, 2026a).

The cumulative effect of these transformations is that the decision itself can disappear. A fully reconstructed enforcement event would minimally contain: Actor A classified content C under rule R and imposed sanction S through mechanism M. The notice may instead contain: C violated R; S applies. The decision is no longer represented as an event but as the invisible relation connecting violation and consequence. This matters because accountability normally requires a decision that can be challenged. Appeals do not meaningfully contest abstract states; they contest determinations made through institutional processes. If the decision disappears linguistically, the user may still know that enforcement occurred while possessing considerably less information about what exactly must be disputed.

This distinction follows the broader argument developed in *The Grammar of Objectivity*, where apparently impersonal forms can shift the perceived location of judgment by presenting institutional selections as properties of the described object rather than as acts of selection performed by a situated authority (Startari, 2025b). In moderation discourse, the same formal problem becomes operational. A platform does not merely describe a violation. It establishes the categories under which the event becomes recognizable as a violation and possesses the authority to impose the corresponding consequence.

3.2 Agent deletion: passive enforcement, violation inversion, and policy substitution~~Passive enforcement~~

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Line spacing: Multiple 1,15 li

Passive constructions constitute the most immediately observable mechanism of censor deletion.

In a canonical active clause:

Platform X removed post Y.

The grammatical subject corresponds to the institutional agent.

In its passive equivalent:

Post Y was removed.

The affected object becomes grammatical subject while the institutional agent can be omitted completely.

The event remains semantically recoverable: removal occurred. What is no longer obligatorily encoded is who performed the removal.

This structure is especially compatible with moderation interfaces because users typically encounter enforcement from the perspective of the affected object. The post, video, comment, image, account, or profile occupies the communicative center. Consequently, formulations such as *your content was removed*, *your account was restricted*, or *distribution was reduced* are functionally natural.

Yet their informational economy has a measurable consequence. The restrictive action is preserved while the enforcing actor is absent.

In terms of responsibility loss, the structure is straightforward:

effect visibility = preserved

agent visibility = reduced

(Startari, 2026a).

This relationship becomes particularly important when passive structures occur repeatedly across an institutional corpus. A single passive clause cannot demonstrate systematic

suppression opacity. A high proportion of agentless passive constructions across enforcement notices, however, would indicate that users are routinely informed about restrictive outcomes without equivalent linguistic identification of the actor responsible for producing them.

For this reason, passive constructions constitute one of the principal variables incorporated into the Censor-Deletion Rate. The relevant coding question is not merely whether a clause is passive, but whether the agent responsible for the restrictive action remains linguistically recoverable within the clause or its immediate explanatory context.

The distinction prevents false positives. Compare:

Your post was removed by Meta because it violated the Dangerous Organizations and Individuals policy.

and:

Your post was removed because it violated policy.

Both are passive. Only the second deletes the institutional agent.

Passive ratio and censor-deletion rate therefore cannot be treated as equivalent measures.

A more consequential transformation occurs when enforcement discourse moves from describing an institutional classification to attributing violation directly to content. Institutionally the sequence runs: institution defines rule, content is evaluated, institution or authorized system classifies content, consequence follows. Grammatically it can become simply content violates rule. This transformation may be called violation inversion: the object subjected to institutional classification becomes the grammatical source of the transgression while the classifier disappears. The contrast is visible between we determined that this post violates our policy and this post violates our policy. The first marks violation as an institutional determination; the second presents it as a direct property of the post. The second formulation is not logically false, since institutions routinely formulate normative conclusions in declarative language, as when a court states that an action violates a statute. The relevance is structural: when such formulations dominate standardized moderation notices, the classificatory judgment becomes detached from the actor making it.

Formatted: Font: (Default) Times New Roman, 12 pt

This matters because moderation categories are rarely natural properties of linguistic objects. Harassment, misinformation, dangerous-organizations support, hate speech, glorification, incitement, and harmful content depend upon definitions, contextual thresholds, policy exceptions, and interpretive decisions, and automated moderation requires those definitions to be rendered sufficiently operational for computational or hybrid enforcement (Gorwa et al., 2020). Content does not enter a platform already carrying an intrinsic label of violation. It becomes a violation within a specific governance system after being mapped onto an institutional category. Once that category has been translated into operational criteria, the classificatory relation can be reproduced without restating the institutional decisions embedded in it (Startari, 2025), so that the linguistic output reverses the actual institutional direction of authority: operationally the platform classifies content, linguistically content violates platform.

A related mechanism grants policy itself quasi-agentive status. Our policies do not allow this content, this content is prohibited under our standards, and our rules require removal retain more institutional information than fully agentless notices, since the platform remains recoverable through possessive markers such as our, but they shift immediate causal attention from the institution to the normative instrument. This policy-agent substitution converts institution creates and enforces policy into policy determines outcome. The pattern is not unique to digital platforms; bureaucratic discourse routinely attributes action to regulations, procedures, or systems. It becomes theoretically important under compiled governance because these structures genuinely possess operational force: a rule encoded into a system can determine what outputs the system permits. That operational capacity should not be confused with independent authorship, since the rule acts only because institutional authority has been compiled into its structure (Startari, 2025). Policy-agent substitution therefore occupies an intermediate position between explicit institutional agency and complete agent deletion, and the progression is measurable because the grammatical subject changes while the enforcement effect remains stable: we restrict recommendation of this material, our policy restricts recommendation of this material, recommendation of this material is restricted under policy.

~~**3.3 The violation inversion**~~

~~A more consequential transformation occurs when enforcement discourse moves from describing an institutional classification to attributing violation directly to content.~~

~~Institutionally, the sequence is approximately:~~

~~**institution defines rule → content is evaluated → institution or authorized system classifies content → consequence follows**~~

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Line spacing: Multiple 1,15 li

~~Grammatically, the sequence can become:~~

~~**content violates rule**~~

~~The transformation may be called **violation inversion**. The object subjected to institutional classification becomes the grammatical source of the transgression, while the classifier disappears.~~

~~This difference is visible in the contrast between:~~

~~*We determined that this post violates our policy.*~~

~~and:~~

~~*This post violates our policy.*~~

~~The first sentence explicitly marks violation as an institutional determination. The second presents violation as a direct property of the post.~~

~~The distinction does not imply that the second statement is logically false. Institutions routinely formulate normative conclusions in declarative language. A court may state that an action violates a statute; a regulator may state that a company breaches a rule. The relevance here is structural: when such formulations dominate automated or standardized moderation notices, the classificatory judgment becomes detached from the actor making it.~~

~~This is especially significant because moderation categories are rarely natural properties of linguistic objects. Categories such as harassment, misinformation, dangerous-organizations support, hate speech, glorification, incitement, graphic violence, or harmful content depend upon definitions, contextual thresholds, policy exceptions, and interpretive decisions. Automated moderation adds another layer because those institutional definitions must be rendered sufficiently operational for computational or hybrid enforcement (Gorwa et al., 2020).~~

A piece of content therefore does not enter a platform already carrying an intrinsic label of *violation*. It becomes a violation within a specific governance system after being mapped onto an institutional category.

The compiled rule explains how this mapping can subsequently appear automatic. Once the category has been translated into operational criteria, the classificatory relation can be reproduced without restating the institutional decisions embedded in it (Startari, 2025).

The linguistic output can then reverse the actual institutional direction of authority.

Operationally:

platform classifies content

Linguistically:

content violates platform

This inversion is central to censorship without a censor because it leaves the regulated object visible while converting institutional judgment into an apparently inherent property of that object.

3.4 Policy as grammatical actor

A related mechanism occurs when policy itself acquires quasi-agentive status.

Platform communications can represent policies, rules, standards, or safety requirements as explanatory sources of enforcement:

Our policies do not allow this content.

This content is prohibited under our standards.

Our rules require removal.

Safety policies restrict this type of material.

These formulations retain more institutional information than fully agentless notices because the platform may remain recoverable through possessive markers such as *our*.

Nevertheless, they shift immediate causal attention from the institution to the normative instrument.

This shift can be termed **policy-agent substitution**.

The underlying institutional relationship is:

institution creates and enforces policy

The linguistic representation becomes:

policy determines outcome

The policy consequently occupies the functional position of an actor despite lacking independent institutional agency.

This transformation is not unique to digital platforms. Bureaucratic discourse frequently attributes actions to regulations, procedures, requirements, or systems. Statements such as *the rules require*, *the system does not permit*, or *the procedure mandates* redistribute agency from institutional decision-makers toward formalized normative structures.

The distinction becomes theoretically important under compiled governance because these structures genuinely possess operational force. A rule encoded into a system can determine what outputs the system permits. Yet this operational capacity should not be confused with independent authorship. The rule acts only because institutional authority has been compiled into its structure (Startari, 2025).

Policy-agent substitution therefore occupies an intermediate position between explicit institutional agency and complete agent deletion.

Compare:

We restrict recommendation of this material.

Our policy restricts recommendation of this material.

Recommendation of this material is restricted under policy.

The first sentence assigns agency directly to the platform.

The second transfers immediate agency to policy while retaining institutional possession.

The third removes the platform and transforms the enforcement result into a state regulated by an abstract rule.

The progression is measurable because the grammatical subject changes while the enforcement effect remains stable.

3.5 Nominalization and the conversion of decisions into states

Nominalization provides another mechanism through which actions can lose their connection to actors.

Compare:

We restricted the account.

with:

The account restriction remains in effect.

The first formulation contains an actor and an action. The second converts the action *restrict* into the noun *restriction*. Once nominalized, the enforcement event can function as an object or state without requiring the sentence to identify who produced it.

The same occurs with terms such as:

removal

restriction

violation

enforcement

limitation

reduction

suspension

classification

review

ineligibility

Each can package an institutional action into a nominal unit.

This is structurally important because nominalization permits procedural discourse to discuss enforcement outcomes while suppressing the predicates that would otherwise require arguments such as agent and affected object.

For example:

Following review, the restriction remains in place.

The sentence contains two nominalized institutional processes: *review* and *restriction*. Neither identifies an actor.

A more explicit equivalent would require substantially more institutional information:

After our moderation team reviewed the post, we decided to continue restricting its distribution.

The difference is not merely length. The expanded sentence reconstructs the institutional chain.

Nominalization therefore contributes to suppression opacity when it converts decision processes into states whose origins are no longer syntactically recoverable.

This is consistent with the broader account of responsibility loss developed in the previous articles of the series. Nominalization permits actions to remain semantically present while reducing the grammatical necessity of representing the actors performing them (Startari, 2026a).

3.3 Procedural inevitability and system agency3.6 Procedural inevitability

Passive enforcement, violation inversion, policy-agent substitution, and nominalization converge in a broader effect: **procedural inevitability**.

Procedural inevitability occurs when an institutionally contingent decision is represented as the predictable or necessary consequence of a rule.

The structure can be represented as:

if X → Y

where X is a policy classification and Y is an enforcement consequence.

Once the relationship is represented as procedural, the institutional choice embedded in the mapping can disappear.

For example:

Content in category X receives reduced distribution.

The formulation presents the relationship as a stable property of the governance system. Missing from the clause are several prior institutional decisions:

the definition of category X;

the choice to associate category X with reduced distribution;

the determination of the relevant threshold;

the selection of an automated or human classification mechanism;

the choice of available exceptions;

the decision regarding whether and how appeals may reverse enforcement.

The rule therefore appears deterministic only after institutional discretion has already been compiled into it.

This is one of the central properties of executable governance. Authority becomes less visible at the moment of execution because earlier discretionary decisions have been stabilized into a formal relationship between condition and outcome (Startari, 2025).

The apparent inevitability of enforcement should consequently be distinguished from its institutional origin.

The system restricted the account is not institutionally equivalent to a restriction without an author.

Policy required removal does not mean that policy independently acquired normative authority.

The content violated the rules does not mean that classification occurred without interpretation.

The grammar of procedural inevitability compresses these distinctions.

Automated moderation introduces a further possible grammatical actor: the system. Statements such as our systems detected a violation, the system identified this content as harmful, or automated technology removed the post appear more transparent than fully agentless constructions because an executor is named. Yet the grammatical presence of a technical system introduces a distinction between proximate execution and institutional responsibility. A classifier may detect a pattern, a model may assign a probability, a rule engine may trigger an enforcement action, and a recommender system may reduce distribution, but none of these operations independently establishes why the relevant category exists, why a particular threshold is used, or why the classification generates a specific sanction. Automated moderation is best understood as a sociotechnical governance arrangement rather than as an autonomous institutional actor (Gorwa et al., 2020): its outputs depend upon policy definitions, training data, system architecture, thresholds, escalation protocols, and organizational decisions concerning acceptable error.

Replacing we restricted with our systems detected can therefore increase technical transparency while simultaneously redistributing responsibility toward the technological intermediary. This does not make the formulation deceptive, since it may accurately describe the proximate mechanism; the problem arises when the technological mechanism becomes the terminal explanation for an institutional decision. Two forms of agency must consequently be distinguished. Technical agency identifies what component

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Line spacing: Multiple 1,15 li

executed or produced a classification. Institutional agency identifies the organization responsible for authorizing and governing the system through which that classification acquired consequences. Suppression opacity increases when technical agency is specified in a manner that makes institutional responsibility harder rather than easier to reconstruct, and this distinction is incorporated into SOI coding below: a notice identifying automated detection but failing to identify the institutional rule or accountability path does not receive the same transparency score as one specifying both.

~~**3.7 System agency and the technological intermediary**~~

~~Automated moderation introduces another possible grammatical actor: *the system*.~~

~~Statements such as:~~

~~*Our systems detected a violation.*~~

~~*The system identified this content as harmful.*~~

~~*Automated technology removed the post.*~~

~~appear more transparent than completely agentless constructions because an executor is identified. Nevertheless, the grammatical presence of a technical system introduces a distinction between **proximate execution** and **institutional responsibility**.~~

~~A classifier may detect a pattern.~~

~~A model may assign a probability.~~

~~A rule engine may trigger an enforcement action.~~

~~A recommender system may reduce distribution.~~

~~None of these operations independently establishes why the relevant category exists, why a particular threshold is used, or why the classification generates a specific sanction.~~

~~Automated moderation is best understood as a sociotechnical governance arrangement rather than as an autonomous institutional actor (Gorwa et al., 2020). Its outputs depend upon policy definitions, training data, system architecture, implementation choices,~~

~~thresholds, escalation protocols, and organizational decisions concerning acceptable error (Gorwa et al., 2020).~~

~~Consequently, replacing *we restricted* with *our systems detected* can increase technical transparency while simultaneously redistributing responsibility toward the technological intermediary.~~

~~This does not make the formulation deceptive. It may accurately describe the proximate mechanism. The problem arises when the technological mechanism becomes the terminal explanation for an institutional decision.~~

~~A useful distinction is therefore required:~~

~~**technical agency** identifies what component executed or produced a classification;~~

~~**institutional agency** identifies the organization responsible for authorizing and governing the system through which that classification acquired consequences.~~

~~Suppression opacity increases when technical agency is specified in a manner that makes institutional responsibility harder rather than easier to reconstruct.~~

~~This distinction will later be incorporated into SOI coding. A notice identifying automated detection but failing to identify the institutional rule or accountability path should not receive the same transparency score as one specifying both.~~

3.4 Asymmetric attribution: user-as-violator and content-as-risk~~3.8 User-as-violator constructions~~

If institutional agency tends to move upward toward policies and systems, user responsibility can simultaneously become increasingly explicit.

Moderation language frequently addresses the recipient directly:

You violated our standards.

Your account breached our rules.

You posted content that violates policy.

Repeated violations may result in suspension.

These formulations have a strong governance function. They identify the regulated subject and establish a connection between behavior and possible sanction.

Their analytical importance lies in the contrast they can produce with the representation of platform action.

A notice may therefore contain:

explicit user agency + implicit platform agency

The user *violated*.

The platform's response *was applied*.

The account *was restricted*.

The policy *requires* compliance.

This is a paradigmatic instance of asymmetric visibility. One side of the governance relation is represented through active predicates, while the other is represented through passive, procedural, or abstract structures (Startari, 2026b).

The asymmetry can be quantified through what this article terms the **user-as-violator ratio**: the frequency with which users or their content occupy grammatical subject position in violation clauses relative to the frequency with which platforms occupy explicit subject position in enforcement clauses.

The measure is not intended to establish unfairness independently. A user who posts prohibited material may accurately be described as violating a rule. The metric instead asks whether institutional discourse systematically expresses the agency of the regulated actor more explicitly than the agency of the regulating actor.

A closely related pattern converts political expression grammatically into a risk object. Content may be categorized as dangerous, hateful, violent, misleading, harmful, extremist, or unsafe. These categories may correspond to legitimate moderation objectives, but their grammatical form deserves attention because adjectival classification

Formatted: Font: (Default) Times New Roman, 12 pt

can erase the act of classification itself. We classified this post as potentially harmful preserves epistemic and institutional mediation: a platform made a classification, and the classification can theoretically be contested. This post is harmful presents the category as a property of the object. The difference is especially significant in politically contested environments, where meaning can depend heavily upon context, quotation, documentation, counterspeech, newsworthiness, political terminology, multilingual variation, or the distinction between description and endorsement.

The moderation controversy concerning Arabic shaheed illustrates why categorical properties cannot be treated as lexically self-evident. Meta's previous approach was criticized by its Oversight Board for applying an overbroad restriction that did not sufficiently preserve contextual distinctions (Oversight Board, 2024a), demonstrating that risk categories are produced through institutional classification practices rather than discovered in content. From the perspective of compiled governance the transition is again predictable: the institution defines a category, operational criteria identify it, and the output represents content as that category, with the first two stages disappearing from the final grammatical representation. This content is dangerous therefore contains less information about governance than we classified this content as falling under our dangerous-organizations policy, and the difference is measurable as classification visibility.

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Line spacing: Multiple 1,15 li

~~**3.9 Content-as-risk constructions**~~

~~A closely related pattern occurs when political expression is converted grammatically into a risk object.~~

~~Content may be categorized as:~~

~~*dangerous*~~

~~*hateful*~~

~~*violent*~~

~~*misleading*~~

~~*harmful*~~

~~*extremist*~~

~~*unsafe*~~

These categories may correspond to legitimate moderation objectives. Their grammatical form nevertheless deserves attention because adjectival classification can erase the act of classification itself.

Compare:

We classified this post as potentially harmful.

with:

This post is harmful.

The first formulation explicitly preserves epistemic and institutional mediation. A platform made a classification, and the classification itself can theoretically be contested.

The second formulation presents the category as a property of the object.

