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

Suffering Without Perpetrators: The Humanitarian Passive in AI-Generated Conflict Discourse.

Agustin V. Startari.

Cita:

Agustin V. Startari (2026). *Suffering Without Perpetrators: The Humanitarian Passive in AI-Generated Conflict Discourse. AI Power and Discourse, 1 (1), 1-10.*

Dirección estable: https://www.aacademica.org/agustin.v.startari/231

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

Suffering Without Perpetrators: The Humanitarian Passive in AI-Generated Conflict Discourse

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 12, 2026

DOI

- Primary archive: https://doi.org/**10.5281/zenodo.20139961**
- Secondary archive: https://doi.org/10.6084/m9.figshare.32253825
- SSRN: Pending assignment (ETA: Q3 2026)

Language: English

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

Word count: 17102

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 the *humanitarian passive* as a machine-mediated syntactic pattern through which civilian suffering is made visible while responsibility is made grammatically optional. In AI-generated conflict discourse, victims, destruction, displacement, sanctions, shortages, and humanitarian catastrophe can remain fully describable, while perpetrators, institutional agents, military actors, sanctioning powers, platform authorities, and decision chains are reduced, displaced, or omitted. The result is not the erasure of suffering, but the weakening of the grammatical pathways through which suffering remains attributable to responsible actors. The paper defines this weakening as *responsibility loss*: the measurable decline of syntactic traceability between civilian harm and the agents, institutions, states, military systems, or geopolitical alliances that produce it. Responsibility loss does not require proving intention, censorship, conspiracy, or ideological coordination. It can be examined as a formal effect of summary compression, platform moderation, safety constraints, geopolitical framing, and machine-generated neutrality. Palestine functions as the central test case because it exposes the conflict between visibility and attribution with unusual intensity. AI systems may recognize humanitarian suffering, civilian death, displacement, infrastructure collapse, bombardment, and crisis conditions while simultaneously avoiding or diluting the grammatical assignment of agency. Iran extends the model to sanctioned suffering, where civilian harm may be described through the neutral grammar of shortages, economic hardship, restricted access, escalation, regional instability, and security concern, while the actors and systems imposing sanctions, financial restrictions, diplomatic pressure, or military threat become syntactically less visible. The article therefore treats Palestine and Iran not as identical cases, but as distinct sites of asymmetric visibility. In Palestine, responsibility loss appears through the grammar of direct violence, occupation, bombardment, displacement, and humanitarian catastrophe. In Iran, it appears through the grammar of sanctions, isolation, overcompliance, diplomatic pressure, securitization, and indirect civilian harm. Together, these cases show how dominant-power discourse can preserve the visibility of suffering while weakening the grammatical conditions under which responsibility can be named.

Methodologically, the article proposes a replicable model for measuring responsibility loss across AI summaries, headlines, platform notices, reports, policy statements, and controlled prompts. The proposed indicators include passive ratio, agent deletion rate, victim subject ratio, perpetrator subject ratio, nominalization density, humanitarian-frame frequency, security-frame frequency, agency-preservation rate, and a *Responsibility Loss Index* (RLI). These metrics allow the paper to move beyond general claims about bias and toward a precise account of how grammar redistributes political visibility. As the foundational article of *Grammars of Asymmetric Visibility: AI, Imperial Power, and the Syntax of Responsibility*, this paper establishes responsibility loss as a central object of linguistic, legal, and political analysis. Its core claim is that AI ethics and platform governance should not only ask whether harmful bias, hate speech, misinformation, or extremist content are present. They must also ask whether the grammar of machine-generated discourse preserves the conditions under which responsibility can still be named. Suffering remains visible. Responsibility disappears. Ultimately, the paper contends that the defining characteristic of plagiarism in predictive systems is no longer duplication, but the disappearance of recoverable epistemic provenance under conditions of synthetic generation. In large language models, originality increasingly emerges as a statistical effect of recombination depth rather than as evidence of autonomous intellectual creation.

