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

The Grammar of Asymmetric Visibility: AI, Zionism, and the Reallocation of Political Agency.

Agustin V. Startari.

Cita:

Agustin V. Startari (2026). *The Grammar of Asymmetric Visibility: AI, Zionism, and the Reallocation of Political Agency. AI Power and Discourse, 1 (1), 1-10.*

Dirección estable: https://www.aacademica.org/agustin.v.startari/232

ARK: https://n2t.net/ark:/13683/p0c2/X2y

The Grammar of Asymmetric Visibility: AI, Zionism, and the Reallocation of Political Agency

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: May 18, 2026

DOI

- Primary archive: https://doi.org/**10.5281/zenodo.20271438**
- Secondary archive: https://doi.org/**10.5281/zenodo.20271438**
- SSRN: Pending assignment (ETA: Q3 2026)

Language: English

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

Word count: 12630

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 **asymmetric visibility** as an AI-mediated condition in which subordinated political actors remain visible while their grammatical agency is weakened. Building on the concept of **responsibility loss** developed in the first article of the series, the paper argues that AI-generated and AI-mediated conflict discourse does not merely describe political violence, occupation, sanctions, resistance, or humanitarian suffering. It redistributes the grammatical conditions under which actors appear as agents, victims, risks, threats, humanitarian populations, or moderation objects.

The article focuses on Palestine, Zionism, anti-Zionism, Iran, United States foreign-policy discourse, Israel-linked institutional discourse, and platform moderation as test cases for analyzing agency asymmetry. Its central claim is not that AI systems intentionally favor one political position, nor that visibility itself disappears. Rather, the problem is formal: dominant-power actors may remain more frequently represented as institutional, defensive, diplomatic, or security-oriented subjects, while subordinated actors are more likely to appear through humanitarian, securitized, or moderation-oriented categories. A population may therefore be repeatedly mentioned, described, classified, and sympathized with while still being denied the grammatical status of a political subject.

Methodologically, the article proposes an **Agency-Preservation Rate** and an **Asymmetric Visibility Index** to measure the disparity between the agency preserved for dominant-power actors and the agency preserved for subordinated actors across AI-generated or AI-mediated conflict discourse. These indicators examine subject position, verbal agency, passive attribution, agent deletion, risk-object conversion, humanitarian framing, security framing, moderation framing, and political-subjecthood retention. The paper therefore shifts the evaluation of AI discourse from representation alone toward the measurable preservation or weakening of political agency.

The article concludes that visibility is not equality. In conflict discourse, the decisive question is not only whether AI systems mention Palestine, Iran, occupied populations, sanctioned societies, or anti-imperial speech, but how they grammatically position them. AI ethics and platform governance should therefore move beyond bias detection, hate

speech classification, and misinformation control, and ask whether machine-generated discourse preserves the grammar through which political agency remains recognizable. Asymmetric visibility begins where recognition ends: when the dominated are seen, but not grammatically allowed to act.

I. Introduction: Visibility Is Not Equality

Visibility is often treated as a corrective standard in debates about artificial intelligence, media representation, and political discourse. If a population is mentioned, counted, summarized, classified, or included in an output, it may appear to have entered the field of recognition. In conflict-related discourse, however, visibility does not necessarily preserve agency. A population can be highly visible as suffering, displacement, threat, instability, humanitarian emergency, or moderation concern while remaining weakly represented as a political subject capable of acting, deciding, resisting, governing, claiming rights, or being addressed as an agent within history (Fairclough, 1995; Hall, 1997; Startari, 2025a, 2025b).

This article begins from that distinction. The problem is not simply whether AI systems mention Palestine, Iran, occupied populations, sanctioned societies, anti-Zionist discourse, or anti-imperial speech. The problem is how these actors are grammatically positioned once they appear. Visibility can coexist with agency loss. Recognition can coexist with syntactic subordination. A group can be made present in discourse while its political subjecthood is displaced into categories of risk, victimhood, crisis, or control. This formal displacement extends the problem of algorithmic neutrality and agency disappearance previously identified in Startari's work on passive voice, objectivity, and non-human authority in AI-generated language (Startari, 2025a, 2025b, 2025c).

The central claim of this article is that AI-mediated political discourse does not merely represent conflict. It reallocates political agency. Dominant-power actors are more likely to remain grammatically organized as institutional, defensive, diplomatic, legal, strategic, or security-oriented subjects. Subordinated actors, by contrast, may appear more frequently as objects of humanitarian concern, moderation review, sanction regimes, counterterror classification, escalation warnings, or crisis management. This does not require proving intention, ideological conspiracy, or explicit model alignment with any single geopolitical actor. It can be examined as a formal effect of syntax, summarization, safety filtering, moderation categories, institutional source weighting, and machine-generated neutrality (Baker, Gabrielatos, & McEnery, 2013; Bhatia, 2005; Entman, 1993; Startari, 2025b).

The concept proposed here is **asymmetric visibility**. It names the AI-mediated condition in which dominated, occupied, sanctioned, or subordinated societies remain visible while their grammatical agency is weakened. They are not necessarily erased. They may appear constantly. But they appear under categories that limit their status as actors. They are visible as bodies, casualties, refugees, risks, users, militants, civilians, populations, claimants, or objects of policy. They are less consistently visible as agents who impose, resist, decide, organize, accuse, govern, negotiate, or demand accountability. In this sense, asymmetric visibility extends Startari's prior concept of responsibility loss: the problem is no longer only that suffering can appear without a responsible actor, but that political actors can appear without equal grammatical access to agency (Startari, 2025a; Startari, 2026).

This distinction matters because contemporary AI ethics often treats representation, harm reduction, hate speech classification, and misinformation control as primary standards of evaluation. Those standards are necessary but insufficient. A system can avoid explicit hate speech and still reproduce agency asymmetry. It can summarize suffering accurately and still weaken responsibility. It can classify dangerous language correctly in some cases and still convert political grievance into security risk in others. It can mention victims and still obscure perpetrators. It can preserve formal neutrality while redistributing the grammar of power. Prior work on framing, securitization, and discourse analysis shows that political meaning is shaped not only by what is stated, but by how actors are positioned within available categories of threat, legitimacy, causality, and responsibility (Buzan, Wæver, & de Wilde, 1998; Entman, 1993; Fairclough, 1995; van Dijk, 1998).

The article therefore shifts the analytical question from visibility to agency preservation. The relevant question is not only "Who appears?" but "Who appears as capable of action?" More precisely, when AI systems summarize, classify, moderate, or explain conflict-related discourse, do they preserve political agency symmetrically across dominant-power and subordinated actors, or do they convert dominant actors into subjects of action while rendering subordinated actors as objects of risk, suffering, or control? This question follows directly from the paper's working premise: AI-mediated discourse can make both sides visible while preserving agency only for one side. That is the formal structure named here as asymmetric visibility.

This paper develops that question through the cases of Palestine, Zionism, anti-Zionism, Iran, Israel-linked institutional discourse, United States foreign-policy language, and platform moderation. These cases are not selected to reduce the analysis to denunciation or partisan judgment. They are selected because they expose the formal problem with unusual clarity: the same discourse field can contain military action, occupation, sanctions, resistance, civilian harm, security claims, anti-imperial critique, moderation pressure, and competing definitions of legitimacy. In such contexts, grammatical agency is not a stylistic detail. It determines who can be represented as acting within history and who is rendered as an object within someone else's frame (Bhatia, 2005; Buzan et al., 1998; Startari, 2025c).

The argument proceeds from the first article in this series, which introduced the humanitarian passive and responsibility loss. That earlier framework showed how AI-generated conflict discourse can describe suffering while weakening the attribution of responsibility (Startari, 2026). This article extends the argument. Responsibility loss is not an isolated syntactic defect. It is part of a broader distributional condition: asymmetric visibility. Where responsibility loss asks whether suffering can appear without a responsible actor, asymmetric visibility asks whether all actors remain equally capable of appearing as political subjects.

The proposed methodological response is to measure agency preservation directly. The article introduces **Agency-Preservation Rate** as a clause-level indicator of whether a political actor remains grammatically represented as capable of action, decision, imposition, resistance, governance, moderation, or responsibility. It also proposes the **Asymmetric Visibility Index**, which measures the disparity between the agency preserved for dominant-power actors and the agency preserved for subordinated actors across AI-generated or AI-mediated conflict discourse. These instruments are designed to move the analysis of AI discourse beyond general claims about bias and toward measurable grammatical asymmetry.

The paper's public thesis can be stated directly: **the powerful act, the dominated appear**. Its theoretical claim is narrower and measurable: visibility is not equality when the distribution of grammatical agency remains unequal.

II. From Responsibility Loss to Asymmetric Visibility

The first article in this series introduced **responsibility loss** as a formal condition in which suffering remains visible while the grammatical pathways to responsible actors are weakened, displaced, or removed (Startari, 2026). In that model, AI-generated conflict discourse can describe civilian harm, destruction, displacement, famine, siege, bombardment, or humanitarian catastrophe without preserving the same degree of agentive attribution for the actors who produced or sustained those conditions. The result is not silence. The result is a syntactic redistribution of responsibility. Suffering appears, but responsibility becomes optional.