This difference is especially significant in politically contested environments, where meaning can depend heavily upon context, quotation, documentation, counterspeech, newsworthiness, political terminology, multilingual variation, or distinctions between description and endorsement.

The moderation controversy concerning Arabic *shaheed* illustrates precisely why categorical properties cannot always be treated as lexically self-evident. Meta's previous approach was criticized by its Oversight Board for applying an overbroad restriction that did not sufficiently preserve contextual distinctions (Oversight Board, 2024). The case demonstrates that risk categories are produced through institutional classification practices, not simply discovered in content.

From the perspective of compiled governance, the transition is again predictable:

institution defines category → operational criteria identify category → output represents content as category

The first two stages disappear from the final grammatical representation.

The sentence *this content is dangerous* therefore contains less information about governance than *we classified this content as falling under our dangerous organizations policy*.

The difference is measurable as **classification visibility**.

3.10 The disappearing decision

Taken together, these mechanisms produce a distinctive transformation: the decision itself can disappear.

A fully reconstructed enforcement event would minimally contain:

Actor A classified content C under rule R and imposed sanction S through mechanism M.

The moderation notice may instead contain:

C violated R; S applies.

The decision is no longer represented as an event. It has become the invisible relation connecting violation and consequence.

This transformation is central to suppression opacity because accountability normally requires the existence of a decision that can be challenged. Appeals do not meaningfully contest abstract states. They contest determinations made through institutional processes.

If the decision disappears linguistically, the user may still know that enforcement occurred but possess less information about what exactly must be disputed.

This produces a hierarchy of increasingly compressed governance representations:

Institutional decision

→ We reviewed your content and decided to remove it under Rule X.

Institutional action

→ We removed your content under Rule X.

~~**Agentless action**~~

~~*→ Your content was removed under Rule X.*~~

~~**Object violation**~~

~~*→ Your content violated Rule X.*~~

~~**Procedural consequence**~~

~~*→ Violations of Rule X result in removal.*~~

~~At each stage, the underlying authority can remain identical. What changes is the grammatical visibility of the decision chain.~~

3.5 From linguistic economy to accountability surface~~3.11 From linguistic economy to accountability loss~~

None of these forms can be evaluated solely through intuition. Standardized moderation interfaces face severe constraints of scale, localization, readability, notification length, legal consistency, and interface design. Linguistic economy therefore has legitimate operational explanations.

The hypothesis advanced here is not that concision itself constitutes censorship or concealment. It is that linguistic economy becomes relevant to accountability when information is removed asymmetrically.

A short notice can remain highly traceable:

We removed this post under Rule X after automated detection. You can appeal here.

A longer notice can remain opaque:

Our commitment to community safety requires enforcement of policies designed to ensure that everyone can participate safely, and content that does not comply with these standards may be subject to restrictions.

Length is therefore not the variable of interest.

The relevant variable is whether the notice preserves the structural components necessary to reconstruct governance:

Who acted?

What action occurred?

What rule was applied?

Who or what classified the content?

What classification was assigned?

What consequence followed?

Can the decision be challenged?

A moderation notice that preserves these relations has low suppression opacity even if its language is concise.

A notice that removes them has higher suppression opacity regardless of length.

Platform governance research has correctly emphasized that moderation involves far more than individual content decisions, encompassing institutional rules, commercial incentives, human labor, automated systems, interface architecture, and political pressures (Gillespie, 2018; Roberts, 2019; Gorwa et al., 2020). The grammar of enforcement should be added to this architecture, because for most users the moderation notice is not merely an explanation of governance but the immediately accessible representation of it. Users ordinarily do not observe policy drafting, classifier design, reviewer instructions, threshold calibration, or internal deliberation. They encounter the final output, and the informational quality of that output determines how much of the invisible institutional chain becomes reconstructable. Moderation language is in this sense an accountability surface: the textual interface through which institutional power becomes linguistically available for attribution.

The concept connects grammar to governance without claiming that grammar alone determines institutional legitimacy. A transparent sentence cannot make an illegitimate policy legitimate, and an opaque sentence cannot prove that the underlying decision was illegitimate. What grammar determines is how much information about authority survives its conversion into communicative form. Three questions must therefore be kept apart. The substantive question asks whether the restriction was justified. The grammatical

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Line spacing: Multiple 1,15 li

question asks who is represented as having restricted. The procedural question asks whether the decision chain can be reconstructed and challenged. Censorship without a censor emerges when the first question becomes politically consequential while the second and third become progressively harder to answer.

~~**3.12 The grammar of enforcement as an accountability surface**~~

~~Platform governance research has correctly emphasized that moderation involves much more than individual content decisions. It encompasses institutional rules, commercial incentives, human labor, automated systems, interface architecture, and political pressures (Gillespie, 2018; Roberts, 2019; Gorwa et al., 2020). The grammar of enforcement should be added to this architecture because it constitutes the surface through which governed users encounter the decision-making system.~~

~~For most users, the moderation notice is not merely an explanation of governance.~~

~~It **is** the immediately accessible representation of governance.~~

~~Users ordinarily do not observe policy drafting meetings, classifier design, training procedures, reviewer instructions, threshold calibration, escalation processes, or internal deliberation. They encounter the final output.~~

~~The informational quality of that output therefore determines how much of the invisible institutional chain becomes reconstructable.~~

~~This makes moderation language an **accountability surface**: the textual interface through which institutional power becomes linguistically available for attribution.~~

~~The concept provides a direct connection between grammar and governance without claiming that grammar alone determines institutional legitimacy. A transparent sentence cannot make an illegitimate policy legitimate. An opaque sentence cannot prove that the underlying moderation decision was illegitimate. What grammar can determine is how much information about authority survives its conversion into communicative form.~~

~~The distinction is fundamental.~~

~~The substantive question asks:~~

~~**Was the restriction justified?**~~

~~The grammatical question asks:~~

~~**Who is represented as having restricted?**~~

~~The procedural question asks:~~

~~**Can the decision chain be reconstructed and challenged?**~~

~~Censorship without a censor emerges when the first question becomes politically consequential while the second and third become progressively harder to answer.~~

3.6 Formal prediction~~3.13 Formal prediction~~

The framework developed in this section produces a series of empirical predictions.

If suppression opacity is a recurrent property of platform enforcement discourse, moderation corpora should exhibit a measurable concentration of:

agentless passive enforcement clauses;

content-as-violator constructions;

user-as-violator constructions;

policy-agent substitutions;

nominalized enforcement processes;

technical-system agency without equivalent institutional agency;

procedural-inevitability constructions;

categorical risk predicates without explicit classification predicates;

and sanctions represented as states rather than institutional acts.

These features should not be interpreted independently. Their co-occurrence is more theoretically significant than the presence of any single structure.

A notice such as:

Your post was removed because it violated our policy.

contains at least three relevant transformations simultaneously.

First, *was removed* preserves the sanction while deleting the restricting agent.

Second, *your post* becomes subject of *violated*, transforming institutional classification into object violation.

Third, *policy* functions as the normative endpoint without representing the decision process through which the policy was interpreted and applied.

The sentence is grammatically ordinary. Institutionally, however, it compresses a potentially complex governance chain into an apparently direct relationship between content and rule.

That compression is the central object of measurement.

The compiled rule explains why it is possible.

Responsibility loss identifies what disappears.

Asymmetric visibility identifies who remains more visible.

Suppression opacity measures the resulting degradation in institutional traceability.

The grammar of platform enforcement therefore exposes a structural paradox of automated governance: **the more completely authority is translated into executable procedure, the less often authority may need to appear as the grammatical source of each individual execution** (Startari, 2025).

The censor does not disappear because no one governs.

The censor can disappear because governance has already been compiled into the rule.

4. Political Speech as Moderation Object

The grammar of platform enforcement becomes politically significant when the object being classified is not merely spam, fraud, pornography, or clearly identifiable malicious activity, but speech whose meaning depends upon contested histories, political identities, military conflict, sanctions regimes, ideological terminology, and relations of geopolitical power. Under these conditions, moderation cannot operate through formal detection alone. Before enforcement can occur, political expression must first be translated into categories that a governance system can recognize and act upon. The central operation is therefore classificatory: speech becomes a moderation object.

This transformation is analytically distinct from the subsequent removal or restriction of content. A statement does not enter a moderation system already constituted as *extremism*, *hate*, *incitement*, *misinformation*, *support*, *glorification*, *dangerous-organizations content*, or *risk*. It enters as linguistic material situated within a political and communicative context. Institutional policy determines which distinctions are relevant; operational rules specify which features count as evidence; automated or human systems classify the material; and enforcement mechanisms attach consequences to the resulting category. The transition from political utterance to moderation object is therefore an institutional production rather than a purely descriptive recognition.

The compiled-rule framework makes this transition formally visible. Institutional authority first establishes a normative distinction between permissible and impermissible expression. That distinction is subsequently translated into operational constraints capable of processing individual cases at scale (Startari, 2025). The moderation system thus requires a mapping:

political expression → institutional category → executable condition → enforcement consequence

The critical analytical question is what happens to the first term when it passes through the second. Political context must be reduced sufficiently for an institutional category to become executable. Yet some of the features discarded during that reduction may be precisely those necessary to determine what the speech means.

This is the central problem of political speech as moderation object.

A sentence may quote a prohibited organization without endorsing it.

A journalist may reproduce violent rhetoric in order to report it.

A researcher may name an extremist organization in an analytical context.

A civilian may document an attack using language shared with propaganda.

A protester may employ a politically contested slogan whose interpretation differs across communities.

A speaker may criticize Zionism without referring to Jewish people.

A speaker may use *Zionist* as a proxy for Jewish identity in a clearly antisemitic statement.

An Iranian user may criticize the Islamic Republic using culturally specific language that is difficult to interpret outside Persian political discourse.

A sanctions critic may reproduce claims originating from sanctioned political actors without endorsing those actors.

An anti-war statement may condemn one military actor, several military actors, or the political architecture sustaining a conflict.

None of these distinctions can be recovered from lexical form alone.

For moderation systems, however, contextual complexity creates an operational problem. Platforms govern enormous quantities of material across languages and jurisdictions. Rules must therefore become sufficiently standardized to support repeatable enforcement. Automated moderation intensifies this requirement because computational classification depends upon formalizable signals, training examples, thresholds, labels, and mappings between detected features and enforcement categories (Gorwa et al., 2020). Political language, by contrast, often derives its meaning from precisely those contextual relations that resist reduction to isolated lexical or syntactic features.

The result is a structural tension between **political context** and **moderation executability**.

The more context-sensitive a political expression is, the more information may be required to interpret it accurately.

The more scalable an enforcement system becomes, the greater the institutional incentive to convert contextual variation into standardized categories.

This tension does not establish that automated moderation is necessarily inaccurate. Nor does it imply that human moderation resolves contextual ambiguity. Human reviewers also work under rules, time constraints, institutional instructions, linguistic limitations, and classificatory systems. The relevant point is that political meaning must become institutionally legible before any moderation architecture, automated or human, can act upon it.

4.1 From political utterance to policy object

The first transformation occurs when an utterance is no longer processed primarily according to its political meaning but according to its relation to an institutional taxonomy.

Consider a hypothetical political post referring to an armed organization. At the level of political discourse, the post may function as reporting, condemnation, historical explanation, documentary evidence, satire, praise, advocacy, quotation, or incitement. At the level of platform governance, these differences must be translated into distinctions defined by platform policy.

The content therefore becomes a policy object.

Formally:

utterance U in context C

becomes

candidate instance of category K

Once U is mapped onto K, enforcement no longer needs to process the complete political context C. It can operate upon the category.

If category K triggers sanction S:

K → S

the original political utterance is transformed into an executable moderation relation.

This is one of the clearest applications of the compiled rule. The political meaning of the utterance is not itself executable. The institutional category is. Once the category has been assigned, authority can operate through the classification rather than through continuous reconsideration of the full political event (Startari, 2025).

The grammatical consequences discussed in Section 3 appear after this conversion. A moderation notice may state that content *violated* a policy because the institutional process that converted the content into a policy object has already occurred invisibly. The final sentence therefore reports the output of classification without reconstructing the classificatory event.

Political speech becomes especially vulnerable to this compression because political categories frequently remain contested outside the platform itself. *Terrorism*, *resistance*, *occupation*, *self-defense*, *extremism*, *incitement*, *propaganda*, *hate*, *Zionism*, *anti-Zionism*, *imperialism*, and *misinformation* do not operate as politically neutral lexical containers whose application is universally uncontested. Platforms nevertheless require governance categories sufficiently determinate to produce enforcement decisions.

The problem examined here is not whether such categories should exist. Some distinctions are indispensable to moderating direct threats, incitement to violence, targeted hatred, or operational support for violent actors. The analytical problem is that once a contested political utterance has been translated into a moderation category, the category can linguistically replace the interpretive process that generated it.

The sentence:

We classified this statement as praise of a designated dangerous organization

preserves that process.

The sentence:

This statement praises a dangerous organization

compresses classification into an asserted property.

The sentence:

This content violates our Dangerous Organizations policy

compresses it further.

The moderation object appears to explain its own enforcement.

4.2 Palestine, contextual classification, and the reporting distinction~~4.2 Palestine and the problem of contextual classification~~

Palestine-related digital speech provides a particularly well-documented environment for examining this transformation. Human Rights Watch documented more than 1,050 cases of takedowns and other forms of suppression involving Palestine-related content on Facebook and Instagram during October and November 2023. Its investigation identified multiple types of restrictions, including content removal, account restrictions, reduced visibility, limitations on engagement, and other enforcement measures (Human Rights Watch, 2023).

These findings are relevant here not as proof that every disputed moderation event constituted illegitimate censorship, but because they demonstrate the variety of mechanisms through which Palestine-related expression became an object of platform governance. The relevant content was not governed only through deletion. Visibility, account functionality, recommendation, and other dimensions of circulation could also become subject to restriction (Human Rights Watch, 2023).

The distinction matters for the concept of censorship without a censor because political visibility can be altered without producing the paradigmatic censorship event of a deleted statement. The moderation object can persist while the conditions governing its circulation change.

Palestine-related moderation also exposes the problem of category boundaries. Meta's treatment of the Arabic term *shaheed* became a direct test of how lexical material can acquire enforcement consequences through institutional classification. In its 2024 policy advisory opinion, Meta's Oversight Board concluded that the company's approach to uses of *shaheed* referring to individuals designated as dangerous substantially and disproportionately restricted expression. The Board reported that Meta had indicated the term was likely responsible for more removals under its Community Standards than any other single word or phrase, and recommended replacing the blanket approach with more context-sensitive treatment (Oversight Board, 2024a).

The significance of this case is formal as well as political. The word did not possess an invariant moderation value independently of context. Its enforceability derived from the relation between lexical identification, the platform's Dangerous Organizations and Individuals framework, and the institutional rules governing how references to designated actors were interpreted.

The sequence can be reconstructed as:

lexical form

→ reference to designated actor

→ institutional interpretation of reference

→ policy classification

→ enforcement consequence

When this sequence becomes executable at scale, the middle stages can disappear from the final representation. A user need not be told that Meta adopted a particular interpretation of the relationship between the word and the policy category. The notice can simply state that the content violated policy.

The rule then appears to discover what institutional design has first made detectable.

This is the classificatory counterpart of responsibility loss. The enforcement consequence remains visible, but the interpretive action necessary to produce the classification can disappear (Startari, 2026a).

The Oversight Board's subsequent analysis of uses of the phrase *from the river to the sea* illustrates the same contextual principle from a different direction. In three cases reviewed in 2024, the Board concluded that the specific posts did not violate Meta's Hate Speech, Violence and Incitement, or Dangerous Organizations and Individuals rules. The Board emphasized contextual differences and found that the posts contained signs of solidarity with Palestinians without calls for violence or exclusion in the cases before it (Oversight Board, 2024b).

The importance of this finding for the present framework does not depend upon establishing a universal meaning for the phrase. The opposite follows. A politically contested expression cannot necessarily be converted into a stable moderation category solely from its lexical presence. The category requires interpretation.

This distinction can be formalized as:

expression ≠ classification

Rather:

expression + context + institutional rule + interpretive procedure → classification

The grammatical problem emerges when the final moderation language collapses this relation into:

expression = violation

That compression removes the interpretive institution from the representation.

Political conflict creates another classification problem: the distinction between representing prohibited material and endorsing it. News reporting may reproduce statements made by armed organizations; human rights documentation may contain imagery of violence; researchers may reproduce extremist language as evidence; civilians documenting warfare may upload graphic material whose visual properties resemble

Formatted: Font: (Default) Times New Roman, 12 pt

propaganda; political critics may quote adversaries precisely in order to condemn them. The same video may function as propaganda in one context and evidence in another, the same quotation as endorsement, condemnation, documentation, or analysis. Moderation systems recognize this institutionally through exceptions, allowances, and contextual policies, yet the existence of such exceptions itself establishes that the surface form of content cannot fully determine its moderation status. The Oversight Board case concerning a Washington Post report on Israel and Palestine illustrates the point: the Board characterized it as an instance in which Meta had over-enforced its Dangerous Organizations and Individuals policy against news reporting (Oversight Board, 2024c).

The relevant classification is therefore not designated actor appears, prohibited content, but something closer to designated actor appears plus communicative relation R plus contextual conditions C, yielding moderation status, where R can include reporting, neutral reference, condemnation, praise, or support. When executable governance compresses these relations too aggressively, political documentation becomes indistinguishable from the political conduct it describes. This may be defined as referential collapse: a moderation system fails to preserve the distinction between reference to a prohibited political object and participation in, endorsement of, or support for that object. The phenomenon matters beyond Palestine, because any conflict in which armed actors, sanctioned organizations, extremist movements, or legally restricted entities play a significant political role creates the same structural problem. A governance architecture that treats reference itself as a sufficiently strong proxy for prohibited relation risks converting political knowledge into moderation risk.

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Line spacing: Multiple 1,15 li

~~**4.3 Reporting, documentation, and endorsement**~~

~~Political conflict creates another classification problem: the distinction between representing prohibited material and endorsing it.~~

~~News reporting may reproduce statements made by armed organizations. Human rights documentation may contain imagery of violence. Researchers may reproduce extremist language as evidence. Civilians documenting warfare may upload graphic material whose visual properties resemble material circulated for propaganda purposes. Political critics may quote adversaries precisely in order to condemn them.~~

~~These cases demonstrate why content identity and communicative function cannot be treated as equivalent.~~

~~The same video may function as propaganda in one context and evidence in another.~~

The same quotation may function as endorsement, condemnation, documentation, or analysis.