I. Introduction: Visibility Without Responsibility

AI-generated conflict discourse does not necessarily erase civilian suffering. In many cases, suffering remains visible, explicit, and narratively central. Civilians are described as displaced. Hospitals are described as overwhelmed. Infrastructure is described as damaged. Food, medicine, water, electricity, shelter, and security are described as scarce or unavailable. Death, injury, displacement, and fear remain present in the text. The problem examined in this article is therefore not the complete disappearance of suffering from machine-generated discourse. The problem is more precise: suffering may remain visible while the grammatical pathways that connect that suffering to responsible agents are weakened, displaced, or removed.

This distinction matters because visibility alone does not guarantee accountability. A text can describe catastrophe without preserving causality. It can name victims without naming perpetrators. It can record damage without identifying the actor, institution, military structure, sanctioning power, platform authority, or political decision that produced it. In such cases, the visible object is not responsibility, but consequence. The event appears as humanitarian condition, emergency, escalation, instability, or crisis. The grammar allows the reader to see that harm occurred, but not necessarily who acted, who authorized, who benefited, or who could be held accountable. This problem extends prior work on passive voice in artificial intelligence language, where agent deletion is not treated as a merely stylistic choice, but as a structural mechanism through which agency can disappear from machine-mediated discourse (Startari, 2025a).

This article calls this pattern the *humanitarian passive*. The term refers to a machine-mediated syntactic structure through which civilian suffering is made visible while responsibility becomes grammatically optional. The structure is not limited to the passive voice in a narrow grammatical sense, although passive constructions are central to it. It also includes nominalizations, agentless clauses, abstract causality, decontextualized humanitarian vocabulary, moderation-oriented reformulations, and security frames that transform political violence into apparently neutral description. The humanitarian passive does not deny that suffering exists. Its effect is different: it separates suffering from the agents and institutions that produce it.

The central object of analysis is *responsibility loss*. Responsibility loss occurs when the grammatical traceability between harm and responsible agency declines. This decline can be observed when "X bombed a residential area" becomes "a residential area was hit," when "sanctions imposed by X restricted access to medicine" becomes "access to medicine became limited," or when "platform Y removed Palestinian content" becomes "content was removed for violating policy." In each case, the factual surface may still refer to harm, restriction, or removal. What changes is the grammatical availability of the agent. This mechanism also connects with the broader problem of authority without visible source, where institutional or algorithmic legitimacy can be simulated even when the responsible subject is absent, displaced, or formally unavailable (Startari, 2025b; Startari, 2025e).

The article does not claim that every passive construction is deceptive, nor that every AI-generated summary intentionally hides responsibility. Such a claim would be too broad and empirically weak. Passive constructions can be necessary, accurate, or stylistically justified when the agent is unknown, irrelevant, or already established. The argument is narrower and more measurable. In contexts of political violence, occupation, sanctions, platform moderation, and great-power pressure, repeated agent deletion can produce a systematic redistribution of visibility. Victims remain describable. Damage remains describable. Humanitarian need remains describable. Responsibility becomes less stable.

This redistribution is not only semantic. It is syntactic. Machine-generated summaries may preserve the vocabulary of suffering while modifying the grammatical structure through which suffering is attributed. An active clause can assign responsibility directly. A passive clause may reduce or remove the agent. A nominalization can convert action into condition. A security frame can convert coercion into instability. A humanitarian frame can convert produced harm into abstract need. These transformations matter because grammar shapes the conditions under which responsibility can be named, contested, or legally reconstructed. The appearance of neutrality may therefore arise from form rather than from balance, objectivity, or evidentiary caution. This claim follows the theory that objectivity in language models can operate as a formal effect of grammar, especially when impersonal constructions simulate neutrality while suppressing the visibility of agency (Startari, 2025c).