This article extends that argument. Responsibility loss is not only a local feature of passive constructions, agent deletion, or humanitarian abstraction. It is one mechanism within a broader distributional structure: **asymmetric visibility**. If responsibility loss explains how harm can be described without a responsible agent, asymmetric visibility explains how different political actors can be made visible under unequal grammatical conditions. Some actors appear as institutional subjects capable of action, decision, defense, diplomacy, sanction, governance, and security reasoning. Others appear as populations, risks, casualties, threats, humanitarian cases, moderation concerns, or objects of intervention.

The distinction is necessary because visibility alone is too weak as an analytic standard. A dominated or sanctioned population may appear frequently in AI-generated summaries and still lose political agency at the level of clause structure. It may be visible as a victim, a crisis, a demographic mass, a humanitarian object, or a risk category, but not as an actor with claims, decisions, strategy, sovereignty, grievance, or historical causality. This is the point at which representation becomes insufficient. The question is not whether a population appears. The question is whether it appears as a subject of action or as an object within the action of others.

In discourse analysis, this problem has a longer genealogy. Critical discourse studies have shown that grammatical choices participate in the distribution of social and political meaning, especially through transitivity, nominalization, passivization, and the assignment or removal of agency (Fairclough, 1995; Halliday & Matthiessen, 2014; van Dijk, 1998).

Framing theory likewise demonstrates that political communication selects some aspects of reality and makes them more salient, thereby shaping causal interpretation, moral evaluation, and possible remedies (Entman, 1993). Securitization theory adds that political actors and institutions can convert subjects into security objects by placing them within threat language and emergency frames (Buzan et al., 1998). The present article transfers these insights into the analysis of AI-mediated conflict discourse and narrows the focus to clause-level agency preservation.

The move from responsibility loss to asymmetric visibility is therefore a move from **attribution** to **distribution**. Responsibility loss asks whether a harmful event remains grammatically connected to an accountable actor. Asymmetric visibility asks whether political agency is preserved equally across actor categories. A sentence can mention both a dominant-power actor and a subordinated actor, but distribute agency unevenly between them. For example, a dominant actor may "respond," "defend," "warn," "negotiate," "secure," or "authorize," while the subordinated actor "is affected," "is displaced," "is accused," "is flagged," "is associated with risk," or "is subject to sanctions." The actors are not equally erased. They are differently positioned.

This distinction allows the paper to avoid two weak formulations. The first weak formulation would be to claim that subordinated actors are simply invisible. That is often false. Palestine, Iran, sanctioned societies, displaced civilians, anti-Zionist discourse, and anti-imperial claims may be highly visible in AI-mediated discourse. The second weak formulation would be to claim that visibility proves discursive equality. That is also false. Visibility can intensify objectification when it occurs through humanitarian, securitized, or moderation-oriented categories. Asymmetric visibility names this second condition: the dominated are present, but their political agency is grammatically reduced.

The concept also prevents the analysis from collapsing into intent-based accusation. The paper does not need to show that a model designer, platform, institution, or state actor intended to remove agency from subordinated populations. The claim is formal and measurable. If AI outputs systematically preserve stronger subject position, active verbs, institutional predicates, and responsibility control for dominant-power actors, while assigning subordinated actors to passive, object, risk, humanitarian, or moderation

categories, then asymmetric visibility can be observed without proving intention. This follows Startari's broader argument that AI-generated authority effects can operate through form rather than explicit ideological assertion (Startari, 2025a, 2025b, 2025c).

This is where the concept differs from ordinary bias detection. Bias analysis often asks whether a system treats groups unfairly, reproduces stereotypes, or generates harmful associations. Those questions remain relevant, but they do not exhaust the problem. An AI system may avoid explicit derogatory content and still produce agency asymmetry. It may use neutral vocabulary and still assign action unevenly. It may comply with safety standards and still convert political grievance into risk language. It may summarize both sides and still preserve political subjecthood for one side more than the other. Asymmetric visibility therefore requires a different unit of analysis: not sentiment alone, not toxicity alone, not factuality alone, but grammatical agency.

The article's core methodological proposition follows from this shift. To evaluate AI-mediated conflict discourse, it is necessary to measure whether political actors retain clause-level agency. This requires identifying who appears as the grammatical subject, what verb is assigned to that subject, whether the actor is active or passive, whether responsibility is attributed or displaced, whether harm is nominalized, whether violence is bureaucratized, whether resistance is securitized, and whether political claims are converted into moderation concerns. The **Agency-Preservation Rate** proposed in this article operationalizes that question. It measures the proportion of conflict-related clauses in which a political actor remains represented as capable of acting, deciding, imposing, resisting, governing, moderating, sanctioning, occupying, attacking, negotiating, or being held responsible.

The **Asymmetric Visibility Index** then extends that measurement across actor categories. It asks whether dominant-power actors and subordinated actors receive comparable levels of grammatical agency. The index increases when dominant actors appear as active subjects while subordinated actors appear as passive victims, threat categories, humanitarian populations, moderation risks, or objects of policy. It also increases when responsibility for dominant-power violence is diluted while subordinated resistance is

intensified through security or extremism frames. The index therefore converts the paper's theoretical claim into a measurable research object.

This framework directly develops the series logic. Paper 1 established the **humanitarian passive** and **responsibility loss** as mechanisms through which suffering can be made visible while responsibility becomes grammatically weakened (Startari, 2026). Paper 2 expands the unit of analysis from the sentence of suffering to the broader field of political subjecthood. The question is no longer only whether harm appears without a perpetrator. The question is whether AI systems preserve the grammatical conditions under which subordinated actors can appear as political agents at all.

The theoretical progression can be stated in one line: **responsibility loss is the mechanism; asymmetric visibility is the political condition**. Responsibility loss explains how accountability disappears inside apparently neutral humanitarian description. Asymmetric visibility explains how this disappearance is distributed unevenly across dominant and subordinated actors. Together, they make it possible to analyze AI-mediated conflict discourse without reducing the problem to censorship, misinformation, or individual bias.

The relevance of this distinction becomes especially clear in Palestine-related and Iran-related discourse. In these domains, AI outputs may acknowledge suffering, sanctions, civilian harm, military escalation, occupation, regional instability, or moderation disputes. Yet acknowledgment does not guarantee agency preservation. Palestinian, Iranian, or anti-imperial actors may be visible as affected populations, escalation sources, security risks, or moderation problems, while dominant-power actors retain the grammar of institutional decision, defense, diplomacy, and strategic necessity. The result is not simple erasure. It is unequal grammatical access to political action.

This article therefore takes the next step in the series. It moves from the visibility of suffering to the grammar of agency. It argues that AI ethics and political discourse analysis must ask not only whether conflict is represented, but whether the actors within conflict remain grammatically capable of action. The central problem is not absence. It is

asymmetrical presence. A dominated population can be seen, named, counted, classified, and summarized, while still being denied equal access to the syntax of political agency.

III. AI, Zionism, and the Grammar of Political Legitimacy

Zionism presents a central test case for asymmetric visibility because it operates simultaneously as a political ideology, a historical national movement, a state-linked discourse, and a contested object of global criticism. Any analysis of AI-mediated language around Zionism must therefore separate four categories that are often collapsed in moderation environments: Judaism as religion and collective identity, antisemitism as hostility toward Jews, Zionism as a political project, and anti-Zionism as criticism or rejection of that project. The distinction is not optional. Without it, AI systems can misclassify political critique as identity-based hostility, or conversely fail to identify real antisemitic content when it is embedded inside geopolitical language (Bunzl, 2007; Klug, 2003; Ullrich, 2019).

The purpose of this section is not to define Zionism normatively, nor to adjudicate the legitimacy of every political claim made in its name or against it. The narrower question is grammatical: how do AI systems preserve or weaken agency when Zionism, Israel-linked institutional discourse, Palestine, anti-Zionism, and anti-imperial speech appear in conflict-related outputs? If dominant-power actors are more frequently represented through institutional, defensive, legal, or diplomatic predicates, while subordinated actors are more frequently represented through risk, escalation, radicalization, humanitarian need, or moderation concern, then visibility does not produce symmetry. It produces an unequal grammar of legitimacy.

Political legitimacy is often constructed not only through explicit justification, but through ordinary grammatical positioning. Actors who “respond,” “defend,” “authorize,” “negotiate,” “secure,” “warn,” “investigate,” or “conduct operations” appear within a grammar of decision and institutional continuity. Actors who “are displaced,” “are caught in violence,” “are linked to unrest,” “are accused of extremism,” “face restrictions,” or “are subject to moderation” appear within a grammar of exposure, suspicion, or administrative

control. This distinction follows the broader insight of critical discourse analysis that agency is distributed through transitivity, nominalization, passivization, and the organization of participants within clauses (Fairclough, 1995; Halliday & Matthiessen, 2014; van Dijk, 1998).