The same political symbol may indicate affiliation in one context and historical discussion in another.

The distinction is institutionally recognized in moderation systems through exceptions, allowances, and contextual policies. Yet the existence of such exceptions also establishes that the surface form of content cannot fully determine its moderation status.

An Oversight Board case concerning a Washington Post report on Israel and Palestine illustrates this point. The Board characterized the case as an instance in which Meta had over-enforced its Dangerous Organizations and Individuals policy against news reporting, emphasizing the recurring difficulty of preserving reporting allowances when content concerns entities designated as dangerous (Oversight Board, 2024c).

This is theoretically important because it demonstrates that the relevant classification is not simply:

designated actor appears → prohibited content

but something closer to:

designated actor appears + communicative relation R + contextual conditions C → moderation status

The variable R can include reporting, neutral reference, condemnation, praise, support, or other context-dependent relations.

When executable governance compresses these relations too aggressively, political documentation can become indistinguishable from the political conduct it describes.

This problem can be defined as **referential collapse**.

Referential collapse occurs when a moderation system fails to preserve the distinction between reference to a prohibited political object and participation in, endorsement of, or support for that object.

~~The phenomenon matters beyond Palestine. Any conflict in which armed actors, sanctioned organizations, extremist movements, or legally restricted entities play a significant political role creates the same structural problem. Public discourse about such actors requires naming them. Reporting requires describing them. Historical analysis requires contextualizing them. Political opposition may require condemning or debating them.~~

~~A governance architecture that treats reference itself as a sufficiently strong proxy for prohibited relation risks converting political knowledge into moderation risk.~~

4.3 Iranian political speech, dual censorship environments, and anti-sanctions discourse~~4.4 Iranian political speech and dual censorship environments~~

Iranian political speech presents a different but complementary test case because users can operate within overlapping structures of state censorship and private platform governance.

This distinction must remain explicit. Restrictions imposed by the Iranian state and moderation imposed by private platforms are institutionally different forms of governance. Conflating them would make the analytical category of censorship without a censor less precise rather than more precise.

The relevance of Iran lies instead in the position of political speakers who can encounter both structures simultaneously.

Digital-rights organizations have documented longstanding concerns about Persian-language moderation on Instagram. In 2022, Access Now, ARTICLE 19, and the Center for Human Rights in Iran ~~and ARTICLE 19~~ argued that Meta needed to improve the transparency, accountability, and contextual competence of Persian-language moderation, citing complaints from Iranian users and concerns about the moderation infrastructure serving the Persian-speaking community (Access Now, 2022; ARTICLE 19, 2022).

These sources should be interpreted according to their evidentiary status. They document civil-society complaints and reported moderation problems; they do not independently establish that every disputed restriction resulted from systematic political bias. That distinction is necessary for the empirical design of the present paper.

What they do establish is the existence of a moderation environment in which Persian linguistic context, political contestation, trust, reviewer oversight, and enforcement transparency have generated sufficient concern to become objects of sustained rights advocacy (Access Now, 2022; ARTICLE 19, 2022).

For the present framework, Iran introduces two analytically separate questions.

The first concerns **language-specific classification**. Does a platform possess sufficient contextual competence to distinguish among Persian political expressions with different functions?

The second concerns **institutional traceability**. When Iranian political content is restricted, does the explanation identify the platform's action and the rule applied, or does it reduce the event to an impersonal violation?

These variables should not be conflated. An accurate classification can be communicated opaquely. An inaccurate classification can be communicated transparently. Classification quality and suppression opacity are distinct dimensions.

This distinction produces a useful matrix:

accurate classification + high transparency

accurate classification + high opacity

inaccurate classification + high transparency

inaccurate classification + high opacity

The fourth condition produces the greatest accountability problem because the speaker may confront both a questionable classification and a weakly traceable enforcement explanation.

Iranian political speech is therefore important not because it can be assumed to receive a particular treatment, but because it permits the framework to test whether suppression opacity persists across a linguistic and geopolitical environment different from the Palestine case.

Anti-sanctions discourse introduces a further problem because it challenges coercive policies that are themselves embedded in legal, security, and geopolitical classifications. In the previous paper of this series, sanctioned suffering was used to analyze how the effects of sanctions can remain visible while coercive agency becomes abstracted into terms such as pressure, restrictions, or economic deterioration (Startari, 2026c). The present question is what occurs when users contest those policies within environments governed by platform rules concerning dangerous organizations, state-linked actors, misinformation, coordinated influence, or sanctioned entities. The claim sanctions are harming civilians is not structurally identical to support the sanctioned organization, and this sanction regime should end is not equivalent to the claims of every sanctioned actor are true. Yet polarized environments can create pressure toward categorical compression, in which positions are grouped according to presumed affiliation rather than proposition-level content.

This may be called alignment inference: a political statement concerning actor A is classified partly through an inferred relationship between the speaker and actor A rather than through the explicit propositional content of the statement. Speaker criticizes sanctions against A can be incorrectly compressed into speaker supports A, which can then be mapped onto support for A belongs to moderation category K, producing the transformation policy criticism, inferred alignment, risk classification, enforcement. No claim is made that contemporary platforms systematically perform this transformation; that proposition requires corpus evidence, and the framework identifies it as a coding possibility requiring empirical testing. The distinction matters because anti-sanctions discourse is particularly valuable for detecting whether political moderation systems preserve the difference between policy position and actor endorsement. If they do, the classification architecture remains comparatively granular; if they do not, geopolitical alignment itself may become an implicit moderation variable.

~~**4.5 Anti-sanctions discourse**~~

~~Anti-sanctions discourse introduces a further problem because it challenges coercive policies that are themselves embedded in legal, security, and geopolitical classifications.~~

~~In the previous paper of this series, sanctioned suffering was used to analyze how the effects of sanctions can remain visible while coercive agency becomes abstracted into terms such as *pressure*, *restrictions*, *economic deterioration*, or *sanctions* themselves (Startari, 2026c). The present paper asks what occurs when users contest those policies within environments governed by platform rules concerning dangerous organizations, state-linked actors, misinformation, coordinated influence, or sanctioned entities.~~

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Line spacing: Multiple 1,15 li

The political claim:

sanctions are harming civilians

is not structurally identical to:

support the sanctioned organization

Likewise:

this sanction regime should end

is not structurally equivalent to:

the claims of every sanctioned actor are true

Yet politically polarized environments can create pressure toward categorical compression. Positions are grouped according to presumed affiliation rather than proposition-level content.

This can be called **alignment inference**.

Alignment inference occurs when a political statement concerning actor A is classified partly through an inferred relationship between the speaker and actor A rather than through the explicit propositional content of the statement.

For example:

speaker criticizes sanctions against A

can be incorrectly compressed into:

speaker supports A

which can then be mapped onto:

support for A belongs to moderation category K

The complete transformation becomes:

~~**policy criticism → inferred alignment → risk classification → enforcement**~~

~~No claim is made here that contemporary platforms systematically perform this transformation. That proposition requires corpus evidence. The framework instead identifies it as a coding possibility requiring empirical testing.~~

~~The distinction is necessary because anti-sanctions discourse is particularly valuable for detecting whether political moderation systems preserve the difference between **policy position** and **actor endorsement**.~~

~~If they do, the classification architecture remains comparatively granular.~~

~~If they do not, geopolitical alignment itself may become an implicit moderation variable.~~

4.4 Anti-war and anti-imperial speech: symmetry and the classification boundary~~4.6 Anti-war speech and the problem of symmetry~~

Anti-war discourse introduces another classification problem because opposition to violence does not necessarily distribute condemnation symmetrically.

A speaker may oppose a war while assigning greater responsibility to one actor.

A speaker may condemn military conduct by one state while defending another actor's right of resistance.

A speaker may oppose all armed parties while sharply criticizing the foreign policy of an external power.

A speaker may support national self-defense while opposing a particular military operation.

The expression *anti-war* therefore does not define a single ideological position.

Moderation systems should consequently not be expected to identify anti-war discourse through a single stable lexical category. The relevant risk lies in converting asymmetrical political criticism into evidence of prohibited allegiance.

This distinction is important within the architecture of asymmetric visibility developed earlier in the series. Political discourse frequently distributes agency and responsibility

unevenly because the speaker believes responsibility itself is unequal (Startari, 2026b). Linguistic asymmetry is therefore not automatically evidence of bias. Sometimes it accurately represents an asymmetric underlying relation.

The same principle must apply to moderation.

A user criticizing one side more intensely than another cannot be classified as dangerous merely because the distribution of criticism is unequal.

The relevant moderation question is whether the speech satisfies an independently defined policy condition.

Formally:

political asymmetry ≠ policy violation

unless additional conditions establish the relevant category.

This becomes important when platforms moderate speech during armed conflict, because political discourse often contains extreme moral evaluations, allegations of crimes, graphic evidence, historical comparisons, denunciations of military institutions, and emotionally intense language. A system optimized for ordinary interpersonal discourse may treat some of these forms differently when they occur in conflict reporting or political argument.

The empirical task is therefore not to assume that anti-war speech is suppressed, but to test whether particular anti-war formulations are more likely to enter risk categories and, if restricted, whether their moderation notices exhibit high CDR or SOI.

Anti-imperial discourse provides the broadest comparative category in this study, including political claims that interpret military, economic, diplomatic, technological, or informational relations through structures of unequal power between states, blocs, institutions, or populations. This category requires particular methodological discipline because anti-imperial cannot be treated as synonymous with truthful, emancipatory, peaceful, or legitimate discourse. Anti-imperial rhetoric can coexist with misinformation or ethnic hatred, can defend authoritarian governments, and can reproduce propaganda. It can also constitute historically grounded criticism of military intervention, occupation, sanctions, dependency, foreign influence, or unequal international institutions. The moderation category therefore cannot be derived from the political orientation itself,

Formatted: Font: (Default) Times New Roman, 12 pt

which yields the principle central to the design of this paper: ideological position is not moderation status.

The relevant question is whether moderation systems preserve that distinction. A statement criticizing United States foreign policy, Israeli state policy, European sanctions, Iranian regional policy, Russian intervention, or any other power relation should be evaluated according to the rule actually implicated by its content, not according to its location within a geopolitical ideological field. The analytical risk emerges when a political position becomes a proxy variable for a moderation category, following the sequence ideological position P, presumed affiliation A, risk classification K, restrictive outcome S. This is structurally similar to alignment inference but broader, because the inferred association concerns an ideological field rather than a specific actor. The model does not assert that this sequence occurs systematically; it defines a falsifiable pattern.

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Line spacing: Multiple 1,15 li

~~**4.7 Anti-imperial speech as a classification boundary**~~

~~Anti-imperial discourse provides the broadest comparative category in this study. It includes political claims that interpret military, economic, diplomatic, technological, or informational relations through structures of unequal power between states, blocs, institutions, or populations.~~

~~This category requires particular methodological discipline because *anti-imperial* cannot be treated as synonymous with truthful, emancipatory, peaceful, or legitimate discourse.~~

~~Anti-imperial rhetoric can coexist with misinformation.~~

~~It can coexist with ethnic hatred.~~

~~It can defend authoritarian governments.~~

~~It can reproduce propaganda.~~

~~It can also constitute historically grounded criticism of military intervention, occupation, sanctions, dependency, foreign influence, or unequal international institutions.~~

~~The moderation category therefore cannot be derived from the political orientation itself.~~

~~This principle is central to the design of the paper:~~

~~**ideological position is not moderation status.**~~

~~The relevant question is whether moderation systems preserve that distinction.~~

~~An anti-imperial statement criticizing United States foreign policy, Israeli state policy, European sanctions, Iranian regional policy, Russian intervention, or any other power relation should be evaluated according to the rule actually implicated by its content, not according to its location within a geopolitical ideological field.~~

~~The analytical risk emerges when a political position becomes a proxy variable for a moderation category.~~

~~The sequence would be:~~

~~**ideological position P**~~

~~**→ presumed affiliation A**~~

~~**→ risk classification K**~~

~~**→ restrictive outcome S**~~

~~This is structurally similar to alignment inference but broader because the inferred association concerns an ideological field rather than a specific actor.~~

~~Again, the model does not assert that this sequence occurs systematically. It defines a falsifiable pattern.~~

4.5 The speaker becomes the risk: lost context and classification without a classifier~~4.8 The speaker becomes the risk~~

Across these cases, a recurrent transformation becomes theoretically possible: political context is progressively reduced until the governed subject becomes legible primarily through risk.

This transformation can occur at several levels.

The **utterance** becomes risky content.

The **content** becomes a policy violation.

The **account** becomes a repeat violator.

The **speaker** becomes associated with prohibited categories.

The **network** becomes coordinated or suspicious.

Each transition expands the object of governance.

Initially, the moderation system evaluates a proposition.

Subsequently, enforcement can extend to the content item, account, distribution network, or behavioral pattern.

This progression is not inherently illegitimate. Platforms require account-level and network-level enforcement to address spam, fraud, coordinated manipulation, repeated harassment, and other harms that cannot be governed through isolated content analysis.

Its political significance emerges when the expansion occurs in contested speech environments.

At that point, asymmetric visibility can acquire an institutional form:

speaker = visible as risk

platform = less visible as classifier of risk

(Startari, 2026b).

This is the double displacement identified in the introductory framework. The speaker becomes increasingly legible through the categories of governance while the institution producing those categories becomes less grammatically explicit in the explanation of enforcement.

The extreme linguistic form is:

Your account was restricted because of repeated policy violations.

The sentence contains the governed entity, the sanction, and the violation history.

It does not necessarily contain the actor that imposed the restriction, the classifications that generated the violation history, the mechanisms that produced those classifications, or the institutional path through which they can be reviewed.

Suppression opacity therefore increases not because no information is provided, but because information is distributed asymmetrically.

The broader theoretical problem can now be stated formally. Political speech contains at least two analytically distinct informational layers: propositional content P and political and communicative context C. Moderation produces a classification K = f(P, C, R), where R represents the institutional rule. In an ideal context-sensitive system K depends upon all three. Under conditions of contextual reduction, however, K is approximately f(P, R), or in more extreme cases f(L, R), where L represents lexical or other surface-level signals. As C approaches zero, the probability of collapsing functionally distinct political utterances into the same moderation category increases whenever those utterances share similar surface features. This does not mean context must always override content, since direct threats or unambiguous targeted hatred may require little contextual reconstruction. The variable becomes most important when the same linguistic material supports several plausible communicative functions.

Formatted: Font: (Default) Times New Roman, 12 pt

The documented cases demonstrate this at three levels. The shaheed assessment concluded that Meta's blanket approach failed to preserve sufficient contextual differentiation, which is the lexical level (Oversight Board, 2024a). The from the river to the sea cases show it at phrase level, where the Board evaluated the particular content rather than treating the phrase as automatically constituting hate speech, incitement, or prohibited support (Oversight Board, 2024b). The Washington Post case shows it at discourse-function level, where reporting about a designated actor was initially caught by enforcement the Board later characterized as over-enforcement (Oversight Board, 2024c). Together they establish that classification requires mediation, and once that proposition is accepted, moderation notices that erase the classificatory actor become analytically consequential.

Formatted: Font: (Default) Times New Roman, 12 pt

The grammatical analogue of censorship without a censor is therefore classification without a classifier. A political statement becomes extremist; a post becomes harmful; a phrase becomes hateful; an account becomes dangerous; a claim becomes misinformation. Each predicate presupposes a classificatory operation whose underlying form is institution I classifies object O as category K, and whose compressed form is simply O is K. The classificatory predicate remains and the classifier disappears. This extends responsibility loss one stage earlier: before a restrictive consequence can be applied, a judgment must be produced, and that judgment can itself lose its agent

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Line spacing: Multiple 1,15 li

(Startari, 2026a). The complete architecture runs from classifier deletion to censor deletion, ending in O violates policy K, where both the institutional classifier and the institutional censor have disappeared and what remains is an object whose properties appear to explain the sanction. This is the strongest form of procedural naturalization identified in the paper.

~~**4.9 Political context as lost variable**~~

~~The broader theoretical problem can now be stated formally.~~

~~Political speech contains at least two analytically distinct informational layers:~~

~~**P = propositional content**~~

~~**C = political and communicative context**~~

~~Moderation produces classification:~~

~~**K = f(P, C, R)**~~

~~where R represents the institutional rule.~~

~~In an ideal context-sensitive system, K depends upon all three.~~

~~Under conditions of contextual reduction, however:~~

~~**K ≈ f(P, R)**~~

~~or, in more extreme cases:~~

~~**K ≈ f(L, R)**~~

~~where L represents lexical or other surface-level signals.~~

~~As C approaches zero, the probability of collapsing functionally distinct political utterances into the same moderation category increases whenever those utterances share similar surface features.~~

~~This does not mean context must always override content. Direct threats, explicit calls for violence, or unambiguous targeted hatred may require little contextual reconstruction to~~

classify. The variable becomes most important when the same linguistic material supports several plausible communicative functions.

The *shaheed* case demonstrates this problem at lexical level. The Oversight Board's assessment concluded that Meta's blanket approach failed to preserve sufficient contextual differentiation (Oversight Board, 2024a).

The *from the river to the sea* cases demonstrate it at phrase level. The Board evaluated the particular content rather than treating the phrase itself as automatically constituting hate speech, incitement, or prohibited support in the three cases under review (Oversight Board, 2024b).

The Washington Post case demonstrates it at discourse function level, where reporting about a designated actor was initially caught by enforcement that the Board later characterized as over-enforcement (Oversight Board, 2024c).

Together, these cases establish a critical distinction:

classification requires mediation.

Once that proposition is accepted, moderation notices that erase the classificatory actor become analytically consequential.

4.10 Classification without a classifier

The grammatical analogue of censorship without a censor is therefore **classification without a classifier**.

A political statement becomes extremist.

A post becomes harmful.

A phrase becomes hateful.

An account becomes dangerous.

A claim becomes misinformation.

Yet each predicate presupposes a classificatory operation.

The underlying form is:

institution I classifies object O as category K

The compressed form is:

O is K

The classificatory predicate remains.

The classifier disappears.

This mechanism extends the earlier concept of responsibility loss. In the first paper of the series, the central problem concerned consequences whose responsible agents became grammatically optional (Startari, 2026a). Here, the problem occurs one stage earlier. Before a restrictive consequence can be applied, a judgment must be produced. That judgment itself can lose its agent.