Palestine serves as the central test case because it exposes the conflict between humanitarian visibility and responsibility attribution with unusual intensity. AI systems may describe civilian death, displacement, infrastructure collapse, bombardment, famine risk, humanitarian crisis, and platform moderation disputes. Yet the grammar of those descriptions may weaken the visibility of military actors, state agency, occupation structures, institutional authorization, and geopolitical support. The analytical point is not that Palestine is invisible. The point is that Palestinian suffering can be made intensely visible while the agents and structures producing that suffering are grammatically reduced.

Iran extends the model to a different but related form of asymmetric visibility. In Iran-related discourse, civilian harm may appear through the vocabulary of sanctions, shortages, restricted access, financial isolation, diplomatic pressure, nuclear tension, regional instability, and military threat. Here the mechanism is often less direct than in descriptions of bombardment or occupation. The relevant harm may be mediated through economic restriction, overcompliance, banking exclusion, supply-chain disruption, and security framing. Yet the syntactic effect can be similar: suffering is described as hardship, instability, or pressure, while sanctioning powers, financial systems, military alliances, and diplomatic actors become grammatically secondary.

Palestine and Iran are therefore not treated as identical cases. They are analytically distinct sites of asymmetric visibility. Palestine foregrounds direct violence, occupation, bombardment, displacement, humanitarian catastrophe, and platform moderation. Iran foregrounds sanctioned suffering, securitization, isolation, coercive diplomacy, and indirect civilian harm. Together, they allow the article to examine how AI-generated discourse may preserve suffering as an object of description while reducing responsibility as an object of attribution. This is not a claim about hidden intention. It is a claim about measurable grammatical transformation.

The broader claim is that AI ethics, platform governance, and computational discourse analysis should not limit themselves to detecting bias, misinformation, hate speech, or extremist content. These categories remain important, but they do not exhaust the problem. A system can avoid hate speech and still erase agency. A summary can be factually cautious and still weaken causality. A moderation notice can be procedurally correct and still hide

the censoring institution behind policy grammar. A conflict summary can sound neutral while redistributing political responsibility through syntax. In this sense, the humanitarian passive belongs to a wider theory of syntactic sovereignty, where authority is produced through linguistic form even when the responsible subject is absent from the surface of the sentence (Startari, 2025d; Startari, 2025f).

The question, then, is not only whether AI-generated discourse mentions suffering. The question is whether it preserves the grammatical conditions under which responsibility can still be named. This is the foundational problem of *Grammars of Asymmetric Visibility*. In machine-generated political speech, power may remain active while appearing grammatically absent. Civilian suffering may remain visible, but the path from suffering to responsibility may disappear.

II. The Humanitarian Passive

The *humanitarian passive* is the syntactic pattern through which civilian suffering is preserved as a visible object of description while the responsible agent becomes grammatically reduced, displaced, or optional. It is not identical to the passive voice as a single grammatical construction. Rather, it names a broader configuration of passive clauses, nominalizations, agentless predicates, abstract causality, institutional euphemisms, and humanitarian frames that allow harm to remain legible while weakening the traceability of responsibility. In this sense, the humanitarian passive extends the problem of passive voice in artificial intelligence language from a general question of agency disappearance to a more specific question of conflict discourse, civilian harm, and political accountability (Startari, 2025a).

The category is necessary because the visibility of suffering can be mistaken for the preservation of responsibility. A sentence may state that civilians were killed, homes were destroyed, hospitals were damaged, or families were displaced. These formulations do not deny harm. They may even intensify the emotional visibility of harm. Yet they can leave unresolved the question of who killed, who destroyed, who damaged, who displaced, who ordered, who funded, who sanctioned, or who moderated the account of the event. The

humanitarian passive therefore does not operate by making violence invisible. Its more precise effect is to make violence visible as consequence while making agency unstable as grammar.

This distinction separates the humanitarian passive from ordinary passive voice. Passive constructions are not inherently evasive. They may be appropriate when the agent is unknown, already established, irrelevant to the immediate focus, or deliberately backgrounded for legitimate analytical reasons. A passive clause can also foreground victims in ways that are ethically necessary, especially when the purpose of a sentence is to document suffering. The argument is not that all passive voice conceals responsibility. The argument is that repeated passive, nominalized, or agentless constructions in asymmetric conflict discourse can produce a patterned loss of attribution. When this pattern appears in machine-generated summaries, platform notices, policy statements, or automated reports, it becomes a measurable feature of discourse rather than a merely stylistic preference.