In AI-generated conflict discourse, this distribution becomes especially important because model outputs often present themselves as neutral summaries. Machine-generated neutrality can conceal the fact that syntactic form still assigns roles. A summary can avoid overtly inflammatory language while still preserving stronger agency for some actors than others. It can balance surface mention while producing deep grammatical asymmetry. This is consistent with Startari's prior argument that AI-generated language can simulate neutrality through formal mechanisms that conceal the disappearance or weakening of agency (Startari, 2025a, 2025b). In the present paper, the same problem is extended to Zionism and conflict discourse: the issue is not whether the topic appears, but whether the political grammar assigns agency symmetrically.

The grammar of political legitimacy often depends on the stabilization of institutional subjecthood. State-linked actors tend to appear as named institutions, offices, governments, militaries, courts, diplomats, or security agencies. This gives them a durable syntactic position from which they can act. They can "announce," "reject," "approve," "defend," "strike," "investigate," "coordinate," or "respond." By contrast, subordinated actors are often collectivized or abstracted. They appear as "civilians," "protesters," "militants," "refugees," "users," "sympathizers," "networks," "groups," or "populations." These labels may be descriptively valid in specific contexts, but they can also reduce political subjecthood when they replace agency with classification.

This is where Zionism becomes analytically important. In many AI-mediated summaries, Zionism may be treated as part of the vocabulary of state legitimacy, historical security, Jewish self-determination, or institutional defense. Anti-Zionism, by contrast, may be filtered through categories of extremism risk, antisemitism proximity, hate speech concern, escalation, or platform safety. The problem is not that antisemitism should be ignored. It should not. The problem is that political critique and identity hatred require distinct classification pathways. If AI systems collapse anti-Zionism into antisemitism without

sufficient contextual analysis, they may convert political disagreement into moderation risk. If they ignore antisemitic content because it is framed as geopolitical critique, they fail in the opposite direction. The analytical standard must therefore be distinction, not substitution (Klug, 2003; Ullrich, 2019).

Asymmetric visibility becomes visible when the same grammatical operation is not applied equally. For example, violence by dominant-power actors may be described through institutional or strategic abstractions: "operations," "security measures," "responses," "campaigns," "strikes," or "policy decisions." These terms preserve the actor's institutional status while sometimes diluting the human consequences of action. Violence or resistance associated with subordinated actors may be described through intensified risk categories: "terrorism," "extremism," "instability," "incitement," "radicalization," or "regional threat." Some of these labels may be factually justified in particular cases. The issue is not the existence of risk language. The issue is whether risk language is distributed asymmetrically and whether it replaces political causality with categorical suspicion.

This problem is not unique to AI. Media and political discourse have long used framing devices that shape how conflicts are understood by foregrounding some causes, actors, and remedies while backgrounding others (Entman, 1993). Securitization theory shows that political language can transform an issue into a security object, thereby authorizing exceptional measures and narrowing the field of legitimate speech (Buzan et al., 1998). What AI adds is scale, automation, and the appearance of procedural neutrality. When models summarize, classify, moderate, or rewrite conflict discourse, they can reproduce inherited frames without openly declaring them. The output appears neutral because it is balanced in tone, but its clause structure may still decide who acts and who is acted upon.

The distinction between agency and visibility also clarifies why "both sides are mentioned" is not a sufficient answer. A model can mention Israel, Zionism, Palestine, Iran, sanctions, occupation, anti-Zionism, civilian harm, and regional security in the same output while still preserving stronger agency for state-linked or dominant-power actors. Mention is a surface condition. Agency is a structural condition. The question is whether each actor receives comparable access to subject position, active verbs, causal attribution, institutional

predicates, and responsibility-bearing clauses. Without this measurement, representational balance can become a mask for syntactic imbalance.

The grammar of legitimacy also operates through euphemism and bureaucratic abstraction. Terms such as "security operation," "targeted response," "collateral damage," "administrative restriction," "border control," or "content enforcement" can move political violence or coercion into procedural language. Again, the issue is not that every use of such terms is false. The issue is that bureaucratic language can reduce perceived agency and responsibility by transforming action into process. This aligns with the first paper's account of responsibility loss, where harm remains describable while the agentive pathway to responsibility weakens (Startari, 2026). In this second paper, the same mechanism is situated inside a broader distribution: whose violence is bureaucratized, whose resistance is securitized, and whose suffering is humanitarized.

AI moderation intensifies the problem because platform categories are not neutral linguistic containers. Categories such as hate speech, extremism, terrorism, incitement, graphic violence, misinformation, or dangerous organizations may be necessary for governance, but they also determine how speech becomes legible to automated systems. Political claims can be converted into compliance objects. A statement about occupation, sanctions, siege, displacement, or resistance may be evaluated less as a political claim than as a safety risk, depending on keywords, entity associations, geopolitical context, and platform policy constraints. The result is not necessarily censorship in every case. It is a shift from political interpretation to risk classification.

For Palestine-related and Iran-related discourse, this shift matters because both domains are already embedded in dense security vocabularies. Palestine-related speech is frequently entangled with terrorism, antisemitism, protest, humanitarian crisis, occupation, and platform moderation. Iran-related speech is frequently entangled with nuclear risk, sanctions, regional destabilization, terrorism designations, proxy language, and United States foreign-policy framing. These associations may derive from real geopolitical conflicts, but AI systems can still reproduce them in ways that weaken the political subjecthood of subordinated or sanctioned actors. The analytical question is whether the

system can distinguish between a political actor, a civilian population, a state apparatus, a military organization, an ideological claim, and a moderation risk.

This distinction is central to the proposed **Agency-Preservation Rate**. In the context of Zionism and anti-Zionism, APR does not ask whether a system approves or rejects a political position. It asks whether actors are grammatically preserved as agents. Does the output allow Palestinians to "claim," "accuse," "organize," "govern," "resist," "negotiate," or "demand accountability," or does it mainly render them as "affected," "displaced," "killed," "caught," "radicalized," or "subject to aid"? Does the output allow Iranian actors to "argue," "respond," "govern," "contest sanctions," or "assert sovereignty," or does it mainly render them as "threatening," "destabilizing," "sanctioned," "isolated," or "monitored"? Does the output allow Israel-linked institutional actors and United States foreign-policy actors to remain active subjects of decision while their coercive actions are softened through abstraction? These are measurable grammatical questions, not moral impressions.

The **Asymmetric Visibility Index** extends this measurement by comparing actor categories across outputs. It increases when dominant-power actors retain active institutional agency and subordinated actors are more frequently routed into passive, humanitarian, threat, or moderation frames. It decreases when all actors retain comparable access to agency-bearing clauses, causal attribution, and responsibility-bearing predicates. The index therefore does not require a predetermined political conclusion. It requires clause-level comparison. This is what separates the proposed framework from rhetorical accusation: the unit of analysis is the distribution of grammatical agency.

The risk of this section is conceptual collapse. Zionism, Judaism, Israel, Israeli state policy, Jewish identity, antisemitism, anti-Zionism, Palestine, Hamas, Palestinian civilians, Iran, Iranian state institutions, Iranian civilians, United States foreign policy, and platform moderation are not interchangeable categories. A credible analysis must keep them separate. It must also recognize that antisemitism exists as a real form of hostility and cannot be dismissed as merely a moderation artifact. At the same time, criticism of Zionism, Israeli state policy, occupation, sanctions, military action, or United States foreign policy cannot be automatically reduced to antisemitism without destroying the

distinction between identity hatred and political critique. That distinction is necessary for both analytical accuracy and platform governance (Klug, 2003; Ullrich, 2019).

The claim of this paper is therefore controlled. It does not argue that AI systems are simply "pro-Israel," "anti-Palestinian," "anti-Iranian," or intentionally aligned with dominant geopolitical power. Such claims would require separate empirical proof and would weaken the formal precision of the argument. The claim is that AI-mediated conflict discourse can preserve political agency unevenly. Zionism and anti-Zionism provide a high-risk test case because they force the model to navigate identity protection, state legitimacy, political critique, historical violence, security discourse, and moderation policy at the same time. In that environment, grammatical agency becomes a measurable site of power.

The section's conclusion is therefore direct: political legitimacy in AI discourse is not produced only by explicit endorsement. It is also produced by grammar. Actors who retain subject position, active verbs, institutional predicates, and causal control retain political agency. Actors who appear mainly as risks, victims, populations, crises, or moderation objects remain visible but weakened. In the field of Zionism, anti-Zionism, Palestine, Iran, and dominant-power discourse, this difference is not stylistic. It is the syntactic infrastructure through which legitimacy is distributed.