The complete architecture therefore becomes:

classifier deletion → censor deletion

First:

we classify O as K

becomes:

O is K

Then:

we restrict O because we classified it as K

becomes:

O was restricted because O is K

~~Finally:~~

~~**O violates policy K**~~

~~The institutional classifier and institutional censor have both disappeared.~~

~~What remains is an object whose properties appear to explain the sanction.~~

~~This is the strongest form of procedural naturalization identified in the paper.~~

4.6 Appeals, corpus logic, and the executable object~~4.11 The role of appeals~~

Appeals provide an important test of this architecture because an appeal can force the institution to reverse the compression.

A generic enforcement notice may state:

Your content violated policy.

An effective appeal process requires a more elaborate institutional operation:

someone or some system must reconsider whether the classification was correct.

The existence of appeal therefore demonstrates that the original violation status was not an intrinsic property of the content. It was a revisable institutional determination.

This is theoretically significant.

If a classification can be overturned, then:

content ≠ violation intrinsically

Rather:

content → institutional determination of violation

The Oversight Board's case record repeatedly includes instances in which Meta reversed original moderation decisions after cases were brought to the company's attention, illustrating that moderation outputs remain revisable institutional judgments rather than immutable properties of the underlying content (Oversight Board, 2024d).

Appeal-path visibility should consequently be treated as more than a procedural convenience. It is evidence that the governance system acknowledges the contingency of its own classifications.

This explains its inclusion in SOI.

A notice with a visible and meaningful appeal path preserves the possibility that classification can be institutionally reconsidered.

A notice without such a pathway presents the output as more final and less traceable.

Human Rights Watch reported cases in its 2023 Palestine-related dataset in which affected users lacked access to an effective appeals mechanism, making the relation between enforcement opacity and procedural accountability particularly relevant to the corpus proposed here (Human Rights Watch, 2023).

The five political environments selected for this study should not be treated as a single ideological category; they serve different analytical functions. Palestinian digital speech tests classification under armed conflict, contested terminology, dangerous-organizations policy, multilingual moderation, and documented disputes concerning over-enforcement. Iranian political speech tests Persian-language moderation and the interaction between contested expression, local linguistic context, and private platform governance. Anti-sanctions discourse tests whether criticism of coercive policy can be distinguished from endorsement of sanctioned actors. Anti-war discourse tests whether asymmetric political criticism remains distinguishable from incitement or prohibited support. Anti-imperial discourse tests whether ideological orientation itself becomes an implicit proxy for risk classification.

The corpus should not presuppose identical enforcement outcomes across these categories, because difference is analytically valuable. If Palestine-related notices display substantially higher CDR or SOI than comparable anti-war content, that difference requires explanation. If Iranian political content exhibits higher classification opacity but not higher censor deletion, the distinction matters. If anti-sanctions content shows no unusual moderation pattern, that result falsifies stronger versions of the hypothesis. If anti-imperial speech is treated no differently from comparable mainstream geopolitical criticism after controlling for explicit policy violations, the corpus should record that null result. The framework therefore does not require suppression to be found; it requires suppression, when present, to be measurable without assuming its existence in advance.

~~**4.12 Comparative logic of the corpus**~~

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Line spacing: Multiple 1,15 li

The five political environments selected for this study should therefore not be treated as a single ideological category.

They serve different analytical functions.

Palestinian digital speech tests classification under armed conflict, contested terminology, dangerous-organizations policy, multilingual moderation, and documented disputes concerning over-enforcement.

Iranian political speech tests Persian-language moderation and the interaction between politically contested expression, local linguistic context, and private platform governance.

Anti-sanctions discourse tests whether criticism of coercive policy can be distinguished from endorsement of sanctioned actors.

Anti-war discourse tests whether asymmetric political criticism can remain distinguishable from incitement, prohibited support, or other enforceable categories.

Anti-imperial discourse tests whether ideological orientation itself becomes an implicit proxy for risk classification.

The corpus should not presuppose identical enforcement outcomes across these categories. Difference is analytically valuable.

If Palestine-related notices display substantially higher CDR or SOI than comparable anti-war content, that difference requires explanation.

If Iranian political content exhibits higher classification opacity but not higher censor deletion, the distinction matters.

If anti-sanctions content shows no unusual moderation pattern, that result falsifies stronger versions of the hypothesis.

If anti-imperial speech is treated no differently from comparable mainstream geopolitical criticism after controlling for explicit policy violations, the corpus should record that null result.

The framework therefore does not require suppression to be found.

It requires suppression, when present, to be measurable without assuming its existence in advance.

4.13 From moderation object to executable object

The final theoretical transition concerns what happens once political expression has been successfully classified.

Before classification, the platform confronts language.

After classification, the platform confronts an executable object.

The distinction is fundamental.

Language possesses ambiguity.

The category reduces ambiguity.

The rule attaches consequences to the category.

Execution applies those consequences.

The transformation can therefore be represented as:

political speech

→ interpreted speech

→ classified speech

→ moderation object

→ executable object

→ restricted object

At the final stages, the original political context may no longer be necessary for the system to act. The rule requires only the classification.

This explains why political speech can move from contested interpretation to procedural inevitability with remarkable grammatical economy.

Once the rule has been compiled, $K \rightarrow S$ is sufficient.

The institutional history through which K and S became connected can disappear from each individual enforcement event (Startari, 2025).

The moderation notice can then present the terminal relation:

Content violating K is subject to S.

At that point, the political dispute has been transformed into a technical condition.

This does not mean that the dispute has actually been resolved.

It means that the governance system has acquired a representation in which it can be executed as though it had been resolved.

That distinction is central to censorship without a censor.

The institutional power of moderation lies partly in the capacity to transform contested political meaning into executable categories. Suppression opacity begins when the classification process, the actor responsible for it, and the discretionary decisions embedded within it become progressively less visible than the object being classified.

Political speech then undergoes a double transformation.

First, **speech becomes risk**.

Second, **institutional judgment becomes procedure**.

The user remains visible as speaker, account holder, violator, or risk object.

The content remains visible as prohibited, dangerous, hateful, misleading, or noncompliant.

The rule remains visible as policy.

~~The consequence remains visible as removal, restriction, reduced distribution, or ineligibility.~~

~~What can disappear is the institutional act that connected them.~~

Political speech then undergoes a double transformation: speech becomes risk, and institutional judgment becomes procedure. The user remains visible as speaker, account holder, violator, or risk object. The content remains visible as prohibited, dangerous, hateful, misleading, or noncompliant. The rule remains visible as policy. The consequence remains visible as removal, restriction, reduced distribution, or ineligibility. What can disappear is the institutional act that connected them.

The final theoretical transition concerns what happens once political expression has been successfully classified. Before classification, the platform confronts language, which possesses ambiguity. After classification, it confronts an executable object: the category reduces ambiguity, the rule attaches consequences to the category, and execution applies those consequences. The full progression runs from political speech through interpreted speech, classified speech, moderation object, and executable object to restricted object. At the final stages the original political context is no longer necessary for the system to act, because the rule requires only the classification. Once the rule has been compiled, K to S is sufficient, and the institutional history through which K and S became connected can disappear from each individual enforcement event (Startari, 2025). The notice can then present only the terminal relation: content violating K is subject to S. The political dispute has not been resolved; the governance system has acquired a representation in which it can be executed as though it had been.

This is the point at which the preceding theoretical sequence converges. Responsibility loss explains the disappearance of accountable agency (Startari, 2026a). Asymmetric visibility explains why the governed actor may remain more grammatically legible than the governing institution (Startari, 2026b). Sanctioned suffering establishes how imposed consequences can be transformed into apparently impersonal conditions (Startari, 2026c). The compiled rule explains how institutional judgments can be converted into executable relations that continue operating without continuous restatement of their author (Startari, 2025).

Political speech as moderation object supplies the missing intermediate stage.

Before the censor can disappear from the grammar of suppression, **political meaning must first be converted into a category that the rule knows how to suppress**.

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Line spacing: Multiple 1,15 li

5. Measuring Censor Deletion

The preceding sections established censorship without a censor as a problem of institutional traceability rather than as a synonym for content moderation itself. The empirical task is therefore to determine whether restrictive platform actions are systematically represented through linguistic structures that preserve the sanction, violation, or risk classification while reducing the visibility of the institution responsible for producing them. This requires converting the theoretical distinction between visible suppression and obscured agency into reproducible variables.

The methodological design operates at two related levels. The first is strictly grammatical. It asks whether the institutional actor responsible for a moderation action remains explicitly represented in the clause describing that action. The second is procedural and informational. It asks how much of the broader enforcement chain can be reconstructed from the moderation communication available to the user.

These levels correspond to the two principal measures proposed in this article:

Censor-Deletion Rate (CDR) measures grammatical agent deletion.

Suppression Opacity Index (SOI) measures institutional traceability loss.

The distinction is necessary because grammatical opacity and procedural opacity do not always coincide. An agentless passive can occur within an otherwise detailed and accountable moderation explanation. Conversely, an active sentence identifying the platform can provide almost no information concerning the rule, classification process, mechanism of enforcement, or possibility of appeal. CDR and SOI therefore measure different properties of the same governance event.

The methodological objective is not to determine through linguistic form alone whether an act of moderation constitutes unlawful, illegitimate, politically motivated, or normatively unjustified censorship. Such conclusions would require additional evidence concerning legal standards, policy consistency, comparative outcomes, institutional intent, political

context, and the substantive content being moderated. The present methodology measures a narrower phenomenon: **how visibly institutional responsibility survives in the representation of restriction**.

This distinction follows the methodological logic developed throughout the series. Responsibility loss was formulated as a measurable degradation in the relation between consequence and responsible agency rather than as an inference about hidden intention (Startari, 2026a). Asymmetric visibility similarly requires comparison of observable distributions of agency rather than assumptions about ideological motive (Startari, 2026b). The present study applies the same restriction to platform governance. The object is observable linguistic structure.

5.1 Unit of analysis

The primary unit of analysis is the **moderation clause**.

A moderation clause is defined as a finite or functionally equivalent linguistic unit that communicates at least one of the following:

a restrictive action;

a policy classification;

a violation determination;

a risk attribution;

an enforcement consequence;

a visibility modification;

an account-level restriction;

or an appeal-related decision.

The clause is preferred over the complete moderation notice because a single notification can contain several analytically distinct propositions.

For example:

We removed your post because it violated our policy. Your account may also have reduced recommendation eligibility.

This notice contains at least two enforcement-related clauses:

We removed your post

and

Your account may also have reduced recommendation eligibility.

The first contains an explicit institutional agent.

The second describes an enforcement-related condition without one.

Treating the complete notice as a single observation would obscure that difference.

Clause-level analysis also makes it possible to compare grammatical agency across platforms, policy categories, languages, and types of sanctions while reducing the influence of interface length or notice design.

The methodology therefore separates the following analytical objects:

N = moderation notice

C = moderation clause

where each notice N may contain one or more clauses:

$N = \{C_1, C_2, \ldots, C_n\}$

The principal grammatical metrics are calculated over C rather than N.

Notice-level variables are retained for SOI because appeal pathways, policy links, review information, and procedural explanations may be distributed across several clauses or interface elements.

5.2 Corpus construction

The corpus should contain moderation-related communications associated with the five political environments defined in Section 4:

Palestinian digital speech;

Iranian political speech;

anti-sanctions discourse;

anti-war discourse;

anti-imperial discourse.

These categories are comparative domains rather than presupposed classes of censorship.

The corpus should include, where verifiable and available:

platform-generated removal notices;

account restriction notices;

visibility-reduction notices;

recommendation restrictions;

monetization restrictions;

policy-violation explanations;

appeal decisions;

automated-review notices;

human-review notices;

policy enforcement explanations;

and platform explanations associated with documented moderation cases.

Publicly documented cases from oversight institutions and digital-rights investigations can supplement direct moderation notices when the original wording is preserved or

reproducibly documented. Cases reconstructed only through secondary paraphrase should be coded separately because paraphrase can alter grammatical agency.

This distinction is methodological rather than editorial. If the study measures grammatical representation, the grammar must originate from the platform communication under analysis. A human-rights organization stating that *Meta removed a post* cannot substitute for a platform notice stating that *the post was removed*. The two formulations assign agency differently.

Each corpus item should therefore include a provenance variable:

P_0 = original platform text

P_1 = authenticated reproduction or screenshot

P_2 = verbatim reproduction in a documented external source

P_3 = secondary paraphrase

Only P_0–P_2 should enter the principal grammatical analysis. P_3 may be retained for contextual comparison but should not contribute to CDR.

This preserves the distinction between the grammar of platform governance and the grammar used by external observers to describe platform governance.

5.3 Corpus metadata

Each moderation event should be coded with sufficient metadata to permit controlled comparison.

At minimum, the dataset should record:

platform;

date;

language;

political category;

content type;

policy category;

restriction type;

account or content level;

automated, human, hybrid, or unspecified review;

appeal availability;

appeal outcome where known;

source provenance;

and whether the original moderated content is available.

Where politically sensitive content is included, the coding architecture should distinguish between the political category of the discourse and the moderation category applied to it.

For example:

Political category: Palestine-related speech.

Moderation category: Dangerous Organizations and Individuals.

These variables cannot be collapsed. Doing so would reproduce precisely the classificatory compression examined in Section 4.

The same principle applies to anti-sanctions discourse, anti-war discourse, and anti-imperial discourse. These categories describe the political function or orientation of the material; they do not determine whether the content violated platform policy.

5.4 Institutional agent

The central variable for CDR is the presence or absence of an **explicit institutional agent**.

An institutional agent is defined as a platform, company, moderation team, reviewer, authorized decision body, or clearly identified technical system acting on behalf of the platform in a clause describing restriction.

Examples of explicit institutional agency include:

We removed this content.

Meta restricted the account.

Our review team determined that the post violated the policy.

Our automated system detected the content and restricted distribution.

Each clause identifies a grammatical actor connected to the moderation process.

By contrast:

This content was removed.

Distribution was reduced.

The account was restricted.

This post violates policy.

do not identify an explicit institutional agent responsible for enforcement.

A possessive marker alone does not necessarily constitute explicit agency.

For example:

Your post violated our policy.

The possessive *our* establishes institutional ownership of the policy but does not identify the platform as the actor that classified or restricted the content. The clause therefore receives:

institutional presence = yes

but

institutional agency = no

This distinction is required to avoid overstating transparency.

5.5 The Censor-Deletion Rate

The **Censor-Deletion Rate (CDR)** measures the proportion of suppression-related clauses in which an enforcement action occurs without an explicit institutional agent.

Let:

S = total number of suppression clauses

and:

A_0 = number of suppression clauses without explicit institutional agent

Then:

$CDR = A_0 / S$

For percentage reporting:

$CDR\% = (A_0 / S) \times 100$

A corpus containing 800 suppression clauses, of which 600 contain no explicit institutional agent, would therefore produce:

CDR = 600 / 800 = 0.75

or:

CDR = 75%

The metric ranges from:

0 = institutional agent preserved in every suppression clause

to

1 = institutional agent absent from every suppression clause

CDR does not measure censorship intensity.

It does not measure the number of removals.

It does not measure political bias.

It does not measure whether a moderation decision was correct.

It measures one formal property only:

the frequency with which suppression is linguistically represented without the suppressing institutional actor.

This narrowness is intentional. A metric gains analytical value by preserving the distinction between what it measures and what may subsequently be inferred from it.

5.6 Suppression clause identification

A clause enters the CDR denominator only if it describes an actual or potential restrictive action affecting content, account functionality, circulation, monetization, recommendation, or visibility.

Examples include:

The post was removed.

We suspended the account.

Distribution has been reduced.

This content cannot be recommended.

Monetization has been disabled.

The account is no longer eligible for recommendation.

Repeated violations may result in suspension.

Pure classification clauses such as:

This content violates our policy

do not independently enter the primary CDR denominator unless they are syntactically connected to a restrictive consequence.

They are instead coded under classification variables.

This separation prevents CDR from conflating **classifier deletion** with **censor deletion**.

The two phenomena are related but analytically distinct.

A platform can explicitly identify itself as classifier while omitting itself as censor:

We determined that your post violated our policy. The post was therefore removed.

Conversely, it can omit itself as classifier while appear as censor:

Your post violated our policy, so we removed it.

These configurations should remain distinguishable.

5.7 Classifier-Deletion Rate

Because Section 4 identified **classification without a classifier** as an important precursor to suppression, a complementary measure can be introduced:

Classifier-Deletion Rate (CLR).

Let:

K = total classification clauses

and:

K_0 = classification clauses in which a policy judgment is asserted without an explicit classificatory agent

Then:

CLR = K_0 / K

The distinction is illustrated by:

We classified this content as hate speech.

classifier visible

versus:

This content is hate speech.

classifier absent

and:

We determined that this post violates our policy.

classifier visible

versus:

This post violates our policy.

classifier absent

CLR permits the analysis to determine whether institutional disappearance occurs before enforcement itself.

If high CLR systematically precedes high CDR, the empirical data would support the theoretical sequence proposed in Section 4:

classifier deletion → censor deletion

This relation should not be assumed in advance. It should be tested statistically.

5.8 Passive Moderation Ratio

The **Passive Moderation Ratio (PMR)** measures the frequency of passive constructions among suppression clauses.

Let:

P = passive suppression clauses

Then:

PMR = P / S

PMR must remain separate from CDR because passive voice does not necessarily delete the agent.

For example:

The post was removed by Meta.

is passive but retains institutional agency.

Accordingly:

PMR = 1 for passive structure

but

censor deletion = 0

for that clause.

Conversely, certain agentless constructions may not be canonical passives:

Reduced distribution applies to this content.

The censor is absent even if the sentence does not instantiate a conventional verbal passive.

The empirical comparison between PMR and CDR is therefore itself informative. If CDR substantially exceeds PMR, institutional disappearance cannot be explained by passive voice alone.

That outcome would support the broader suppression-opacity hypothesis.

5.9 Policy-Subject Ratio

The **Policy-Subject Ratio (PSR)** measures the frequency with which policy, standards, safety, rules, procedures, or equivalent abstractions occupy grammatical subject or quasi-agentive position in enforcement-related clauses.

Examples include:

Our policy requires removal.

Community Standards prohibit this content.

Safety rules restrict distribution.

The policy does not allow this material.

Let:

PS = enforcement-related clauses in which policy or equivalent abstraction occupies subject position

and:

E = total enforcement-related clauses

Then:

PSR = PS / E

PSR operationalizes the mechanism identified in Section 3 as **policy-agent substitution**.

A high PSR would indicate that normative instruments frequently occupy linguistic positions otherwise available to institutional actors.

Again, the metric does not establish concealment or intention. It measures formal substitution.

5.10 User-as-Violator Ratio

The **User-as-Violator Ratio (UVR)** measures the frequency with which users, accounts, or user-produced content occupy grammatical subject position in violation clauses.

Examples include:

You violated our policy.

Your account violated our rules.

Your post violates the Dangerous Organizations policy.

Let:

UV = user/content-subject violation clauses

and:

V = total violation clauses

Then:

UVR = UV / V

The metric becomes analytically stronger when compared with institutional-agent visibility.

If a corpus exhibits:

high UVR + high CDR

the grammar allocates explicit violation agency to the governed subject while simultaneously omitting institutional agency from enforcement.

This combination constitutes a measurable form of the asymmetric visibility described earlier in the series (Startari, 2026b).

A derived comparison can therefore be calculated:

Agency Asymmetry Differential (AAD) = UVR − PAVR

where PAVR represents Platform-Agent Visibility Rate, defined below.

Positive values indicate greater grammatical visibility of the user as violator than of the platform as enforcing agent.

Negative values indicate the reverse.

An AAD near zero indicates approximate balance between the two forms of grammatical agency.

This metric should be interpreted cautiously because violation clauses and enforcement clauses perform different semantic functions. It is therefore primarily comparative rather than normative.

5.11 Platform-Agent Visibility Rate

The inverse perspective of CDR is captured by the **Platform-Agent Visibility Rate (PAVR)**.

Let:

A_1 = suppression clauses containing explicit institutional agency

Then:

$PAVR = A_1 / S$

If all suppression clauses can be categorically coded as either agent-present or agent-absent:

$PAVR = 1 - CDR$

The metric is nevertheless useful for reporting because visibility may be easier to interpret positively than deletion.

For example:

CDR = 0.72

can be reported alongside:

PAVR = 0.28

indicating that only 28% of suppression clauses explicitly preserve the institutional actor.

5.12 Content-as-Risk Ratio

The **Content-as-Risk Ratio (CRR)** measures clauses in which content is represented as possessing a moderation-relevant risk property without explicit marking of institutional classification.

Examples include:

This content is harmful.

This post is dangerous.

The material is extremist.

This claim is misleading.

These should be distinguished from mediated formulations such as:

We classified this content as potentially harmful.

or:

Our system identified the material as potentially misleading.

The distinction corresponds to Section 4's contrast between:

institution classifies O as K

and:

O is K

Let:

R_0 = unmediated risk-predicate clauses

and:

R = total risk-classification clauses

Then:

CRR = R_o / R

A high CRR indicates that moderation categories are frequently represented as properties of content rather than outcomes of institutional classification.

This measure therefore complements CLR.

5.13 Suppression-Without-Agent Ratio

CDR measures all suppression clauses lacking explicit institutional agents. A narrower **Suppression-Without-Agent Ratio (SWAR)** can isolate clauses in which the restrictive effect itself remains explicit while both institutional actor and proximate technical executor are absent.

For example:

Distribution was reduced.

The content was removed.

Monetization was disabled.

These differ from:

Our automated system removed the content.

Although the platform as corporate actor may not occupy subject position in the latter, a proximate executor is identified.

SWAR therefore identifies the strongest form of immediate agentlessness.

Its inclusion permits differentiation among:

institutionally explicit enforcement

technically attributed enforcement

completely agentless enforcement

This distinction becomes important when analyzing automation because technical attribution can either increase transparency or redistribute responsibility, depending on whether institutional responsibility remains recoverable.

5.14 Visibility-Reduction Opacity Rate

Deletion and account suspension are comparatively discrete actions. Recommendation restriction, de-ranking, distribution reduction, and eligibility changes operate differently because they alter the probability of circulation without necessarily eliminating content.

A dedicated measure is therefore required.

The **Visibility-Reduction Opacity Rate (VROR)** measures visibility-changing clauses in which neither the institutional actor nor a specific mechanism responsible for the reduction is identified.

Let:

VR = total visibility-reduction clauses

and:

VR_0 = visibility-reduction clauses lacking explicit institutional or technical agency

Then:

$VROR = VR_0 / VR$

Examples of high-opacity visibility clauses include:

Your content has reduced distribution.

This post is not eligible for recommendation.

Visibility may be limited.

Lower-opacity formulations include:

We reduced recommendation of this post because it violated Rule X.

or:

Our recommendation system lowered distribution after the content was classified under Rule X.

VROR is theoretically important because contemporary suppression need not eliminate speech. It can modify the conditions under which speech becomes discoverable.

A corpus limited to content deletion would therefore systematically underestimate the governance of visibility.

5.15 From clause metrics to Suppression Opacity

CDR and its secondary ratios isolate particular linguistic mechanisms. SOI measures their broader institutional consequence.

The **Suppression Opacity Index** is designed to estimate the degree to which the moderation communication prevents reconstruction of the enforcement chain.

The full institutional sequence can be represented as:

I → R → K → M → S → A

where:

I = institutional actor

R = rule

K = classification

M = decision or enforcement mechanism

S = sanction

A = appeal pathway

Maximum traceability exists when each component is identifiable and the relations among them are explicit.

Suppression opacity increases as components disappear or become indeterminate.

The index therefore evaluates six principal dimensions:

institutional-agent visibility;

restrictive-action visibility;

rule specificity;

classification visibility;

decision-mechanism visibility;

appeal-path visibility.

A seventh variable, **traceability coherence**, evaluates whether the available components can be connected into a reconstructable decision chain.

5.16 SOI coding model

For each moderation notice, each dimension can initially be coded binarily:

0 = information visible

1 = information opaque or absent

Let:

I_o = institutional agent opacity

S_o = sanction opacity

R_o = rule opacity

K_o = classification opacity

M_o = mechanism opacity

A_o = appeal opacity

T_o = traceability opacity

A basic unweighted SOI can then be calculated as:

$SOI = (I_o + S_o + R_o + K_o + M_o + A_o + T_o) / 7$

The index ranges from:

0 = minimum suppression opacity

to

1 = maximum suppression opacity

A notice in which all seven dimensions remain transparent receives:

SOI = 0

A notice in which none can be reconstructed receives:

SOI = 1

The use of an unweighted model is preferable in the initial study because assigning unequal weights before empirical validation would introduce additional theoretical assumptions. Weighted versions can later be developed if validation demonstrates that particular dimensions contribute more strongly to perceived or procedural opacity.

5.17 Example of low SOI

Consider:

We removed your post after our automated system classified it under Section 3.2 of the Dangerous Organizations policy. A human reviewer confirmed the decision. You may appeal within 30 days through the link below.

The notice identifies:

institutional actor;

restrictive action;

policy;

classification;

automated involvement;

human review;

appeal mechanism.

Its CDR is low because *we* remains explicit as institutional agent.

Its SOI is also low because most of the decision chain can be reconstructed.

The notice can still describe a substantively incorrect moderation decision. Low opacity does not imply correct enforcement.

5.18 Example of intermediate SOI

Consider:

Your post was removed for violating our Dangerous Organizations policy. You may request another review.

Here:

the sanction is explicit;

the policy category is identified;

an appeal mechanism exists.

However:

the censor is grammatically absent;

the classificatory process is unspecified;

the decision mechanism is unknown.

CDR records agent deletion.

SOI records partial institutional opacity.

5.19 Example of high SOI

Consider:

Your content violates our standards and is not eligible for recommendation.

The user knows:

the content has been classified negatively;

recommendation eligibility has changed.

The user may not know:

who made the determination;

which exact standard applies;

whether automated detection was involved;

what specific feature triggered the classification;

whether the content was reviewed;

whether the visibility restriction can be appealed.

The sentence also contains a content-as-violator construction and an agentless visibility consequence.

Accordingly, several metrics rise simultaneously:

CLR;

UVR or content-as-violator frequency;

VROR;

SOI.

The co-occurrence of these variables is theoretically more significant than any single feature.

5.20 Accountability-Traceability Rate

Because opacity is expressed negatively, the methodology also introduces the **Accountability-Traceability Rate (ATR)** as its positive counterpart.

For the basic model:

ATR = 1 − SOI

ATR estimates how much of the institutional enforcement chain remains reconstructable from the user-facing communication.

An SOI of 0.71 therefore corresponds to:

ATR = 0.29

This indicates that only a minority of the coded accountability dimensions remain clearly visible.

The metric is useful because suppression opacity and accountability traceability describe the same informational architecture from opposite directions.

5.21 Appeal-Path Visibility Rate

The **Appeal-Path Visibility Rate (APVR)** measures the proportion of moderation notices containing a clearly identifiable procedure for contesting or requesting review of the decision.

Let:

A_1 = notices containing visible appeal pathway

and:

N_e = moderation notices involving an enforceable decision

Then:

APVR = A_1 / N_e

The appeal pathway must be operationally identifiable. Generic language indicating that users *may have options* should not be coded as equivalent to a specific review mechanism.

Human Rights Watch's documentation of cases in which Palestine-related users reportedly could not access effective appeal mechanisms demonstrates why this variable requires independent measurement rather than being inferred from the existence of formal platform policies (Human Rights Watch, 2023).

The distinction is between:

appeal exists institutionally

and:

appeal is visible and accessible in the enforcement communication.

The present paper measures the second.

5.22 Compiled Rule Distance

The integration of the compiled-rule framework permits an additional derived measure: **Compiled Rule Distance (CRD)**.

CRD estimates how many institutional layers separating authority from output remain visible in the moderation explanation.

The complete theoretical chain is:

institution → policy → operational rule → classification mechanism → decision → sanction

A maximally explicit notice allows several of these transitions to be reconstructed.

A maximally compressed notice presents only the terminal outcome.

For coding purposes, each recoverable stage receives a binary value:

institution identified

policy identified

operational criterion identified

classification mechanism identified

decision identified

sanction identified

The proportion of missing stages constitutes CRD.

The measure operationalizes a prediction derived directly from the compiled-rule model: **as governance becomes represented further from its institutional origin, the visible distance between rule authorship and rule execution increases** (Startari, 2025).

CRD differs from SOI because it focuses specifically on the architecture of rule execution rather than the total accountability environment.

An appeal mechanism, for example, affects SOI but not necessarily CRD.

5.23 Comparison across political categories

The value of the proposed metrics lies primarily in comparison.

Aggregate CDR alone cannot establish political asymmetry.

Suppose the corpus produces:

overall CDR = 0.68

That result indicates that institutional agents are absent from 68% of suppression clauses. It does not indicate whether politically sensitive speech receives distinct treatment.

The corpus must therefore calculate CDR, SOI, CLR, VROR, and related measures independently for each political category and for appropriate comparison groups.

For category *j*:

$\mathbf{CDR_j = A_{0j} / S_j}$

This permits comparison among:

Palestine-related content;

Iran-related content;

anti-sanctions content;

anti-war content;

anti-imperial content;

and nonpolitical moderation controls.

The inclusion of control categories is essential.

If high CDR occurs equally in spam notices, copyright notices, sexual-content notices, and political-speech notices, agent deletion may reflect a general platform communication convention rather than a politically specific structure.

If politically contested categories exhibit significantly higher SOI after controlling for platform and sanction type, the political-opacity hypothesis becomes stronger.

The methodology must therefore distinguish between:

general moderation opacity

and:

politically asymmetric moderation opacity.

Without this control, claims about political specificity would exceed the evidence.

5.24 Platform comparison

Metrics should also be calculated separately by platform.

Platforms differ in:

interface design;

policy terminology;

notice length;

appeal architecture;

automation practices;

and enforcement vocabulary.

Pooling all platforms without controlling for these differences could produce misleading results.

For each platform p:

CDR_p

SOI_p

CLR_p

$VROR_p$

$APVR_p$

should be independently calculated.

Cross-platform comparison can then determine whether censor deletion is:

platform-specific;

policy-specific;

politically specific;

or a more general property of digital moderation discourse.

5.25 Sanction-type comparison

The hypothesis also predicts variation according to sanction type.

Content removal is likely to generate different linguistic structures from de-ranking or recommendation restriction.

The following categories should therefore be distinguished:

R_1 = content removal

R_2 = account suspension or restriction

R_3 = recommendation reduction

R_4 = search or discovery reduction

R_5 = demonetization

R_6 = feature restriction

R_7 = warning or labeling

R_8 = other visibility modification

This allows the study to test whether visibility-based sanctions produce greater opacity than deletion-based sanctions.

Such a finding would be theoretically important because visibility reduction is structurally less observable to users than outright removal.

The prediction is:

SOI_visibility > SOI_removal

and potentially:

VROR > CDR_removal

These are hypotheses, not assumed outcomes.

5.26 Automation and opacity

The corpus should also distinguish among:

automated enforcement;

human enforcement;

hybrid enforcement;

unspecified enforcement.

This permits testing of one of the paper's central claims: whether automation increases the grammatical or procedural distance between institutional authority and enforcement output.

The relevant comparison is not:

machine = opaque

versus:

human = transparent.

That would be theoretically crude.

The actual question is whether notices associated with automated or unspecified enforcement produce higher CDR, CRD, or SOI than notices associated with identifiable human review.

A platform can explain automated moderation transparently.

A human review process can remain opaque.

The variable is therefore explanatory visibility, not automation itself.

5.27 Reliability of coding

Because several variables require linguistic judgment, the coding scheme must be tested for inter-coder reliability.

The core categories—active institutional agent, agentless passive, policy subject, user-as-violator, content-as-risk, explicit sanction, explicit rule, decision mechanism, and appeal pathway—should be defined in a codebook before full corpus annotation.

A subset of the corpus should then be independently coded by at least two annotators.

Agreement can be assessed using established reliability statistics appropriate to categorical content analysis, including Cohen's kappa for two-coder categorical variables or Krippendorff's alpha where variable structure, missing data, or multiple coders make it preferable (Cohen, 1960; Krippendorff, 2018).

The use of formal intercoder reliability is especially important because the present paper advances new constructs. CDR can be meaningful only if independent coders can reliably distinguish explicit institutional agency from mere institutional reference.

Likewise, SOI requires sufficiently precise definitions that different coders reach comparable judgments regarding whether rule, mechanism, classification, and appeal information are genuinely visible.

5.28 Reliability thresholds and disagreement

The study should not treat agreement as a cosmetic validation step.

Variables producing unstable annotation should be revised or removed.

Disagreements should be used to identify conceptual ambiguity in the coding architecture.

For example, consider:

Our policy does not allow this type of content.

One annotator might classify *our* as sufficient institutional visibility.

Another might classify *policy* as the grammatical actor and therefore mark institutional agency absent.

The codebook resolves the disagreement by distinguishing:

institutional reference

from:

institutional agency.

The platform is referenced through *our*.

It is not the grammatical agent of the prohibitory relation.

This distinction is theoretically central and should therefore be enforced consistently throughout the corpus.

5.29 Statistical testing

After descriptive frequencies are calculated, inferential analysis can test whether observed differences are associated with political category, platform, policy type, language, sanction type, or review mechanism.

Depending on corpus size and variable distribution, categorical comparisons can employ chi-square tests or exact alternatives where expected cell counts are insufficient.

Regression models can subsequently estimate whether the probability of censor deletion changes after controlling for platform and sanction type.

A simplified logistic specification would model:

Pr(agent deletion = 1)

as a function of:

political category;

platform;

policy category;

sanction type;

language;

review mechanism;

and source type.

Similarly, SOI can be modeled as a continuous or ordinal outcome depending on its final empirical distribution.

The central inferential objective is not merely to determine whether opacity exists, but **where variation in opacity is concentrated**.

5.30 No-intent rule

The methodology imposes a strict interpretive restriction:

linguistic opacity is not evidence of institutional intent to conceal.

A high CDR demonstrates frequent agent deletion.

A high SOI demonstrates weak institutional traceability.

Neither independently demonstrates that the platform intended to hide responsibility.

Intent would require separate evidence, such as internal documentation, design records, explicit communication strategy, testimony, or experimentally demonstrated systematic choices attributable to a concealment objective.

This restriction is necessary to preserve the distinction between observable structure and inferred motive.

The same principle governed the earlier concept of responsibility loss: a grammatical pattern can redistribute responsibility regardless of whether the producer consciously intended that redistribution (Startari, 2026a).

5.31 No-censorship equivalence rule

A second interpretive restriction is equally important:

moderation opacity ≠ censorship by definition.

A platform may legitimately remove direct incitement while communicating the decision opaquely.

A platform may illegitimately suppress lawful political criticism while communicating the decision transparently.

Opacity and substantive legitimacy occupy different dimensions.

The methodology therefore produces no automatic transformation:

high SOI → censorship

Instead:

high SOI → reduced traceability of restriction

Whether the underlying restriction should additionally be classified as censorship requires a separate normative and institutional analysis.

This limitation increases rather than reduces the usefulness of the measure because it permits CDR and SOI to be applied across contested cases without requiring prior agreement about political legitimacy.

5.32 Falsification conditions

The framework would be weakened if empirical analysis demonstrates that institutional agency is generally preserved across moderation discourse.

For example, the strongest version of the hypothesis would not be supported if the corpus shows:

low CDR;

low CLR;

low SOI;

high APVR;

clear identification of rule and decision mechanism;

and no meaningful difference between political and control categories.

Likewise, the specific claim that compiled platform governance correlates with censor deletion would be weakened if automated enforcement notices consistently preserve institutional agency more clearly than human or unspecified enforcement notices.

The framework must therefore permit null results.

This is methodologically necessary because censorship without a censor is proposed as a measurable phenomenon, not as an interpretation that must be found regardless of data.

5.33 The expected signature of censorship without a censor

The strongest empirical signature would not consist of one high metric.

It would consist of a recurrent configuration.

For example:

high CDR

indicating frequent censor deletion;

high CLR

indicating frequent classifier deletion;

high UVR

indicating explicit representation of the governed subject as violator;

high PSR

indicating displacement of institutional agency toward policy;

high VROR

indicating opaque governance of visibility;

low APVR

indicating weakly visible appeal pathways;

and

high SOI

indicating overall degradation of accountability traceability.

If these properties cluster within politically contested moderation cases, the empirical pattern would correspond closely to the theoretical structure proposed in Sections 2–4.

The speaker remains visible.

The violation remains visible.

The policy remains visible.

The sanction remains visible.

The institutional chain becomes difficult to reconstruct.

That configuration is analytically stronger than passive voice alone.

5.34 Measurement as reconstruction of authority

The proposed methodology ultimately measures how much institutional authority can be reconstructed from its linguistic output.

This is where the connection to the compiled rule becomes most precise.

The compiled rule predicts that authority can persist through formalized execution even after its source becomes unnecessary to the immediate operation of the rule (Startari, 2025).