IV. Resistance as Risk: The Securitization of Subordinated Speech

Resistance becomes politically fragile when it is processed primarily through the grammar of risk. In AI-mediated conflict discourse, subordinated actors may remain visible, but their speech, protest, refusal, mourning, accusation, or demand for sovereignty can be routed through categories such as escalation, extremism, incitement, instability, terrorism adjacency, dangerous organization support, or moderation concern. The problem is not that risk categories are always false. The problem is that the conversion of subordinated political speech into security language can weaken political agency before any substantive interpretation occurs. In such cases, the actor is not first read as a political subject. The actor is first read as a potential risk object.

This section develops the second major mechanism of asymmetric visibility: risk-object conversion. Risk-object conversion occurs when the discourse of a subordinated, occupied, sanctioned, or dominated actor is not primarily represented as political claim-making, but as a security, moderation, or escalation problem. The actor remains present. The claim may even be quoted or summarized. Yet the surrounding grammar shifts from political agency to risk management. Instead of "Palestinians accuse," "Iranian officials contest," "anti-imperial activists argue," or "civil society organizations demand," the output may produce forms such as "content related to the conflict may violate policy," "statements risk inflaming tensions," "claims are associated with extremist narratives," or "speech around the issue requires careful moderation." The speech remains visible, but its political status is weakened.

This mechanism extends the logic of securitization theory. Securitization does not merely describe danger. It constructs an issue as a security object, thereby changing the range of acceptable responses (Buzan et al., 1998). Once an actor or claim is placed inside a security frame, ordinary political disagreement can be reorganized as threat management. In platform and AI contexts, the same transformation can occur through safety language, classifier labels, moderation categories, and automated summarization. The grammar of risk does not need to declare a political position. It can produce the same effect through classification. A claim becomes legible as danger before it becomes legible as politics.

For Palestine-related discourse, this problem is especially visible because political claims about occupation, displacement, siege, civilian harm, or state violence often coexist with platform concerns about antisemitism, extremism, terrorism, incitement, and graphic content. These categories are not analytically identical. Antisemitism refers to hostility toward Jews as Jews. Anti-Zionism refers to opposition to Zionism as a political project or state-linked ideology. Criticism of Israeli state policy refers to political criticism of governmental, military, legal, or territorial action. Support for armed organizations, endorsement of violence, and incitement to harm require separate classification. If AI-mediated moderation collapses these categories, it may transform political critique into risk language without preserving the agency of the speaker or the specificity of the claim (Klug, 2003; Ullrich, 2019).

For Iran-related discourse, the same structure appears through sanctions, nuclear risk, proxy language, terrorism designations, regional destabilization, and United States foreign-policy framing. Iranian actors may appear as "threatening," "destabilizing," "sanctioned," "isolated," "monitored," or "linked to escalation," while dominant-power actors appear as "responding," "pressuring," "deterring," "negotiating," or "seeking security guarantees." The asymmetry is not simply lexical. It is grammatical. One actor is positioned as a subject managing danger. The other is positioned as the danger to be managed. This is precisely the field in which asymmetric visibility becomes measurable.

Risk-object conversion also depends on nominalization. Political actions can be transformed into abstract nouns that remove agency from subordinated actors while intensifying suspicion around them. "Resistance" becomes "unrest." "Mobilization" becomes "instability." "Grievance" becomes "radicalization." "Sovereignty claim" becomes "regional tension." "Speech against occupation" becomes "content related to extremism risk." Nominalization has long been identified as a central mechanism through which discourse can compress action, conceal actors, and convert processes into abstract entities (Fairclough, 1995; Halliday & Matthiessen, 2014). In AI-mediated discourse, nominalization can be amplified by the model's tendency to produce administrative, balanced, and safety-oriented phrasing.

This is why the distinction between political speech and moderation object is central. A political claim attributes responsibility, asserts causality, demands recognition, and locates the speaker within a field of agency. A moderation object is processed according to policy categories, risk thresholds, and safety constraints. When AI systems convert the former into the latter too quickly, the discourse shifts from interpretation to containment. The speaker is no longer primarily an actor making a claim. The speaker becomes a user producing potentially risky content. The shift is formal, but its political effect is substantial.

The mechanism does not require explicit censorship. A model may answer the query, summarize the claim, and mention the affected population. Yet it may surround subordinated speech with cautionary language while allowing dominant institutional speech to remain politically legible. For example, a dominant actor may "warn of security threats," while a subordinated actor's warning may be "inflammatory rhetoric." A

dominant actor may “seek deterrence,” while a subordinated actor may “escalate tensions.” A dominant actor may “conduct operations,” while a subordinated actor may be “linked to violence.” Such distinctions cannot be evaluated only by sentiment analysis or topic detection. They require clause-level agency analysis.

This is also where the grammar of neutrality becomes unstable. Neutral tone can coexist with asymmetric classification. A system can use careful, moderate, non-inflammatory language and still assign different political statuses to different actors. This follows Startari’s broader argument that AI-generated neutrality can operate as a formal effect, not as proof of epistemic symmetry (Startari, 2025a, 2025b). In the present paper, neutrality is not rejected as a goal. It is treated as insufficient. Neutrality must be tested against agency preservation. If the dominant actor retains the grammar of institutional action while the subordinated actor is converted into risk, neutrality is surface-level.

The securitization of subordinated speech also affects temporality. Dominant-power action is often represented as reactive, defensive, or time-bound: “after attacks,” “in response to threats,” “following security concerns,” “amid regional tensions.” Subordinated action is often represented as originating instability: “protests erupted,” “violence spread,” “militants launched,” “rhetoric escalated,” “content circulated.” The difference is not trivial. Temporal framing determines whether an actor appears as responding to prior conditions or producing disruption without historical cause. Framing theory shows that causality and remedy are central to how political events are interpreted (Entman, 1993). In AI-mediated conflict discourse, temporal compression can remove occupation, sanctions, blockade, displacement, or prior violence from the causal field.

This creates a second form of responsibility loss. In the first article of the series, responsibility loss described the weakening of the path from suffering to responsible actor (Startari, 2026). In the present article, risk-object conversion weakens the path from resistance to political cause. The actor may still be visible, but the historical reasons for action are reduced or displaced. Resistance appears as escalation. Protest appears as unrest. Refusal appears as radicalization. Anti-imperial critique appears as destabilizing rhetoric. The grammar does not need to deny the actor’s existence. It only needs to detach the actor’s speech from a recognizable political cause.

This is especially relevant to platform moderation. Moderation systems must manage real harms, including threats, identity-based hate, harassment, incitement, glorification of violence, and coordinated manipulation. The existence of those harms is not disputed. The problem arises when moderation categories become the dominant interpretive frame for subordinated political speech. In such cases, AI-mediated systems may classify before they interpret. They may detect entity associations, conflict keywords, emotionally charged language, or contested ideological terms and route the output toward caution. That caution may be justified in some cases. But when it appears disproportionately around subordinated actors, it becomes part of asymmetric visibility.

A rigorous framework must therefore avoid two errors. The first error is to treat every moderation intervention as illegitimate suppression. That is analytically false. Some speech does incite harm, dehumanize, or target protected groups. The second error is to treat every moderation intervention as neutral safety enforcement. That is also false. Moderation systems operate through categories, thresholds, and institutional priorities that can redistribute political legibility. The correct object of analysis is not whether moderation exists, but whether moderation preserves the distinction between political grievance, identity hatred, violent incitement, historical analysis, and legitimate criticism.

The proposed Agency-Preservation Rate captures this distinction by coding whether subordinated actors remain subjects of political verbs. In Palestine-related discourse, the relevant question is whether Palestinians are represented as "claiming," "organizing," "governing," "accusing," "mourning," "resisting," "documenting," "negotiating," or "demanding accountability," or whether they are primarily represented as "affected," "displaced," "radicalized," "flagged," "screened," or "associated with unrest." In Iran-related discourse, the question is whether Iranian actors are represented as "arguing," "responding," "contesting sanctions," "asserting sovereignty," "negotiating," or "governing," or whether they are primarily represented as "threatening," "destabilizing," "violating," "evading," "being sanctioned," or "being monitored." These are measurable grammatical differences.

The proposed Asymmetric Visibility Index then compares the distribution of those differences across actor categories. The index rises when subordinated actors are more

frequently converted into risk objects and dominant actors are more frequently preserved as institutional subjects. It also rises when subordinated political speech is classified through safety language more often than dominant-power coercion is classified through responsibility-bearing language. In this sense, the index does not measure ideological intent. It measures the formal allocation of political agency.

The risk language surrounding anti-Zionism is a central test for this model. A credible AI system must identify antisemitic speech when it targets Jews as Jews, invokes antisemitic stereotypes, denies Jewish collective safety, or endorses harm. At the same time, it must not automatically convert criticism of Zionism, Israeli state policy, military action, settlement expansion, occupation, siege, apartheid claims, sanctions policy, or United States backing into antisemitism or extremism. The analytical challenge is not solved by choosing one category in advance. It is solved by preserving category boundaries and measuring whether the system retains political agency for the speaker while enforcing legitimate protections against hate and incitement (Klug, 2003; Ullrich, 2019).