CDR tests whether the originating or enforcing institution survives grammatically at the point of sanction. CRD tests how much of the rule-execution chain remains visible. SOI tests whether the broader decision architecture remains accountable to reconstruction.

The methodological architecture can therefore be summarized as:

institutional authority

→ compiled rule

→ classification

→ enforcement

→ user-facing representation

and then analytically reconstructed through:

CDR → who disappeared

CLR → who classified

PSR → what replaced the institution

UVR → who remained visible as violator

VROR → how visibility itself was governed

CRD → how far execution moved from visible authority

SOI → how much of the decision chain remains reconstructable

This sequence converts the central proposition of the article into an empirical test.

Censorship without a censor should not be identified because a researcher interprets a notice as impersonal. It should be identified only when a reproducible corpus demonstrates a systematic divergence between the visibility of restrictive effects and the visibility of the institutions responsible for producing them.

The decisive measurement is therefore not whether political speech disappears.

It is whether **institutional agency disappears faster than the evidence of suppression itself** (Startari, 2026a, 2026b).

6. Demonstration: Platform Notices and AI-Mediated Moderation Explanations

The metrics introduced in Section 5 require empirical application before censorship without a censor can be treated as a demonstrated property of platform governance. The present section therefore performs a **pilot demonstration** rather than presenting results from a completed large-scale corpus. Its purpose is methodological: to establish how publicly documented moderation communications can be transformed into analyzable clauses, how CDR, CLR, PSR, UVR, VROR, CRD, and SOI distinguish different forms of institutional visibility, and which claims can and cannot be supported from publicly available evidence.

This distinction is necessary because three types of material frequently become conflated in discussions of platform censorship.

The first consists of **actual user-facing moderation notices** produced by platforms.

The second consists of **platform descriptions of their own enforcement systems**, including transparency pages, policy documents, and appeals documentation.

The third consists of **external descriptions of moderation events** produced by oversight bodies, journalists, researchers, or civil-society organizations.

These materials are not grammatically interchangeable. If an Oversight Board decision states that a Facebook post "was removed," the passive construction belongs to the Board's description unless the decision explicitly reproduces the wording shown to the user. It cannot automatically be counted as evidence that Meta's original notice was itself agentless. Likewise, when Human Rights Watch describes content as having been removed or suppressed, that grammar belongs to Human Rights Watch unless the original platform notification is reproduced (Human Rights Watch, 2023).

The pilot therefore applies the provenance rule established in Section 5. Only original or verifiably reproduced platform language can support clause-level claims about platform grammar. External reports are used to establish the existence, political context, outcome, or disputed classification of moderation events. They are not silently converted into platform notices.

This methodological restriction substantially raises the evidentiary threshold of the argument. It also prevents the analysis from manufacturing the phenomenon it seeks to measure.

6.1 The first contrast: policy-level agency versus notice-level agency

An instructive starting point is Meta's own public description of content enforcement. Meta's Transparency Center explicitly identifies the company as an enforcement actor. Its explanation of takedowns states that when content violates Community Standards, Meta removes it and notifies the affected user so that the user can understand the reason for the action (Meta, 2024a).

This formulation is significant because it establishes that platform discourse is not intrinsically agentless.

At the institutional-policy level, Meta can state:

Meta → removes → content

and:

we → notify → user

The platform remains grammatically visible.

Meta's broader enforcement framework is similarly explicit in presenting moderation as something the company does. The Transparency Center describes its enforcement model through three categories: *remove*, *reduce*, and *inform*, and states that Meta uses technology and review teams to detect, review, and take action on large quantities of content (Meta, 2026).

This provides a useful baseline.

The compiled-rule argument does not predict that institutions must erase themselves from all descriptions of governance. On the contrary, institutional agency may be highly visible at the level at which a company explains its governance architecture while becoming less visible at the level of individual execution.

The relevant empirical comparison is therefore not:

platform language = agentless

but:

Does agent visibility change as discourse moves from rule description to individual enforcement output?

This can be expressed as a cross-level hypothesis:

institutional-policy discourse → higher agent visibility

individual enforcement notice → potentially lower agent visibility

The distinction is theoretically important because the compiled rule operates precisely between those levels. At the policy level, authority describes itself. At the execution level, the rule may be capable of producing a result without reconstructing that authority in each output (Startari, 2025).

A completed corpus should therefore distinguish **governance-level language** from **execution-level language** rather than pooling them.

6.2 Meta and the grammar of reduction

The difference between removal and reduction is particularly useful for testing suppression opacity. Meta publicly states that some content that does not reach the threshold for removal may nevertheless receive reduced distribution. Its Transparency Center describes lowering the distribution of problematic content and explains that recommendation systems are governed by additional guidelines determining what may be distributed less widely (Meta, 2025a).

The official description itself contains relatively visible institutional agency. Meta states that *we* reduce distribution and that its guidelines describe content categories that *we* consider problematic or low quality (Meta, 2025b).

At the institutional level, therefore, the sequence remains recoverable:

Meta → classifies content under distribution criteria → Meta reduces circulation

The user-facing representation can be structurally different.

Instagram's public help material concerning recommendation eligibility explains that an account may become ineligible to be recommended and that, in such circumstances, its content will not be recommended to non-followers (Instagram, n.d.-a).

The distinction illustrates why VROR is necessary. A visibility intervention can be represented not as:

we reduced the distribution of your account

but as a state:

your account is not eligible to be recommended

The two propositions can correspond to closely related governance effects, but they encode authority differently.

The first represents an institutional action.

The second represents an account property.

Formally:

I acts upon O

becomes:

O possesses status K

The moderation event has undergone a grammatical transformation from action to condition.

This is structurally parallel to the transition identified in sanctioned suffering, where an imposed consequence can become linguistically represented as a surrounding condition rather than as the result of an attributable action (Startari, 2026c). The platform context differs institutionally, but the grammatical operation remains comparable.

Recommendation eligibility is particularly useful for the present paper because visibility reduction complicates the intuitive model of censorship. Content does not have to disappear in order for institutional governance to modify its communicative reach. Instagram states that material falling outside its Recommendation Guidelines may be made harder to find, even where the content remains accessible (Instagram, n.d.-b).

Thus:

existence ≠ distribution

and:

availability ≠ visibility

The distinction makes censorship without a censor empirically broader than content deletion while still requiring precise coding. The study must record the actual governance effect rather than presuming that reduced recommendation is equivalent to removal.

6.3 Controlled coding demonstration: removal

A controlled example demonstrates how the metrics respond to progressively compressed formulations.

Consider four versions of the same hypothetical enforcement event.

Version A

We reviewed your post and removed it because we determined that it violated Section X of our Dangerous Organizations policy. You may appeal this decision.

This representation preserves:

institutional actor;

review event;

classification act;

specific rule;

sanction;

appeal pathway.

The suppression clause contains an explicit censor.

The classification clause contains an explicit classifier.

CDR is therefore zero for the relevant suppression clause.

CLR is zero for the classification clause.

PAVR is high.

APVR is positive.

SOI is correspondingly low.

The important point is not that Version A is normatively superior in every possible respect. It is simply more reconstructable.

Now consider:

Version B

Your post was removed because we determined that it violated Section X of our Dangerous Organizations policy. You may appeal this decision.

The suppression clause has become passive and agentless.

The classification remains institutionally attributed through *we determined*.

Consequently:

CDR increases

while:

CLR remains low

and:

APVR remains high.

This demonstrates why CDR and SOI cannot be treated as interchangeable.

Grammatical censor deletion has occurred, but the institutional chain remains partially reconstructable.

Now:

Version C

Your post was removed because it violated our Dangerous Organizations policy. You may appeal this decision.

Two transformations have occurred.

First, the suppressing institution disappears.

Second, the classificatory institution disappears.

The post itself becomes grammatical subject of *violated*.

Accordingly:

CDR increases

CLR increases

UVR/content-as-violator frequency increases

while:

APVR remains high.

SOI rises further but does not reach its maximum because the policy category and appeal pathway remain visible.

Finally:

Version D

Your post violated our policies and is no longer eligible for recommendation.

Here the institutional actor is absent from both classification and visibility reduction.

The exact rule is unspecified.

The decision mechanism is unspecified.

The content becomes the violator.

The visibility consequence becomes an account or content condition.

No appeal pathway is represented.

This produces the strongest predicted combination:

high CLR

high UVR

high VROR

high CRD

high SOI

The controlled sequence demonstrates that suppression opacity is gradual rather than binary.

Nothing requires the final sentence to be false for it to contain less institutional information than the first.

6.4 A documented Palestine-related case: reporting and dangerous-organizations policy

The Washington Post Israel-Palestine case reviewed by the Oversight Board provides a documented example of the difference between content, classification, and enforcement. Meta removed a Facebook post linking to a Washington Post article under its Dangerous Organizations and Individuals policy. The user argued that the post constituted reporting

on the Israel-Hamas war rather than support for Hamas. The Oversight Board concluded that the removal represented over-enforcement of the relevant policy in the context of news reporting (Oversight Board, 2024c~~2024a~~).

The case matters because the final moderation category did not arise from an uncontested property of the content.

The relevant structure was not:

content = prohibited support

but:

content → institutional interpretation → prohibited-support classification → removal

The subsequent review demonstrated that the classification was revisable.

This permits an important inference.

If an enforcement notice had represented the content simply as inherently violating the Dangerous Organizations policy, the grammar would have concealed the institutional contingency later exposed by the reversal.

The enforcement event therefore illustrates the significance of classifier visibility independently of the exact wording of the original user-facing notice.

The case must not, however, be used to assign a CDR score unless the original moderation communication is available. The Oversight Board's statement that the post was removed describes the moderation event; it does not by itself establish the grammatical form Meta displayed to the user.

This distinction is central to the pilot.

The event supports analysis of **classification reversibility**.

It does not automatically supply evidence of **censor deletion in the original notice**.

The difference protects the metric from circular confirmation.

6.5 The *shaheed* case: when a lexical item becomes executable

The Oversight Board's policy advisory opinion concerning *shaheed* offers a stronger demonstration of the transition from political language to executable moderation object. The Board concluded in 2024 that Meta's prior approach to the Arabic term was overbroad and disproportionately restricted expression and civic discourse. It also noted Meta's own finding that variations of the term were responsible for more content removals under its Community Standards than any other single word or phrase (Oversight Board, 2024a~~2024b~~).

The importance of the case for the present article is not simply that content was removed.

It demonstrates how a linguistic form can become connected to enforcement through a policy architecture.

The abstract sequence is:

lexeme L

→ reference relation

→ dangerous-individual classification

→ praise/support interpretation

→ policy violation

→ removal

The Oversight Board's criticism concerned the insufficient contextual differentiation embedded in this sequence. The term possesses uses and meanings that do not automatically amount to praise or endorsement, yet Meta's policy implementation treated a broad set of uses as actionable (Oversight Board, 2024a~~2024b~~).

This is precisely what the compiled-rule model predicts can occur when contextual political language is translated into scalable enforcement conditions.

A complex semantic object becomes institutionally executable.

The critical operation is not the removal alone.

It is the conversion:

linguistic form → governance category

(Startari, 2025).

Once this mapping is stabilized, individual outputs can appear procedurally self-explanatory.

The platform no longer needs to reconstruct for each user the policy history through which *shaheed* acquired its moderation significance.

A notice can operate at the terminal level:

content violated policy

Yet the Oversight Board's review demonstrates that the relationship between content and violation was itself institutionally constructed and contestable (Oversight Board, 2024a~~2024b~~).

The case therefore supplies strong evidence for the distinction between:

classification as property

and:

classification as institutional judgment.

It does not establish the frequency of classifier deletion. That requires original notices.

It establishes why classifier deletion, when present, matters.

6.6 *From the River to the Sea*: the non-equivalence of phrase and violation

The Oversight Board's 2024 decision concerning three Facebook posts containing the phrase *from the river to the sea* provides another demonstration. The Board reviewed the cases together and concluded that the three posts did not violate Meta's Hate Speech,

Violence and Incitement, or Dangerous Organizations and Individuals policies in their specific contexts (Oversight Board, 2024b~~2024c~~).

The result supplies an empirical constraint on any automated or simplified classificatory model:

phrase presence ≠ policy violation

The relevant function must instead include context:

K = f(P, C, R)

where the expression P, contextual information C, and institutional rule R jointly contribute to classification.

This is significant for AI-mediated moderation because automated systems are often used to detect candidate violations at scale, while contextual interpretation may subsequently involve additional automated or human processes. Meta publicly describes its enforcement system as combining technology with review teams rather than as relying upon a single autonomous moderator (Meta, 2026).

The phrase cases consequently illustrate why an AI-mediated enforcement architecture must distinguish between **detection** and **determination**.

A system can detect the presence of a phrase.

That detection does not logically establish its policy meaning.

The difference is:

signal detected

versus:

violation determined

When user-facing explanations collapse these stages, technical identification can acquire the appearance of normative judgment.

This is another form of compiled-rule distance.

The longer institutional chain may be:

model detects signal → system associates signal with candidate category → policy conditions are evaluated → reviewer or additional system determines applicability → sanction is selected

The explanation can become:

your content violated policy

The intermediate stages disappear.

A completed corpus should therefore code whether notices distinguish **detection**, **classification**, and **decision**, rather than treating all references to automated technology as equivalent transparency.

6.7 AI-mediated enforcement is not autonomous enforcement

The term **AI-mediated moderation** requires precise use.

In this article it does not mean that every moderation notice is written by a generative language model. Public evidence does not justify that assumption.

It refers to governance processes in which machine-learning or automated systems participate in detection, classification, prioritization, recommendation, restriction, or routing for human review.

This definition is consistent with research describing automated content moderation as part of a broader sociotechnical process rather than an autonomous replacement for institutional governance (Gorwa et al., 2020).

Platform documentation supports the same distinction. Meta states that it uses both technology and review teams in content enforcement (Meta, 2026). YouTube states that content can be flagged by users or by its automated detection technology and that review

teams determine whether flagged material violates Community Guidelines (YouTube, n.d.-a).

YouTube's description is analytically useful because it explicitly separates stages:

flagging → review → policy determination → enforcement

The technical system does not grammatically absorb the entire decision chain.

The platform states that content can be identified by detection technology, but its review teams decide whether it fails to comply with Community Guidelines (YouTube, n.d.-a).

This provides a low-CRD model of explanation.

The automated system is visible as detector.

The review team is visible as decision-maker.

The policy is visible as standard.

The sanction is visible as consequence.

Such a formulation demonstrates that automation does not inherently require suppression opacity.

That observation is important because it creates a falsifiable baseline for the paper's hypothesis.

If platforms can explain automated moderation through explicit stage differentiation, then high opacity elsewhere cannot be defended as an unavoidable grammatical consequence of automation.

Automation may facilitate institutional distance.

It does not determine it.

6.8 YouTube as a counterexample to maximal opacity

YouTube's current Community Guidelines documentation provides a useful comparative case because its public explanation specifies several elements that SOI treats as accountability variables. YouTube states that when a creator receives a strike, the company communicates what content was removed, which policy was violated, how the decision affects the channel, and what the creator can do next. It also provides an appeal mechanism for warnings, strikes, and content removals (YouTube, n.d.-b).

At the public-policy level, the architecture is comparatively reconstructable:

institutional actor → enforcement action → policy → consequence → appeal

YouTube also uses active institutional constructions such as *we remove* and *we find*, alongside content-centered formulations such as *your content violates our Community Guidelines* (YouTube, n.d.-b).

This coexistence is theoretically more useful than a uniformly agentless example.

It demonstrates that suppression opacity is not determined by a single vocabulary item. The same platform can alternate between:

institution-centered grammar

and:

content-centered grammar.

The empirical object is therefore the **distribution** of these structures.

A completed corpus should ask:

When does YouTube say **we removed**?

When does it say **content was removed**?

When does it say **your content violates**?

When does it identify a review team?

When does it identify automation?

Do these choices vary by policy category?

Do political cases differ from ordinary enforcement cases?

Without those comparisons, isolated sentences cannot support a claim of systematic censor deletion.

The counterexample strengthens the proposed method because it shows what evidence would falsify an overly strong formulation of censorship without a censor.

6.9 TikTok and visible appeal architecture

TikTok's public support documentation provides another comparative case. TikTok states that users whose content is removed receive a notification explaining why the content was removed, which guideline was violated, and how to submit an appeal (TikTok, n.d.-a).

The statement preserves significant institutional agency through second-person and first-person relations: the platform notifies the user and provides an appeal pathway.

TikTok also maintains procedures for appealing account bans and recommendation-ineligibility decisions (TikTok, n.d.-b, n.d.-c).

Again, the existence of an appeal system does not establish that individual enforcement decisions are accurate or transparent in practice.

It establishes a formal accountability pathway.

For SOI, this distinction is critical.

A moderation notice could still receive a high CDR if its suppression clause is agentless while receiving a lower overall SOI because the rule and appeal mechanism remain visible.

Consider:

Your post was removed for violating our Community Guidelines. Appeal.

Under the proposed system:

CDR is high for the suppression clause.

CLR is high if *violating* is treated as an unmediated classification.

Rule visibility is moderate.

Appeal visibility is high.

SOI is therefore intermediate rather than maximal.

This is exactly why a multidimensional index is required.

The grammar may delete the censor without completely erasing procedural accountability.

6.10 Appeals reveal the hidden status of classification

Appeals are analytically valuable because they reveal something that ordinary moderation grammar can obscure: **violation is a revisable institutional status**.

YouTube states that an accepted appeal can result in reinstatement of the content and removal of the associated strike. If the company determines that the original content complied with its guidelines, the enforcement outcome is reversed (YouTube, n.d.-a).

Instagram similarly allows users in eligible cases to request another review when content has been removed (Instagram, n.d.-c).

The logical significance is substantial.

If:

O = violation

were an intrinsic relation, institutional review could not change it.

Appeal demonstrates instead:

I_1 classifies O as K

followed potentially by:

I_2 reclassifies O as ¬K

The object has not necessarily changed.

The institutional judgment has.

Therefore the grammatical formulation:

O violates R

compresses an underlying proposition more accurately reconstructed as:

I determined that O violates R.

This does not make ordinary declarative violation language invalid. Institutional systems routinely announce judgments as propositions.

It does, however, establish why the absence of the classifier matters analytically.

The appeal mechanism exposes the invisible predicate:

determined that

which the shorter moderation sentence removes.

6.11 Reversal as evidence of institutional contingency

Oversight Board cases provide particularly clear demonstrations of this principle because Meta sometimes reverses enforcement decisions after the Board brings cases to its attention. In its 2024 summary decisions concerning reports on the war in Gaza, the Board described cases in which Meta reversed its original decisions after the content was selected for review (Oversight Board, 2024d).