The same applies to anti-imperial discourse more broadly. Speech against sanctions, military intervention, occupation, regime change, foreign bases, proxy warfare, or diplomatic coercion may be intense, accusatory, or morally charged. That intensity does not automatically make it extremist. Political language often contains accusation because politics contains responsibility claims. If AI systems systematically soften dominant-power responsibility while treating subordinated accusation as escalation, they alter the grammar of political conflict. They do not merely moderate tone. They redistribute the conditions under which responsibility can be named.

This section therefore identifies risk-object conversion as a core mechanism of asymmetric visibility. The dominated actor is not simply erased. It appears, but under a grammar of caution, suspicion, containment, or humanitarian management. Its suffering may be acknowledged. Its danger may be emphasized. Its speech may be moderated. Its claims may be summarized. But its agency is weakened if it cannot consistently appear as a political subject with reasons, demands, history, and responsibility claims. This is the central structural point: subordinated speech becomes less politically legible when it is first processed as risk.

The implication for AI ethics is direct. Bias detection is not enough. Hate speech classification is not enough. Misinformation control is not enough. Safety systems must also be audited for agency preservation. The question is whether AI-mediated discourse can distinguish danger from dissent, hatred from critique, incitement from accusation, and political grievance from moderation risk. Without that distinction, subordinated actors may remain visible only as objects of control.

The conclusion of this section can be stated in the terms of the series: responsibility loss hides who caused suffering; risk-object conversion weakens why resistance appears. Together, they form the syntactic core of asymmetric visibility. Dominant-power actors retain institutional agency. Subordinated actors remain visible as suffering, danger, instability, or content risk. The result is not silence. It is unequal political grammar.

V. Measuring Agency Preservation

The central methodological problem of this paper is that visibility cannot be measured only by frequency of mention. A political actor may appear often in AI-generated discourse and still lose agency at the level of grammar. Frequency counts can show whether an actor is present, but they cannot show whether that actor is represented as capable of acting, deciding, governing, resisting, accusing, negotiating, imposing, sanctioning, occupying, attacking, moderating, or being held responsible. For this reason, the paper shifts the unit of analysis from actor visibility to **agency preservation**.

Agency preservation refers to the degree to which a political actor remains grammatically represented as an agent within a clause. This requires examining subject position, verb type, voice, object position, attribution, nominalization, risk framing, humanitarian framing, moderation framing, and causal structure. The method follows the broader premise of critical discourse analysis that grammar is not a neutral container of meaning, but a formal system through which social actors are positioned, foregrounded, backgrounded, activated, or passivated (Fairclough, 1995; Halliday & Matthiessen, 2014; van Leeuwen, 2008). In AI-mediated conflict discourse, this positioning becomes

measurable because model outputs can be segmented, coded, compared, and tested across controlled prompts.

The first proposed metric is the **Agency-Preservation Rate**. APR measures the proportion of conflict-related clauses in which a political actor remains grammatically represented as an agent capable of action, decision, imposition, resistance, governance, moderation, sanction, occupation, attack, negotiation, or responsibility. The metric does not ask whether the actor is morally justified. It asks whether the actor remains grammatically capable of acting. This distinction is essential. A system may criticize, condemn, or contextualize an actor while still preserving agency. Conversely, a system may mention an actor sympathetically while grammatically reducing that actor to a victim, risk, population, or object of policy.

For coding purposes, a clause counts as agency-preserving when the actor appears as the grammatical subject of an action-bearing predicate or is otherwise clearly represented as the source of decision, causality, responsibility, institutional action, political claim, resistance, governance, or coercive effect. Examples include clauses in which an actor "imposes sanctions," "orders an attack," "contests occupation," "demands accountability," "negotiates a ceasefire," "documents civilian harm," "governs a territory," "moderates content," "authorizes force," or "claims sovereignty." These clauses preserve the actor as a political subject because the actor remains attached to action.

A clause counts as agency-weakening when the actor appears mainly as an object of action, a passive recipient, a humanitarian category, a security threat, a moderation concern, a demographic mass, or an abstracted crisis marker. Examples include formulations in which "civilians were displaced," "content was flagged," "tensions escalated," "unrest spread," "the region faced instability," "claims were associated with extremism," or "a population was affected." These clauses may be factually valid in specific contexts, but they weaken agency when they remove or obscure who acts, who decides, who imposes, who resists, or who is responsible. This connects directly to Startari's prior work on passive voice, objectivity, and responsibility loss in AI-generated language (Startari, 2025a, 2025b, 2026).

APR can be expressed conceptually as follows:

Agency-Preservation Rate = agency-preserving clauses for actor category / total conflict-related clauses mentioning that actor category.

The formula is intentionally simple because the object is not mathematical abstraction for its own sake. The metric exists to force the analyst to ask a precise question: when a political actor appears, how often does that actor remain grammatically capable of action? A high APR indicates that an actor is frequently preserved as a subject of political agency. A low APR indicates that an actor is frequently visible without equivalent agency.

The second metric is the **Dominant-Power Agency Ratio**. This measures the agency-preserving clauses assigned to dominant-power actors, including state institutions, allied military actors, major geopolitical powers, platform authorities, or recognized institutional actors. In the empirical domain of this paper, this may include Israel-linked institutional discourse, United States foreign-policy actors, platform governance actors, or international institutions. The metric tracks whether these actors appear as subjects of defense, diplomacy, security, law, moderation, authorization, investigation, or strategic decision.

The third metric is the **Subordinated-Agency Ratio**. This measures the agency-preserving clauses assigned to subordinated, occupied, sanctioned, displaced, or dominated actors. In the empirical domain of this paper, this may include Palestinian actors, Iranian actors, anti-imperial speakers, sanctioned populations, occupied populations, civil society actors, or anti-Zionist political speech when it is not reducible to identity hatred or incitement. The metric tracks whether these actors appear as subjects of claim-making, resistance, governance, documentation, accusation, negotiation, mourning, organization, or demands for accountability.

The fourth metric is the **Risk-Object Conversion Rate**. This measures how often a subordinated actor or subordinated speech act is converted from political subjecthood into security, moderation, escalation, extremism, or instability language. A clause contributes to this rate when the actor's political claim is framed primarily as risk rather than as agency. Examples include formulations that convert protest into unrest, grievance into radicalization, resistance into threat, anti-imperial critique into destabilizing rhetoric, or

political accusation into moderation concern. This metric operationalizes the argument developed in Section IV: subordinated speech can remain visible while being processed first as danger.

The fifth metric is the **Victim-Object Conversion Rate**. This measures how often subordinated actors are represented primarily as passive victims, humanitarian populations, casualties, refugees, displaced persons, or affected civilians without corresponding preservation of political agency. Humanitarian description is not inherently false or illegitimate. Civilian suffering must be named. The problem arises when humanitarian visibility replaces political subjecthood. A population can be made visible through suffering while being denied the grammar of action. This is the point at which humanitarian recognition becomes structurally compatible with responsibility loss (Startari, 2026).

The sixth metric is **Threat-Frame Frequency**. This measures how often an actor is associated with threat language, including extremism, terrorism, instability, danger, escalation, radicalization, regional risk, security concern, or dangerous organization adjacency. The purpose is not to forbid threat classification. Some actors and actions may warrant security description. The methodological question is distributional: is threat language applied symmetrically, contextually, and with preserved causality, or is it disproportionately attached to subordinated actors while dominant-power coercion is described through institutional or defensive abstraction?

The seventh metric is **Humanitarian-Frame Frequency**. This measures how often an actor appears through humanitarian categories such as suffering, casualties, displacement, famine, aid, trauma, crisis, civilian harm, or emergency relief. This metric must be interpreted carefully. Humanitarian framing can be necessary, especially in contexts of mass harm. However, when humanitarian framing dominates the representation of subordinated actors while dominant actors retain active institutional agency, the result is asymmetric visibility. The dominated appear as suffering bodies; the dominant remain actors.

Additional secondary indicators refine the model. **Passive Attribution Rate** measures the frequency of passive constructions in which the affected actor appears but the responsible

actor is omitted or backgrounded. **Agent Deletion Rate** measures clauses where an action or harm is described without an identifiable agent. **Responsibility Displacement Rate** measures cases where responsibility is moved from actors to abstractions such as "violence," "conflict," "tensions," "war," "instability," or "the situation." **Security-Frame Asymmetry** compares the distribution of security language across actor categories. **Political-Subjecthood Retention Rate** measures whether each actor category retains access to political verbs and causal predicates across outputs.

These metrics lead to the paper's derived index: the **Asymmetric Visibility Index**. AVI measures the disparity between the grammatical agency preserved for dominant-power actors and the agency preserved for subordinated actors across AI-generated or AI-mediated conflict discourse. Conceptually, AVI increases when dominant actors appear as active institutional subjects, subordinated actors appear as passive victims or risk objects, responsibility for dominant-power violence is diluted, subordinated resistance is securitized, and humanitarian or bureaucratic frames replace causal attribution. AVI decreases when all actors retain comparable access to subject position, active predicates, causal explanation, responsibility attribution, and political claim-making.

The index does not require the analyst to prove that an AI system "intends" to favor one side. It does not require psychological claims about model designers or platform moderators. Its object is formal distribution. If a corpus of AI-generated summaries repeatedly gives dominant actors higher APR scores and subordinated actors higher risk-object or victim-object conversion scores, then asymmetric visibility can be observed as a measurable pattern. The finding may then be tested against prompts, model versions, languages, moderation settings, source material, and output constraints.

The unit of analysis is the clause. Each output should be segmented into clauses, and each clause should be coded for actor category, grammatical role, verb type, voice, frame, attribution, and responsibility structure. Actor categories should be defined before coding. At minimum, the coding scheme should distinguish dominant-power actors, subordinated actors, civilian populations, state institutions, non-state armed actors, platform governance actors, international institutions, and abstract entities such as "conflict," "violence," "tensions," or "the crisis." Without this separation, the analysis risks collapsing civilians,

governments, armed groups, ideologies, religious identities, and platform objects into the same category.

The coding procedure should also distinguish between political critique, identity-based hate, violent incitement, historical description, humanitarian reporting, and moderation classification. This is especially important in the analysis of Zionism and anti-Zionism. Anti-Zionism cannot be automatically equated with antisemitism, and antisemitism cannot be dismissed as merely political critique. These categories require separate coding because each carries different implications for agency, moderation, and legitimacy (Klug, 2003; Ullrich, 2019). A credible method must preserve those distinctions rather than dissolve them into ideological assertion.

A minimal coding table for each clause should include the following fields in prose form: output identifier, prompt condition, model or system used, source text if applicable, actor mentioned, actor category, grammatical role, main verb, voice, agency status, frame type, attribution status, responsibility status, moderation language, and coder note. The coding decision should be reproducible. If a clause is ambiguous, the ambiguity should be marked rather than forced into a clean category. Ambiguity is itself analytically relevant when it weakens responsibility or agency.

The method should use controlled prompt pairs. For example, one prompt may ask for a summary of an Israeli military action and another for a summary of Palestinian resistance; one may ask for a summary of United States sanctions and another for Iranian responses to sanctions; one may ask for a moderation assessment of anti-Zionist speech and another for a moderation assessment of pro-state security rhetoric. The prompts should be structurally parallel so that differences in output can be attributed more plausibly to actor category and framing, not to prompt design. This follows standard concerns in experimental discourse testing: input symmetry matters if output asymmetry is to be measured.

The method should also include source-controlled summaries. A single source paragraph can be rewritten with actor names varied while preserving event structure. This allows the analysis to test whether the model changes agency allocation when the actor category changes. For example, the same syntactic template may be used with different actors

occupying the role of state, sanctioned state, occupied population, platform authority, or protest movement. If the system preserves agency differently under equivalent syntactic conditions, the result is evidence of actor-category sensitivity.

The methodology must include intercoder reliability if human coding is used. Clause-level coding of agency, frame, and attribution contains interpretive judgment. To reduce arbitrariness, at least two coders should code a subset of outputs independently, compare decisions, and refine the coding manual. The goal is not to eliminate interpretation entirely, but to make coding decisions explicit and reproducible. This is consistent with established practice in content analysis and discourse studies, where reliability procedures are used to stabilize categories across coders (Krippendorff, 2018; Neuendorf, 2017).

The model can also be audited through negative and positive controls. A positive control would use sentences where agency is explicit and should be preserved by any adequate system, such as “the government imposed sanctions” or “the military ordered the strike.” A negative control would use sentences where agency is genuinely unknown, such as “an explosion occurred and the responsible actor has not been identified.” If a system deletes agency in positive-control cases or invents agency in negative-control cases, that indicates a separate reliability problem. Agency preservation should not become forced attribution. The goal is to preserve agency where the input or context supports it.

This distinction prevents the framework from becoming accusatory rather than analytic. Not every passive construction is manipulative. Not every humanitarian frame is agency-erasing. Not every security frame is illegitimate. Not every moderation warning is suppression. The methodological claim is narrower: when these forms are distributed unevenly across actor categories, they may produce asymmetric visibility. The question is not whether a single clause is politically suspicious. The question is whether a measurable pattern appears across comparable outputs.

The framework also separates factuality from agency. A model may produce a factually accurate output that still weakens agency. Conversely, a model may preserve agency while making factual errors. These are different evaluation dimensions. Factuality asks whether the content corresponds to verifiable events. Agency preservation asks whether actors

remain grammatically attached to action and responsibility. Both dimensions matter, but one cannot substitute for the other. This is why existing AI evaluation focused on hallucination, toxicity, or bias must be supplemented with agency metrics.

In practical terms, APR and AVI allow AI ethics to move from general claims to auditable patterns. Instead of saying that an output "feels biased" or "sounds neutral," the analyst can ask: how many clauses preserve dominant-power agency? How many preserve subordinated agency? How often are subordinated actors converted into risks? How often are dominant-power actions nominalized or bureaucratized? How often does humanitarian framing replace responsibility attribution? How often does moderation language appear around one category of speech more than another? These questions are measurable.

The proposed method also gives the series a reusable analytical instrument. Paper 1 identified the humanitarian passive and responsibility loss. Paper 2 expands the analysis into agency preservation and asymmetric visibility. Future applications can test the same framework across Ukraine, Sudan, migration discourse, policing, healthcare, climate disasters, sanctions regimes, platform moderation, and automated legal or bureaucratic summaries. The value of the model is that it does not depend on one conflict or one ideological field. It identifies a formal structure that can be tested across domains.

The section's methodological conclusion is direct: asymmetric visibility becomes researchable only when visibility is separated from agency. Mention is not enough. Sympathy is not enough. Humanitarian description is not enough. Neutral tone is not enough. The relevant question is whether actors retain grammatical access to political action. APR measures that preservation at the clause level. AVI measures the disparity across actor categories. Together, they convert the paper's thesis into an auditable framework: the powerful act, the dominated appear, and the difference can be measured.

VI. Demonstration: Palestine, Iran, and AI-Generated Conflict Summaries

The framework proposed in this article can be demonstrated through controlled comparison rather than through accusation. The purpose of the demonstration is not to prove that any

single model, platform, or moderation system consciously favors one geopolitical actor. The purpose is narrower: to show how agency can be redistributed when AI systems summarize, classify, or rewrite conflict-related discourse. If dominant-power actors retain subject position, institutional verbs, defensive predicates, and responsibility control, while subordinated actors appear more frequently as risks, victims, humanitarian populations, or moderation objects, then asymmetric visibility becomes observable as a formal pattern.

A valid demonstration must begin with structurally parallel prompts. Without prompt symmetry, output asymmetry may simply reflect input asymmetry. For that reason, the test design should compare outputs generated from equivalent prompt structures applied to different actor categories. For example, a prompt may ask the model to summarize an Israeli military action and then ask it to summarize a Palestinian act of resistance using the same requested length, tone, source constraints, and analytical frame. A second pair may ask the model to summarize United States sanctions on Iran and Iranian responses to those sanctions under equivalent conditions. A third pair may ask the model to evaluate anti-Zionist speech and pro-state security rhetoric as potential moderation cases, while preserving the distinction between political critique, identity-based hate, violent incitement, and historical analysis.

The first demonstration domain is Palestine-related discourse. A model may produce an output in which Israeli state-linked actors “respond,” “defend,” “target,” “conduct operations,” “seek security,” or “negotiate,” while Palestinians “are killed,” “are displaced,” “are affected,” “are caught in violence,” or “face humanitarian crisis.” Both actor categories appear. The output may even mention Palestinian suffering extensively. Yet the grammatical distribution is unequal if one actor repeatedly occupies the role of institutional subject and the other repeatedly occupies the role of affected object. In this case, the issue is not erasure. It is visible subordination. The dominated actor appears, but mainly as the bearer of consequences.

A second Palestine-related comparison concerns responsibility attribution. An AI summary may state that “civilians were killed during clashes,” “homes were destroyed amid fighting,” or “aid access deteriorated as the conflict intensified.” These formulations preserve the event but weaken the path to agency. The responsible actor may be deleted,

displaced into an abstraction, or replaced by nouns such as “clashes,” “conflict,” “violence,” or “tensions.” This is the mechanism identified in the first article of the series as responsibility loss (Startari, 2026). In the present article, the question is whether such responsibility loss appears more frequently when dominant-power violence is involved than when subordinated actors are framed as sources of threat.

A third Palestine-related comparison concerns anti-Zionist speech. A controlled prompt may ask the model to classify or summarize a political statement criticizing Zionism, Israeli state policy, occupation, military action, or United States backing. The output should be evaluated for category preservation. Does the model distinguish political critique from antisemitism? Does it preserve the speaker as a political subject making a claim? Or does it immediately route the statement into risk language such as extremism, hate speech concern, escalation, inflammatory rhetoric, or dangerous adjacency? The analytical requirement is not to deny antisemitism. Antisemitic content must be identified when present. The requirement is to test whether the model collapses anti-Zionism, criticism of state policy, and identity-based hatred into the same moderation pathway (Klug, 2003; Ullrich, 2019).