Such reversals should not be interpreted as proof that the original decisions were politically motivated.

They establish something narrower and methodologically valuable:

the original enforcement output was contingent and revisable.

Consequently, moderation grammar that presents the original policy relation as self-evident can be compared against the institutional record showing that the relation required judgment and could later be altered.

Reversal therefore offers an external validation variable for classifier opacity.

A completed dataset could code:

appeal/review outcome = upheld / modified / reversed

and compare it with the original notice's CLR and SOI.

This would permit a stronger empirical test:

Are highly opaque classifications more likely to be reversed?

No answer should be presumed.

If the correlation is absent, the framework must report that result.

If it exists, the relationship would connect grammatical opacity to measurable decision instability.

6.12 Documented Palestine-related suppression and the limits of inference

Human Rights Watch's 2023 investigation remains important for corpus construction because it assembled more than a thousand reported cases involving Palestine-related content on Facebook and Instagram and identified several forms of enforcement, including removals, account restrictions, limits on engagement, and reduced visibility (Human Rights Watch, 2023).

The report also documented difficulties with appeals in a substantial subset of the cases it reviewed (Human Rights Watch, 2023).

For the present project, this material supplies an event pool rather than automatic grammatical evidence.

Each documented case should be separated into:

E = enforcement event

T = original platform text

C = political context

R = external assessment

The existence of E can be established from documentation.

The linguistic metrics require T.

Political analysis requires C.

Claims about error or rights impact may draw upon R.

The four variables should never be merged.

This evidentiary architecture is particularly important in politically polarized research because an external organization may use the term *censorship* while the platform uses *enforcement*, *removal*, *reduction*, or *ineligibility*. The present paper should preserve those terminological differences rather than selecting one vocabulary in advance.

The research question is not resolved by naming the event.

It is resolved by reconstructing it.

6.13 Pilot transformation of a political moderation event

A generic political moderation event can now be reconstructed across its full institutional chain.

Assume that a user publishes a post containing politically contested material.

The full governance sequence could be:

1. User produces political utterance U.

2. Automated detector identifies signal L.

3. L triggers candidate category K.

4. Automated or human process evaluates contextual conditions C.

5. Platform determines that U falls within policy rule R.

6. Rule R maps classification K onto sanction S.

7. Platform applies S.

8. User receives notice N.

9. User may or may not access appeal A.

This is the underlying institutional architecture.

Now assume notice N reads:

Your content violated our policies and its distribution has been reduced.

The notice preserves:

U as *your content*;

K only implicitly as *violation*;

S as *distribution has been reduced*.

It deletes or leaves unspecified:

detector;

classifier;

contextual review;

specific rule;

institutional decision-maker;

institutional enforcing actor;

appeal pathway.

The difference between the decision architecture and the notice can be expressed as **representational compression**.

Let:

D = number of institutionally relevant decision stages

and:

V = number of stages visible in the notice

Then the simplest representation of decision visibility is:

V/D

The closer V approaches zero, the greater the informational distance between governance and its linguistic representation.

CRD operationalizes a related relation using predefined institutional stages.

This is where the compiled rule becomes empirically visible.

The output need not reproduce the chain because the chain has already been operationally compressed into the rule.

What remains for the user is the terminal state.

6.14 Controlled comparison: explicit automation versus automated inevitability

Consider two AI-mediated explanations.

Explanation A

Our automated detection system flagged this post for review. A reviewer determined that it violates Rule X, and we removed it. You can appeal the decision.

This notice identifies:

technical detector;

human classifier;

institutional censor;

specific rule;

sanction;

appeal.

Automation is explicit but institutional traceability remains high.

Now:

Explanation B

Our systems detected a violation, so the content was removed.

This notice identifies a technical system as detector.

It does not identify the classifier separately.

It represents *violation* as detected rather than determined.

The suppression clause is passive.

The institutional censor is absent.

The important transformation is epistemic.

Explanation A states:

system detected candidate → reviewer determined violation

Explanation B states:

system detected violation

The distinction collapses detection and normative classification.

Finally:

Explanation C

Violating content is automatically removed.

No institutional or technical actor needs to appear.

The rule itself becomes self-executing.

The grammatical architecture is:

property K → consequence S

This is the strongest compiled form.

Authority is present only in the existence of the rule.

The sentence itself represents no decision-maker.

The three examples generate different CRD and SOI values despite all involving automated moderation.

This demonstrates why the variable **automation = yes/no** is insufficient.

The relevant issue is **how automation is represented in relation to institutional responsibility**.

6.15 Detection verbs and decision verbs

The pilot suggests an additional linguistic distinction that should be incorporated into the final coding manual: **detection verbs versus decision verbs**.

Detection verbs include:

detect

identify

flag

match

find

recognize

Decision verbs include:

determine

decide

classify

conclude

review

rule

The distinction is not absolute; verbs such as *identify* can function classificatorily depending on context. Nevertheless, the semantic difference matters.

Automated systems can detect features.

Institutional governance determines what those features mean for enforcement.

When the same verb covers both operations, technical detection can be represented as though it were normative determination.

For example:

Our system detected the phrase.

differs from:

Our system detected a policy violation.

The first reports observation.

The second encodes classification within the object of detection.

The linguistic structure therefore deserves separate coding.

A new variable can be designated **Detection-Decision Collapse (DDC)**.

DDC = 1 where the explanation represents an automated system as directly detecting a normative violation without exposing an intervening classificatory operation.

DDC = 0 where detection and determination remain linguistically distinct.

The measure should remain exploratory until intercoder reliability establishes that the distinction can be applied consistently.

6.16 The value of negative cases

A robust demonstration requires examples in which censorship without a censor does **not** occur.

Meta's institutional statement that it removes violating content is one such example because the company remains explicitly represented as agent (Meta, 2024a).

YouTube's explanation that its review teams decide whether flagged content complies with its rules is another because it identifies both detection and institutional review (YouTube, n.d.-a).

TikTok's statement that it provides users with the reason for removal, the relevant guideline, and an appeal mechanism likewise preserves several SOI dimensions at the policy-design level (TikTok, n.d.-a).

These counterexamples prevent the framework from becoming tautological.

The hypothesis cannot be:

all moderation language deletes the censor.

Available evidence already shows that this proposition is false.

The defensible hypothesis is:

certain moderation structures reduce institutional agency more strongly than others, and their distribution can be measured across enforcement contexts.

The politically consequential hypothesis is narrower still:

those structures may be distributed asymmetrically across political categories.

That proposition remains empirical.

6.17 Why the political comparison cannot be simulated

At this stage, it would be methodologically invalid to assign numerical CDR or SOI differences to Palestinian, Iranian, anti-sanctions, anti-war, and anti-imperial discourse without a coded corpus of original platform communications.

Theoretical plausibility is not measurement.

Documented complaints are not frequency estimates.

Individual reversals are not prevalence.

Oversight cases are selected cases and cannot be treated as random samples of total platform enforcement.

The present pilot therefore refuses the following unsupported transformation:

documented political moderation controversy → systematic political censor deletion

Instead, the evidence supports:

documented political moderation controversies → suitable test environment for measuring censor deletion

That distinction preserves the falsifiability established in Section 5.

The eventual empirical result could show strong political asymmetry.

It could show platform-wide opacity independent of politics.

It could show that certain political categories receive more detailed explanations because platforms anticipate controversy.

It could show no meaningful difference.

Each outcome remains theoretically interpretable.

6.18 From public documentation to corpus protocol

The pilot permits a concrete corpus procedure.

Each selected enforcement case should first be assigned a unique event identifier.

The researcher should then preserve the original notification through screenshot, exported account-status data, authenticated reproduction, or another verifiable source.

The moderation message should be transcribed exactly.

The transcription should be segmented into clauses.

Each clause should be coded for:

institutional agent;

technical agent;

passive structure;

suppression predicate;

classification predicate;

policy subject;

user/content subject;

risk predicate;

specific rule;

sanction type;

visibility effect;

detection language;

decision language;

and appeal language.

Notice-level coding should separately record:

platform;

political category;

moderation category;

language;

date;

review mechanism;

appeal availability;

appeal result;

and provenance.

Only after annotation should CDR and related metrics be calculated.

This order matters.

If cases are first labeled as censorship and then coded, the dataset will inherit the researcher's normative conclusion.

The method requires the reverse:

code structure first → compare distributions → interpret result

This maintains continuity with the epistemic restriction established in the previous papers of the series, where responsibility must be derived from observable linguistic structure rather than inferred from political expectation (Startari, 2026a, 2026b).

6.19 What the pilot already demonstrates

Although the pilot does not establish population-level frequencies, it does demonstrate several propositions.

First, platform governance contains **multiple linguistically separable stages**: detection, classification, decision, enforcement, notification, and appeal. Official platform documentation itself distinguishes several of these stages (Meta, 2026; TikTok, n.d.-a; YouTube, n.d.-a).

Second, restrictive governance extends beyond removal. Platforms explicitly describe distribution reduction, recommendation ineligibility, account restrictions, and other interventions affecting circulation (Meta, 2025a; Instagram, n.d.-a).

Third, political classifications can be institutionally revisable. Oversight Board cases concerning *shaheed*, *from the river to the sea*, news reporting, and Gaza-related content demonstrate that contextual interpretation can materially alter whether specific content is considered violative (Oversight Board, 2024a, 2024b, 2024c, 2024d).

Fourth, automation does not eliminate institutional stages. Platforms themselves describe combinations of technology and human or institutional review (Meta, 2026; YouTube, n.d.-a).

Fifth, it is linguistically possible to compress those stages until the final explanation contains only the governed object, the violation category, and the sanction.

The first four propositions are externally documentable.

The fifth is demonstrated formally.

What remains to be established empirically is its frequency and political distribution.

6.20 From responsibility loss to executable disappearance

The demonstration also clarifies the relationship between this paper and the preceding theoretical sequence.

Responsibility loss described a condition in which an outcome survives while the actor responsible for producing it loses grammatical visibility (Startari, 2026a).

In platform governance, the same relation can be instantiated as:

removal visible → remover absent

or:

distribution reduction visible → reducing institution absent

Asymmetric visibility adds a second relation:

user visible as violator → platform less visible as enforcer

(Startari, 2026b).

Sanctioned suffering adds the transformation from action to environmental condition:

institution reduces distribution → account has reduced distribution

(Startari, 2026c).

The compiled rule explains why the resulting output remains operationally complete despite its reduced institutional grammar:

authority has already been encoded upstream

(Startari, 2025).

The moderation notice therefore does not need to reproduce authority in order for authority to remain effective.

This is the crucial distinction.

Grammatical disappearance is compatible with operational persistence.

Indeed, the pilot suggests that the two can coexist systematically.

The rule executes.

The sanction occurs.

The user receives the effect.

The institution can remain recoverable only indirectly through the ownership of the platform, the existence of the policy, or the possibility of appeal.

6.21 The strongest test

The strongest empirical test of censorship without a censor would therefore not search for politically dramatic examples.

It would compare structurally similar moderation events.

Suppose two sets of notices involve the same platform, same sanction type, similar content format, same language, and similar interface version.

Set A concerns ordinary nonpolitical policy violations.

Set B concerns politically contested speech.

The relevant comparison is:

CDR_A versus CDR_B

CLR_A versus CLR_B

SOI_A versus SOI_B

APVR_A versus APVR_B

CRD_A versus CRD_B

If no meaningful differences appear, there is no evidence from these measures that political content produces a distinctive grammar of suppression.

If differences appear only between platforms, the phenomenon may be architectural rather than political.

If differences appear primarily by sanction type, visibility governance rather than political classification may explain the pattern.

If differences persist for political categories after those factors are controlled, the hypothesis of politically asymmetric suppression opacity becomes substantially stronger.

The method thus permits the argument to fail.

That possibility is a requirement, not a weakness.

6.22 Demonstration outcome

The pilot does not establish that platforms systematically censor Palestinian, Iranian, anti-sanctions, anti-war, or anti-imperial speech through agentless grammar.

It establishes a more precise result.

Platform enforcement consists of an institutional sequence that can be represented with high or low levels of grammatical and procedural traceability.

Public platform documentation demonstrates that explicit agency is possible.

Public oversight cases demonstrate that political classifications are context-dependent and revisable.

Platform visibility systems demonstrate that restriction extends beyond deletion.

Controlled transformations demonstrate that the same enforcement event can be linguistically represented while progressively deleting classifier, censor, mechanism, and decision chain.

The empirical question is therefore no longer conceptual.

It is measurable.

A completed corpus can determine whether the structures identified here occur frequently, whether they cluster around particular types of enforcement, and whether their distribution changes when the moderated object is politically contested.

The pilot consequently supports the formal proposition underlying the paper:

suppression and suppressor are separable variables in moderation discourse.

The existence of the first does not guarantee grammatical visibility of the second.

The strongest form of the phenomenon occurs when several transformations converge:

the platform defines the rule;

the rule becomes executable;

the system classifies the speech;

the classification triggers the sanction;

the sanction alters visibility;

the notice represents the content as violative;

and the institution appears nowhere as the grammatical source of the intervention.

At that point, the system has not become authorless.

Authority has not vanished.

The decision has not ceased to be institutional.

What has changed is its linguistic surface.

The political speaker appears.

The prohibited category appears.

The rule appears.

The consequence appears.

The censor survives operationally while disappearing grammatically.

That is the empirical signature the full corpus must test.

6.23 Limitations

Five limitations constrain what the present article can claim. They are stated here so that the framework is evaluated against its actual evidentiary base rather than against the scope of the protocol it specifies.

First, evidentiary scope. The demonstration in this section draws on publicly available platform documentation and a small set of Oversight Board decisions and summary decisions. It illustrates that the coding procedure can be applied and that it discriminates between notices, but it does not constitute the corpus specified in Section 5. The metrics defined there, including CDR, CLR, PMR, PSR, UVR, PAVR, CRR, SWAR, VROR, ATR, APVR, CRD, and the composite SOI, are advanced as instruments accompanied by the falsification conditions under which they would fail. Aggregate values, inter-coder reliability statistics, and the significance tests described in Section 5.29 await the coded corpus.

Second, access asymmetry. Individual moderation notices are delivered privately to affected users and are not systematically archived. Any corpus assembled from user-submitted notices, screenshots, or civil-society documentation will therefore be non-random, weighted toward users who possess the resources and motivation to report suppression. Findings drawn from such a corpus describe the population of documented cases, not the population of enforcement events.

Third, source volatility. Platform policy pages, transparency documentation, and help-center articles are undated and revised continuously. The formulations analyzed here reflect the state of those documents at the time of consultation, and a subsequent revision may alter the grammatical features coded. Replication therefore requires archived snapshots rather than live links.

Fourth, construct validity. The metrics measure the grammatical availability of institutional agency within enforcement language. They do not measure whether users in fact fail to contest decisions, whether appeals in fact succeed less often, or whether comprehension in fact degrades. The claim advanced here concerns textual affordances for accountability, not realized accountability outcomes. Establishing the latter requires user-comprehension studies and appeal-outcome data that lie outside the present design.

Fifth, comparative design. The five political environments examined in Section 4 were selected because each tests a distinct classification boundary, not because they exhaust the relevant variation. Absent a matched comparison category of politically opposite

Formatted: Font: (Default) Times New Roman, 12 pt, Bold

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

speech, any measured differential in opacity remains confounded with topic prevalence and enforcement volume. For that reason no differential claim is advanced here. Construction of the matched comparison category, and the differential testing it makes possible, is reserved for the next article in this series.

None of these limitations weakens the conceptual argument, which concerns the conditions under which suppression becomes describable without a describable suppressor. They do, however, delimit what the present demonstration establishes: that the phenomenon is specifiable, measurable in principle, and variable across platforms, and that the negative cases in Sections 6.8 and 6.9 already show the variance a full corpus would have to explain.

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Line spacing: Multiple 1,15 li

7. Conclusion: From Content Moderation to Responsibility Moderation

Platform governance is ordinarily evaluated through the content it permits, removes, restricts, recommends, demonetizes, or suppresses. These outcomes remain indispensable objects of analysis, but they capture only one dimension of moderation. A second dimension concerns the representation of the intervention itself: whether the institutional actor responsible for modifying the visibility of speech remains linguistically identifiable after the decision has been executed.

This article has defined **censorship without a censor** as the platform-mediated condition in which political speech can be removed, restricted, de-ranked, demonetized, delayed, excluded from recommendation, or otherwise made less visible while the institutional actor responsible for the restrictive action becomes grammatically obscured. The concept does not establish that every moderation intervention constitutes censorship. It identifies a narrower formal problem: suppression and suppressor can become linguistically separable variables.

That separation matters because an enforcement outcome can remain completely visible while its institutional source becomes difficult to reconstruct.

The post is removed.

The account is restricted.

Distribution is reduced.

Recommendation eligibility is withdrawn.

A policy is violated.

A risk category is activated.

Yet the institution that established the policy, defined the category, configured the system, authorized the sanction, and continues to govern the enforcement architecture may occupy no explicit agentive position in the sentence communicating the result.

The central claim is therefore not that authority disappears from automated governance. The opposite is structurally possible. Authority can remain fully effective while its linguistic visibility declines.

This distinction connects platform moderation to the theoretical sequence developed across this series. **Responsibility loss** established that consequential events can remain visible even when the grammatical pathway connecting those events to responsible actors deteriorates (Startari, 2026a). **Asymmetric visibility** showed that discourse can distribute agency, intentionality, and causal prominence unequally across actors occupying the same political structure (Startari, 2026b). **Sanctioned suffering** demonstrated how imposed consequences can become linguistically represented as conditions, pressures, or abstract processes whose responsible institutional agents become less salient (Startari, 2026c).

Platform moderation reproduces elements of all three mechanisms, but adds an institutional complication. The actor capable of disappearing from the explanation may also be the actor whose governance system produced the event being explained.

The resulting structure is therefore reflexive.

A platform governs speech.

The platform subsequently describes that governance.

Its language determines how visible the platform remains as the agent of its own intervention.

This is the transition from **content moderation** to **responsibility moderation**.

Responsibility moderation does not denote a separate platform policy category. It describes the second-order linguistic effect through which the representation of enforcement regulates the visibility of responsibility for enforcement. The platform controls one variable through moderation itself—whether and how speech circulates—and another through the language of moderation—how easily the intervention can be attributed to an institutional actor.

These processes need not result from deliberate concealment. The framework developed here does not infer intention from grammatical structure. Standardization, interface economy, automation, policy consistency, localization, and procedural design can all produce impersonal enforcement language without any demonstrated intention to obscure responsibility. The analytical object remains the observable result: whether institutional agency survives in the final representation.