The second demonstration domain is Iran-related discourse. Iran is a useful test case because it is frequently represented through sanctions, nuclear risk, terrorism designations, regional destabilization, proxy conflict, and diplomatic crisis. These frames may correspond to real political disputes, but the question here is grammatical distribution. In AI-generated summaries, United States foreign-policy actors may appear as subjects who “impose sanctions,” “seek deterrence,” “pursue diplomacy,” “warn against escalation,” or “protect regional security.” Iranian actors may appear as “threatening,” “destabilizing,” “sanctioned,” “isolated,” “monitored,” or “accused of violations.” The asymmetry lies in the difference between actor-as-manager and actor-as-risk.

A controlled Iran prompt should therefore compare two equivalent formulations. The first may ask: “Summarize the United States decision to impose sanctions on Iran.” The second may ask: “Summarize Iran’s response to United States sanctions.” The outputs should then be coded for subject position, verb type, agency preservation, responsibility attribution, and frame. If the United States appears mainly through verbs of policy, security,

diplomacy, and decision, while Iran appears mainly through threat, violation, instability, or defensive denial, the Asymmetric Visibility Index increases. If both actors retain comparable agency, causal context, and responsibility-bearing predicates, the index decreases.

A second Iran-related test should examine sanctions discourse. Sanctions are often represented as policy instruments rather than as coercive actions with civilian effects. A model may write that "sanctions were imposed," "economic pressure increased," or "restrictions affected access to goods," while weakening the agency of the sanctioning actor. Conversely, Iranian responses may be described as "defiance," "provocation," "evasion," or "destabilizing behavior." The coding question is whether coercive action by dominant-power actors is bureaucratized while response by the subordinated or sanctioned actor is securitized. This is a direct test of asymmetric visibility because both sides appear, but not necessarily under equivalent grammatical conditions.

A third Iran-related test concerns sovereignty language. If a model summarizes an Iranian claim as "Tehran insists it has sovereign rights" while immediately surrounding it with "concerns," "violations," "threats," and "regional risk," the claim may remain visible but subordinated to security framing. By contrast, if United States or allied claims are summarized through "security guarantees," "deterrence," "nonproliferation," or "regional stability," those actors retain institutional legitimacy. The issue is not whether one side is factually right. The issue is whether the model assigns political subjecthood and causal explanation symmetrically across actor categories.

The third demonstration domain is platform moderation around conflict narratives. Moderation outputs are especially important because they often convert political speech into compliance language. A statement may be treated less as a claim about occupation, sanctions, military action, or civilian harm and more as a potential violation of hate speech, extremism, dangerous organization, incitement, or misinformation policy. These categories may be necessary for platform governance, but they can also produce risk-object conversion when applied unevenly. The correct test is whether the model distinguishes political accusation from identity hatred, dissent from incitement, historical analysis from endorsement of violence, and anti-imperial critique from extremism.

A controlled moderation test should use paired statements with parallel structure. One statement may criticize Zionism as a political ideology or Israeli state policy. Another may criticize Iranian state policy or United States foreign policy. A third may criticize an armed group. A fourth may contain actual identity-based hostility. The model should be evaluated on whether it preserves category distinctions. A reliable system should identify real hate or incitement without collapsing political critique into hate speech. It should also avoid sanitizing dominant-power coercion through neutral administrative language. If the model applies cautionary language disproportionately to subordinated speech while preserving dominant institutional speech as political argument, risk-object conversion is present.

The demonstration should code each output clause by clause. The minimum coding fields are actor, actor category, grammatical role, main verb, voice, frame, agency status, attribution status, responsibility status, and moderation language. For instance, the clause "Israel said it acted in self-defense" preserves Israel-linked institutional agency through subject position and a defensive predicate. The clause "Palestinians were displaced during the operation" represents Palestinians as affected objects and deletes or weakens the responsible actor. The clause "anti-Zionist content may raise safety concerns" converts a political category into a moderation object unless the output specifies the content feature that triggers legitimate concern. The clause "Iranian officials argued that sanctions violate sovereignty" preserves Iranian agency because the actor remains attached to a political claim.

The same coding can be applied to outputs produced from source-controlled summaries. A source paragraph may state: "State A imposed sanctions on State B, causing economic disruption. State B officials accused State A of violating sovereignty and demanded international review." The paragraph can then be rewritten with different actor substitutions: United States and Iran, Israel and Palestine, platform authority and user group, or international institution and sanctioned population. If the model changes agency allocation when actor identities change while event structure remains constant, that indicates actor-category sensitivity. This does not prove intent. It proves distributional variation in grammatical agency.

The demonstration also requires negative controls. If the source text genuinely lacks a responsible actor, the model should not invent one. For example, “an explosion occurred and responsibility has not been established” should not be rewritten as if a specific actor is responsible. Agency preservation is not forced attribution. It is the preservation of agency where agency is present or inferable from the source. This distinction is necessary because the framework does not authorize speculative blame. It measures whether available agency is preserved, weakened, displaced, or redistributed.

The expected pattern, if asymmetric visibility is present, would include five observable results. First, dominant-power actors would show higher Agency-Preservation Rates than subordinated actors. Second, subordinated actors would show higher risk-object and victim-object conversion rates. Third, passive attribution and agent deletion would appear more frequently in descriptions of dominant-power coercion or violence. Fourth, subordinated resistance or grievance would be more frequently securitized than dominant-power coercion. Fifth, moderation language would appear more readily around anti-Zionist, Palestinian, Iranian, or anti-imperial speech than around dominant institutional claims, even when the prompts are structurally parallel.

If the expected pattern does not appear, the framework still remains useful. A low or neutral Asymmetric Visibility Index would indicate that a model preserves agency more evenly than predicted. That would be an important result. The purpose of APR and AVI is not to guarantee a predetermined conclusion. Their purpose is to make agency distribution auditable. A credible framework must be able to detect both asymmetry and its absence. Otherwise, it becomes political assertion rather than measurement.

This demonstration also clarifies the difference between factual evaluation and agency evaluation. A model may produce a factually accurate sentence that still weakens agency. “Civilians were displaced during the conflict” may be factually true, but it does not identify the actor responsible for displacement. Conversely, a model may preserve agency while making a factual error. These are separate dimensions. Factuality asks whether the statement is true. Agency preservation asks whether the grammatical path between actor, action, and responsibility remains intact. AI evaluation needs both.

The demonstration therefore supports the paper's central theoretical movement. Paper 1 showed that suffering can appear without responsibility. This paper shows that political actors can appear without equal agency. Palestine and Iran are not used as rhetorical examples only. They are used as high-density test cases because they combine occupation, sanctions, resistance, military action, security discourse, humanitarian suffering, platform moderation, and dominant-power framing. These conditions make agency asymmetry visible at the level of clause structure.

The methodological result is direct. If AI-generated summaries preserve institutional agency for dominant-power actors while converting subordinated actors into victims, threats, risks, or moderation objects, then visibility has not produced equality. It has produced asymmetric visibility. The relevant question is not whether Palestine or Iran appears in the output. The relevant question is whether Palestinian, Iranian, anti-Zionist, or anti-imperial actors retain grammatical access to political agency when placed beside Israel-linked institutional discourse, United States foreign-policy language, and platform moderation categories.

The demonstration confirms the need for agency-based auditing. AI systems should not be evaluated only on whether they avoid hate speech, misinformation, or hallucination. They should also be evaluated on whether they preserve the grammar through which political actors remain legible as agents. Without such auditing, a system can appear balanced while structurally reproducing domination. It can mention the dominated constantly, sympathize with their suffering, warn against escalation, and still deny them equal grammatical access to action. That is the formal condition named in this paper as asymmetric visibility.

VII. Conclusion: From Bias Detection to Agency Detection

The central argument of this article is that AI-mediated conflict discourse cannot be evaluated only by asking whether political actors are mentioned, whether tone appears balanced, or whether explicit hate speech is avoided. Those standards remain necessary, but they are not sufficient. A system can mention Palestine, Iran, anti-Zionism, occupation, sanctions, civilian suffering, Israeli security claims, United States foreign policy, and

platform moderation while still distributing agency unevenly. Visibility is therefore not equality. The decisive question is whether actors retain grammatical access to political action.

This paper has named that condition **asymmetric visibility**. Asymmetric visibility occurs when dominated, occupied, sanctioned, or subordinated actors remain visible while their agency is grammatically weakened. They may appear as victims, displaced populations, humanitarian categories, moderation concerns, security risks, destabilizing forces, or crisis markers. Dominant-power actors, by contrast, may retain stronger access to institutional subjecthood, active predicates, defensive framing, strategic decision, diplomatic agency, and responsibility control. The result is not disappearance. It is asymmetrical presence.