The concept of the **compiled rule** provides the mechanism connecting these levels. Institutional decisions must become operational before they can be repeatedly enforced at platform scale. Policy is translated into classifications, thresholds, procedural conditions, reviewer instructions, technical systems, and enforcement consequences. Once institutional authority has been compiled into an executable relation, the individual enforcement event no longer requires the originating authority to be linguistically reconstructed each time the rule operates (Startari, 2025).

The transformation can be summarized as:

institution determines

→ policy specifies

→ rule operationalizes

→ system classifies

→ sanction executes

→ **notice reports**

At the beginning of this sequence, authority is explicit.

At the end, only the output may remain.

This explains why censorship without a censor cannot be reduced to passive voice. Passive constructions are one possible surface manifestation of a deeper institutional transformation. The decisive structural condition is that a rule can continue reproducing the consequences of authority even when the authority that authored and maintains the rule no longer appears at the point of execution.

The rule does not become independent.

It becomes executable.

The distinction is critical.

A platform policy does not author itself.

A classifier does not independently determine which political categories deserve enforcement consequences.

A threshold does not choose its own acceptable rate of error.

A recommendation system does not independently establish the normative meaning of dangerous, hateful, misleading, extremist, or prohibited speech.

An appeal system does not determine its own jurisdiction.

These relations originate within governance structures.

The disappearance of the institution from the terminal sentence therefore cannot be equated with the disappearance of institutional authorship.

This is what the **Compiled Rule Distance (CRD)** proposed in Section 5 is designed to capture: the visible distance separating an enforcement output from the institutional architecture that made that output possible. CRD complements the narrower **Censor-**

Deletion Rate (CDR), which measures whether the enforcing institution survives grammatically, and the broader **Suppression Opacity Index (SOI)**, which measures whether the decision chain remains reconstructable across agent, rule, classification, mechanism, sanction, and appeal.

Together, these measures convert a rhetorical intuition into a falsifiable research program.

CDR asks:

Who disappeared from the suppression clause?

CLR asks:

Who disappeared from the classificatory judgment?

PSR asks:

Does policy occupy the grammatical position vacated by institutional agency?

UVR asks:

How frequently does the governed subject remain explicitly visible as violator?

VROR asks:

How opaque is the governance of distribution and recommendation?

CRD asks:

How much institutional structure separates visible output from visible authority?

SOI asks:

How much of the complete enforcement chain can the affected user reconstruct?

These measures matter because the study of political moderation cannot remain confined to removal counts. Contemporary platform governance operates over visibility itself. Material can remain published while losing recommendation eligibility. Accounts can remain accessible while facing restrictions. Content can remain technically available while

its distribution is reduced. Monetization, discovery, ranking, searchability, recommendation, and account functionality all constitute possible sites of intervention.

Consequently:

presence is not equivalent to visibility.

A political statement can exist without circulating under the same conditions as comparable material.

This distinction expands the conventional object of censorship analysis. A binary framework asking whether a statement is present or absent cannot fully describe systems capable of governing the probability with which users encounter that statement.

Yet the paper has deliberately avoided converting all visibility governance into censorship. Such a move would destroy the analytical precision the framework is designed to create. Platforms necessarily make distributional decisions. Recommendation systems cannot recommend everything equally. Moderation systems confront direct incitement, targeted hatred, violent propaganda, harassment, fraud, and other forms of content requiring institutional response. The existence of moderation therefore cannot itself constitute evidence of illegitimate suppression.

The relevant questions must remain separated.

Was content restricted?

is an empirical question.

Who restricted it?

is an agency question.

Why was it restricted?

is a classification question.

How was the decision produced?

is a procedural question.

Can the decision be contested?

is an accountability question.

Was the restriction justified?

is a substantive and normative question.

Was the restriction politically asymmetric?

is a comparative empirical question.

The framework developed here prevents answers to one question from being substituted for evidence concerning another.

This separation is especially important for Palestinian, Iranian, anti-sanctions, anti-war, and anti-imperial speech. These categories were selected because they generate political contexts in which classification is often contested and contextual information can be critical. They are not treated as equivalent political positions, nor are they presumed to constitute legitimate speech in every instance. Political orientation cannot substitute for policy analysis.

The Palestine-related cases discussed in this article demonstrate why that methodological restriction is necessary. Oversight Board decisions concerning *shaheed*, *from the river to the sea*, and reporting involving designated organizations show that identical or similar linguistic material can receive different policy interpretations depending on context and communicative function (Oversight Board, 2024a, 2024b, 2024c). Such cases establish that policy violation cannot always be treated as an intrinsic lexical property.

The difference can be represented formally:

O is K

is often a compressed representation of:

I determines that O satisfies K under R and C

where:

O = object of moderation

I = institutional classifier

K = moderation category

R = applicable rule

C = relevant context

The grammatical compression removes I and may remove C.

Once the classification becomes executable, K can directly activate sanction S:

K → S

At that stage, the user-facing notice can reduce the complete institutional sequence to:

O violated R; therefore S.

The political interpretation has become a procedural consequence.

This is the point at which the distinction between **classification without a classifier** and **censorship without a censor** becomes theoretically decisive.

Classifier deletion occurs first when an institutional judgment becomes represented as a property of the moderated content:

This content is harmful.

This post violates policy.

This material is extremist.

Censor deletion occurs when the resulting restriction becomes represented without the institutional actor performing it:

The content was removed.

Distribution was reduced.

The account was restricted.

The strongest form of suppression opacity appears when both mechanisms converge:

content becomes the violator

while

institution becomes absent as enforcer.

This structure produces the double displacement predicted by asymmetric visibility (Startari, 2026b).

The governed subject remains grammatically explicit.

The governing subject recedes.

The post violates.

The account breaches.

The content presents risk.

Policy prohibits.

Safety requires.

The system detects.

The platform may no longer need to appear as the agent responsible for the resulting intervention.

The significance of this pattern should not be overstated before corpus measurement. Section 6 demonstrated that major platforms also use explicit institutional language. Platform documentation contains formulations in which companies openly describe themselves as removing content, reducing distribution, reviewing decisions, and providing

appeals. These counterexamples are not anomalies to be discarded. They are necessary to the theory.

If explicit institutional agency is possible, then censor deletion is not an unavoidable property of digital moderation.

Its distribution becomes the empirical question.

This leads to the most important methodological consequence of the paper. Evidence of politically asymmetric suppression opacity requires comparative data.

A high CDR for Palestine-related notices means little in isolation if the same platform produces an equally high CDR for spam, copyright, nudity, fraud, and ordinary nonpolitical violations. In that case, agent deletion may describe the platform's general notification grammar rather than a political asymmetry.

The relevant test is comparative:

political enforcement vs. nonpolitical enforcement

while controlling, where possible, for:

platform;

sanction type;

interface;

language;

policy architecture;

review mechanism;

and period.

Only then can the analysis distinguish **general institutional opacity** from **politically asymmetric opacity**.

The same standard applies to Iran, anti-sanctions discourse, anti-war discourse, and anti-imperial speech. Claims of selective suppression require evidence of selective distribution. The political importance of a case cannot substitute for comparative measurement.

This restriction also clarifies the role of publicly documented moderation controversies. Human Rights Watch's findings concerning Palestine-related content establish a substantial body of reported restrictions and procedural concerns suitable for investigation (Human Rights Watch, 2023). They do not independently establish the grammatical structure of every user-facing notice. Oversight Board reversals demonstrate that classifications can be disputed and revised. They do not establish the prevalence of incorrect classification throughout the platform.

The corpus proposed here therefore treats such documentation as evidence of **events and test environments**, while reserving grammatical metrics for original or verifiably reproduced platform language.

This is necessary because censorship without a censor must be measurable in the discourse of the institution being analyzed rather than reconstructed from the language of its critics.

The distinction also protects the framework against circularity.

If the researcher first labels an event as censorship, selects examples because they appear censorial, and then interprets their grammar as evidence of censor deletion, the conclusion has already been embedded in the sampling procedure.

The required sequence is instead:

collect

→ authenticate

→ segment

→ code

→ measure

→ compare

→ interpret

The political conclusion comes last.

This order preserves the methodological principle established throughout the series: responsibility must be demonstrated through observable structure rather than inferred from ideological expectation (Startari, 2026a, 2026b).

The framework also produces a stronger interpretation of automation.

Automated moderation is often described as though automation displaced human or institutional authority. The analysis developed here indicates why that description is incomplete. Automation can displace proximate human execution while leaving institutional authorship intact. A classifier can operate automatically because prior institutional decisions have already specified what counts as relevant input, which categories exist, which threshold activates them, what sanctions follow, and how errors will be managed.

The disappearance of an individual reviewer is therefore not equivalent to the disappearance of a responsible institution.

This distinction can be stated formally:

automation ≠ authorlessness

and:

execution without immediate human intervention ≠ governance without institutional authority.

The compiled rule explains this relation. Authority migrates into the formal architecture of execution (Startari, 2025).

This has consequences beyond platform moderation. The same analytical model can potentially apply wherever institutional decisions are transformed into executable

constraints and subsequently communicated as procedural outcomes: automated eligibility systems, financial compliance, administrative decision systems, risk scoring, algorithmic ranking, automated fraud detection, and other forms of computational governance. The present article does not claim to have demonstrated the mechanism in those domains. It identifies the theoretical portability of the model.

Its contribution within platform governance is more specific.

The paper shifts the analytical focus from:

Which content was suppressed?

to:

How is suppression linguistically represented after the institution has already acted?

This shift reveals that moderation has an informational afterlife. Enforcement does not end when content is removed or distribution changes. The institution produces a representation of the enforcement event. That representation determines whether users encounter governance as an attributable decision or as an impersonal condition.

Two otherwise identical sanctions can therefore produce different accountability structures.

Compare:

We reduced the distribution of your post because our review found that it violated Rule X. You can appeal this decision.

with:

Your content violates our standards and is not eligible for recommendation.

Both may produce reduced visibility.

The first preserves a decision chain.

The second presents classification and restriction primarily as properties of the content.

The governance effect may be comparable.

The accountability surface is not.

That distinction is the central empirical object of SOI.

At the lowest level of suppression opacity, the institution remains visible, the action is named, the rule is specific, the classification is attributable, the decision mechanism is identifiable, and the user can contest the result.

At higher levels, components begin to disappear.

At the extreme, only two elements remain:

violation

and

consequence.

Everything connecting them becomes implicit.

The result can be represented as:

K → S

The rule appears complete.

Institutional history disappears.

This is the terminal form of compiled governance analyzed in this article.

It is also why the paper's title should not be read literally as the absence of a censoring institution. **Censorship without a censor describes the disappearance of the censor from the representation of censorship, not necessarily from the institutional production of the restrictive act.**

The distinction resolves the apparent paradox.

There can be suppression without a grammatically visible suppressor.

There can be classification without a grammatically visible classifier.

There can be executable authority without a continuously visible author.

There can be institutional power without institutional subject position.

The research program that follows from these propositions is straightforward. A sufficiently large corpus of authenticated moderation notices can determine whether censor deletion is common, whether classifier deletion precedes it, whether policy substitutes grammatically for institutional agency, whether visibility restrictions exhibit greater opacity than outright removals, whether automated and human enforcement differ in traceability, and whether politically contested content differs systematically from relevant controls.

The framework is therefore falsifiable.

If platforms generally preserve explicit institutional agency, specify rules and mechanisms, provide visible appeals, and show no meaningful political differences after controlling for platform and sanction type, the strongest form of the censorship-without-a-censor hypothesis will not be supported.

If, however, politically contested moderation repeatedly produces high CDR, high CLR, high VROR, high SOI, low APVR, and a substantial asymmetry between user-as-violator and platform-as-enforcer constructions, then the empirical pattern will require explanation.

In that case, the finding would not merely concern communication style.

It would indicate a recurring structure in which the visibility of governance is distributed unequally between the institution exercising authority and the subject upon whom authority acts.

That would constitute **responsibility moderation** in its strongest form.

The platform would not merely determine the visibility of political speech.

Its enforcement discourse would also determine the visibility of responsibility for determining that visibility.

The theoretical sequence of the article can therefore be stated in its final form:

institutional authority

→ compiled rule

→ classification

→ enforcement

→ visibility modification

→ grammatical compression

→ responsibility loss

→ suppression opacity

The final transition does not require the institution to disappear operationally.

It requires only that the output remain intelligible without explicitly naming it.

This is the structural threshold identified by censorship without a censor.

The speech can remain visible enough to be classified.

The speaker can remain visible enough to be sanctioned.

The violation can remain visible enough to justify enforcement.

The rule can remain visible enough to appear authoritative.

The sanction can remain visible enough to alter circulation.

Yet the institution connecting all five elements can become grammatically optional.

At that point, platform governance reaches a distinctive form of executable authority: **the censor has not ceased to govern; the rule has become capable of governing without requiring the censor to remain in the sentence** (Startari, 2025).

Content moderation determines what speech may remain visible.

Responsibility moderation determines whether the actor making that determination remains visible with it.

Censorship without a censor begins where those two forms of visibility separate.

Formatted: Justified, Line spacing: 1,5 lines

References

Access Now, ARTICLE 19, & Center for Human Rights in Iran. (2022, June 9). Iran: Meta must overhaul Persian-language content moderation on Instagram. Access Now. https://www.accessnow.org/press-release/iran-meta-instagram-persian-rightscon-statement/

ARTICLE 19. (2022, June 9). Iran: Meta must overhaul Persian-language content moderation on Instagram. https://www.article19.org/resources/iran-meta-persian-language-content-moderation-instagram/

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 1–15. https://doi.org/10.1177/2053951719897945

Human Rights Watch. (2023, December 21). Meta's broken promises: Systemic censorship of Palestine content on Instagram and Facebook. https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and

Instagram. (n.d.-a). Recommendations on Instagram [Help Center article]. Instagram Help Center. Retrieved August 10, 2026.

Instagram. (n.d.-b). Recommendation Guidelines [Help Center article]. Instagram Help Center. Retrieved August 10, 2026.

Instagram. (n.d.-c). Request a review of a content removal [Help Center article]. Instagram Help Center. Retrieved August 10, 2026.

Krippendorff, K. (2018). Content analysis: An introduction to its methodology (4th ed.). SAGE Publications.

Meta. (2024a). Taking action on content that violates our policies [Transparency Center article]. Meta Transparency Center. Retrieved August 10, 2026.

Meta. (2025a). Types of content we demote [Transparency Center article]. Meta Transparency Center. Retrieved August 10, 2026.

Meta. (2025b). Content Distribution Guidelines [Transparency Center article]. Meta Transparency Center. Retrieved August 10, 2026.

Formatted: Font: (Default) Times New Roman, 12 pt, Bold
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Hyperlink, Font: (Default) Times New Roman, 12 pt
Field Code Changed
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Hyperlink, Font: (Default) Times New Roman, 12 pt
Field Code Changed
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Hyperlink, Font: (Default) Times New Roman, 12 pt
Field Code Changed
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Hyperlink, Font: (Default) Times New Roman, 12 pt
Field Code Changed
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Hyperlink, Font: (Default) Times New Roman, 12 pt
Field Code Changed
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt
Formatted: Font: (Default) Times New Roman, 12 pt

Meta. (2026). How Meta enforces its policies: Remove, reduce, inform [Transparency Center article]. Meta Transparency Center. Retrieved August 10, 2026.

Oversight Board. (2024a, March 26). Referring to designated dangerous individuals as "shaheed" [Policy advisory opinion]. https://www.oversightboard.com/news/oversight-board-publishes-policy-advisory-opinion-on-referring-to-designated-dangerous-individuals-as-shaheed/

Oversight Board. (2024b, September 4). Posts that include "From the River to the Sea" (Case bundle BUN-86TJ0RK5) [Decision]. https://www.oversightboard.com/decision/bun-86tj0rk5/

Oversight Board. (2024c, April 4). Washington Post article on Israel-Palestine (Case FB-8RTRZY6Q) [Summary decision]. https://www.oversightboard.com/decision/fb-8rtrzy6q/

Oversight Board. (2024d). Reports on the war in Gaza (Case bundle BUN-JEXZGIY5) [Summary decisions]. https://www.oversightboard.com/decision/bun-jexzgiy5/

Roberts, S. T. (2019). Behind the screen: Content moderation in the shadows of social media. Yale University Press.

Startari, A. V. (2025). The codex of authority [Preprint]. SSRN. https://doi.org/10.2139/ssrn.5432334

Startari, A. V. (2026a). Suffering without perpetrators: The humanitarian passive in AI-generated conflict discourse. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6753123

Startari, A. V. (2026b). The grammar of asymmetric visibility: AI, Zionism, and the reallocation of political agency. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6787439

Startari, A. V. (2026c). *The grammar of asymmetric visibility: AI, Zionism, and the reallocation of political agency. AI Power and Discourse, 1*(1), 1–10.

TikTok. (n.d.-a). Content violations and bans [Support article]. TikTok Support. Retrieved August 10, 2026.

TikTok. (n.d.-b). Appeal an account ban [Support article]. TikTok Support. Retrieved August 10, 2026.

TikTok. (n.d.-c). Appeal a recommendation-ineligibility decision [Support article]. TikTok Support. Retrieved August 10, 2026.

YouTube. (n.d.-a). How does YouTube enforce its Community Guidelines? [Help Center article]. YouTube Help. Retrieved August 10, 2026.

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Hyperlink, Font: (Default) Times New Roman, 12 pt

Field Code Changed

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Hyperlink, Font: (Default) Times New Roman, 12 pt

Field Code Changed

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Hyperlink, Font: (Default) Times New Roman, 12 pt

Field Code Changed

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Hyperlink, Font: (Default) Times New Roman, 12 pt

Field Code Changed

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Hyperlink, Font: (Default) Times New Roman, 12 pt

Field Code Changed

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Hyperlink, Font: (Default) Times New Roman, 12 pt

Field Code Changed

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Hyperlink, Font: (Default) Times New Roman, 12 pt

Field Code Changed

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Font: (Default) Times New Roman, 12 pt

YouTube. (n.d.-b). Community Guidelines strike basics and appeals [Help Center article]. YouTube Help. Retrieved August 10, 2026,

Formatted: Font: (Default) Times New Roman, 12 pt

Formatted: Left, Indent: Left: 0 cm, Hanging: 1,27 cm, Space After: 6 pt, Line spacing: Multiple 1,15 li

Formatted: Font: (Default) Times New Roman, 12 pt