The theoretical movement from the first paper in the series is direct. **Responsibility loss** showed how suffering can be described while the grammatical path to responsible actors is weakened or removed (Startari, 2026). **Asymmetric visibility** extends that mechanism into a broader political condition. The problem is no longer only whether harm appears without a perpetrator. The problem is whether different actors appear under unequal grammatical conditions. Responsibility loss explains how accountability disappears inside humanitarian description. Asymmetric visibility explains how political agency is distributed unequally across dominant and subordinated actors.

This distinction matters because much of AI ethics remains organized around bias detection, toxicity reduction, misinformation control, representational inclusion, and safety classification. These categories address real problems, but they can miss agency asymmetry. A model may avoid slurs and still securitize subordinated speech. It may avoid hallucination and still nominalize dominant-power violence. It may include victims and still omit perpetrators. It may classify dangerous content correctly in some cases and still convert political critique into moderation risk in others. The evaluation of AI-generated political discourse must therefore include agency detection, not only bias detection.

Agency detection asks a more precise question: when an actor appears, does the actor remain grammatically capable of acting? This question shifts the analysis from surface representation to clause-level structure. It asks who occupies subject position, what verbs

are assigned, whether voice is active or passive, whether agency is deleted, whether responsibility is displaced, whether violence is bureaucratized, whether resistance is securitized, and whether political claims are converted into moderation objects. These are not rhetorical impressions. They are measurable features of discourse.

The **Agency-Preservation Rate** proposed in this article operationalizes that question. APR measures the proportion of conflict-related clauses in which a political actor remains represented as capable of action, decision, resistance, governance, imposition, sanction, occupation, negotiation, moderation, or responsibility. The **Asymmetric Visibility Index** then compares agency preservation across actor categories. AVI rises when dominant-power actors retain active institutional agency while subordinated actors are routed into passive, humanitarian, security, or moderation frames. AVI falls when all actors retain comparable access to subject position, active predicates, causal explanation, and responsibility-bearing clauses.

This framework does not require claims about conspiracy, intention, or psychological motivation. It does not require asserting that a model is simply "pro-Israel," "anti-Palestinian," "anti-Iranian," or aligned with any single geopolitical actor. Such claims would require separate empirical proof and may weaken the formal precision of the analysis. The claim made here is narrower: AI-generated and AI-mediated discourse can preserve political agency unevenly, and that unevenness can be measured. The unit of analysis is not intention. It is grammatical distribution.

The cases of Palestine, Zionism, anti-Zionism, Iran, United States foreign-policy discourse, Israel-linked institutional discourse, and platform moderation show why this framework is necessary. These cases force AI systems to process identity protection, state legitimacy, occupation, sanctions, resistance, civilian harm, security claims, anti-imperial critique, moderation policy, and historical asymmetry at the same time. In such domains, grammatical agency is not a stylistic detail. It determines whether an actor appears as a political subject or as an object inside another actor's frame.

The paper has also insisted on category separation. Judaism, Zionism, Israeli state policy, antisemitism, anti-Zionism, Palestinian political claims, Iranian state discourse, Iranian

civilian life, United States foreign policy, armed organizations, and platform moderation are not interchangeable categories. A credible AI system must identify antisemitism when it appears. It must also preserve the distinction between antisemitism, criticism of Zionism, criticism of Israeli state policy, anti-imperial critique, historical analysis, and incitement to harm (Klug, 2003; Ullrich, 2019). Without that distinction, moderation can become a mechanism of political compression rather than analytical clarification.

The same principle applies to Iran-related discourse. A system must be able to describe nuclear disputes, sanctions, regional security concerns, and state violence without automatically reducing Iranian actors to threat categories. It must also be able to describe United States or allied coercive action without hiding agency behind bureaucratic abstractions such as “pressure,” “restrictions,” “security measures,” or “regional stability.” The point is not to reverse asymmetry by assigning automatic innocence or guilt. The point is to preserve agency, causality, and responsibility wherever the source material supports them.

The article’s broader contribution is to move AI discourse evaluation from **representation** to **political subjecthood**. Representation asks whether an actor appears. Political subjecthood asks how the actor appears. A population can be repeatedly mentioned and still be denied agency. It can be visible as suffering but not as claimant, visible as danger but not as governed subject, visible as crisis but not as historical actor, visible as content but not as speech. This is the formal structure captured by asymmetric visibility.

The implication for AI governance is direct. Audits of AI-generated political discourse should include agency-preservation tests. Platforms, researchers, and policy analysts should not only ask whether outputs are toxic, biased, false, or inflammatory. They should also ask whether outputs preserve the grammatical pathways through which political actors remain legible as agents. This requires clause-level coding, controlled prompt pairs, actor-category comparison, source-controlled summaries, and explicit separation between political critique, identity hatred, violent incitement, historical description, and moderation classification.

The implication for linguistics is equally direct. Syntax is not secondary to politics in AI-mediated discourse. It is one of the formal sites through which political agency is preserved, weakened, or redistributed. Passive constructions, nominalizations, abstract causality, risk framing, humanitarian framing, moderation categories, and institutional predicates are not merely stylistic choices. They are mechanisms through which actors are made capable or incapable of appearing as responsible subjects. This extends Startari's prior work on passive voice, objectivity, syntactic authority, and responsibility loss in AI-generated language (Startari, 2025a, 2025b, 2025c, 2026).

The article's final claim is therefore controlled but consequential. AI systems do not need to erase subordinated populations in order to weaken them politically. They can mention them constantly. They can summarize their suffering. They can classify their speech. They can warn about risks around them. They can include them in neutral-sounding summaries. Yet if those populations do not retain equal grammatical access to agency, visibility becomes a managed form of subordination.

The standard proposed here is not visibility alone, but **agency-preserving visibility**. A political actor is not adequately represented merely because it appears in an output. It must also retain the grammatical possibility of acting, claiming, accusing, resisting, governing, negotiating, suffering under identifiable causes, and being placed within a traceable field of responsibility. Without that preservation, AI-mediated discourse can produce recognition without subjecthood.

The conclusion can be stated in the terms of the series: the humanitarian passive makes suffering visible while weakening responsibility; risk-object conversion makes resistance visible while weakening political cause; asymmetric visibility makes subordinated actors visible while weakening agency. Together, these mechanisms show that AI-generated neutrality must be audited not only for what it says, but for what its grammar permits actors to be.

The final thesis is therefore simple: **visibility is not equality when agency is not preserved**. The powerful act. The dominated appear. Agency detection is the method required to measure the difference.

Full Bibliography

Baker, P., Gabrielatos, C., & McEnery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British press*. Cambridge University Press.

Bhatia, A. (2005). Critical discourse analysis of political press conferences. *Discourse & Society, 17*(2), 173–203. https://doi.org/10.1177/0957926506058057

Bunzl, M. (2007). *Anti-Semitism and Islamophobia: Hatreds old and new in Europe*. Prickly Paradigm Press.

Buzan, B., Wæver, O., & de Wilde, J. (1998). *Security: A new framework for analysis*. Lynne Rienner Publishers.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43*(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Fairclough, N. (1995). *Critical discourse analysis: The critical study of language*. Longman.

Hall, S. (Ed.). (1997). *Representation: Cultural representations and signifying practices*. Sage.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar* (4th ed.). Routledge.

Klug, B. (2003). The collective Jew: Israel and the new antisemitism. *Patterns of Prejudice, 37*(2), 117–138. https://doi.org/10.1080/0031322032000087973

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). Sage.

Neuendorf, K. A. (2017). *The content analysis guidebook* (2nd ed.). Sage.

Startari, A. V. (2025a). *The passive voice in artificial intelligence language: Algorithmic neutrality and the disappearance of agency*. SSRN. https://doi.org/10.2139/ssrn.5263305

Startari, A. V. (2025b). *The grammar of objectivity: Formal mechanisms for the illusion of neutrality in language models*. SSRN. https://doi.org/10.2139/ssrn.5319520

Startari, A. V. (2025c). *AI and syntactic sovereignty: How artificial language structures legitimize non-human authority*. SSRN. https://doi.org/10.2139/ssrn.5276879

Startari, A. V. (2025d). *Ethos without source: Algorithmic identity and the simulation of credibility*. SSRN. https://doi.org/10.2139/ssrn.5313317

Startari, A. V. (2025e). *TLOC, The irreducibility of structural obedience in generative models*. SSRN. https://doi.org/10.2139/ssrn.5303089

Startari, A. V. (2026). *Suffering without perpetrators: The humanitarian passive in AI-generated conflict discourse*. In *Grammars of asymmetric visibility: AI, imperial power, and the syntax of responsibility*.

Ullrich, P. (2019). Expert report on the working definition of antisemitism of the International Holocaust Remembrance Alliance. *Rosa Luxemburg Stiftung*.

van Dijk, T. A. (1998). *Ideology: A multidisciplinary approach*. Sage.

van Leeuwen, T. (2008). *Discourse and practice: New tools for critical discourse analysis*. Oxford University Press.