The humanitarian passive appears most clearly when an active harm-event is transformed into an agentless condition. An active formulation such as “the army bombed the residential block” assigns agency, action, and object. A humanitarian passive formulation such as “the residential block was damaged during the escalation” preserves the damaged object but weakens the responsible subject. A further nominalized formulation, “damage to residential areas increased amid tensions,” removes the action almost entirely and converts violence into an abstract condition. In each case, the informational surface may remain compatible with the fact that harm occurred. The syntactic structure, however, reduces the path from harm to agent.

This mechanism can be described through four basic operations. First, passive reduction moves the affected object into subject position while deleting or backgrounding the actor. Second, nominalization converts actions into nouns, allowing events such as bombing, sanctioning, displacing, censoring, or restricting to appear as damage, hardship, displacement, moderation, or instability. Third, causal abstraction replaces direct agency with vague environmental causes such as “amid tensions,” “during the crisis,” “following escalation,” or “as conditions deteriorated.” Fourth, frame substitution changes the

interpretive field of the sentence, converting political violence into humanitarian need, military action into security response, sanctions into economic pressure, or platform suppression into policy enforcement. These operations do not need to be intentional to be politically consequential. Their effect lies in the redistribution of grammatical visibility.

The concept also connects to the problem of ethos without source. In AI-generated discourse, authority can be simulated through impersonal balance, procedural tone, and neutralized grammatical form. When a model produces a sentence in which responsibility is absent but the statement still sounds institutionally credible, the text acquires an appearance of objectivity without preserving a visible source of judgment or attribution (Startari, 2025b). The humanitarian passive is one conflict-specific form of that broader mechanism. It allows the statement to sound neutral, cautious, and humanitarian while reducing the grammatical presence of responsible actors.

The humanitarian passive also belongs to the grammar of objectivity. A sentence can appear balanced not because it distributes responsibility accurately, but because it removes agency from the sentence structure. For example, “civilian casualties occurred during clashes” can appear less politically charged than “military forces killed civilians,” but the reduction of political charge is produced by syntax, not necessarily by superior evidentiary accuracy. The sentence becomes “neutral” by lowering attribution. This is why objectivity must be analyzed formally rather than assumed from tone. In AI-generated conflict discourse, machine neutrality can become a grammatical effect: the sentence appears impartial because the responsible agent has been removed from the position where responsibility would normally be assigned (Startari, 2025c).

The pattern is especially relevant in contexts of asymmetric power. In symmetrical conflict descriptions, agent deletion may affect all parties in comparable ways. In asymmetric conflicts, however, the deletion of agency may have unequal effects. If a dominant military, state, platform, or sanctioning power is systematically backgrounded, while the harmed population remains visible mainly as victim, crisis, demographic mass, or humanitarian object, the discourse produces asymmetric visibility. The subordinated population does not disappear. It remains visible as suffering, damage, need, risk, or instability. What becomes less visible is the chain of agency that links that suffering to responsible actors.

This structure is central to Palestine-related AI summaries because the discourse often contains several competing pressures at once: humanitarian recognition, accusations of bias, antisemitism enforcement, anti-Zionist speech, security vocabulary, platform moderation, military reporting, and diplomatic caution. The humanitarian passive can operate as a compromise formation among these pressures. It allows the text to acknowledge civilian harm while avoiding explicit attribution to military, state, institutional, or geopolitical actors. The result is not silence. It is a grammar of suffering with weakened causality.

The same category can be extended to Iran, although the mechanism differs. In Iran-related discourse, the humanitarian passive may appear less through direct descriptions of bombardment and more through the grammar of sanctions, shortages, restriction, banking isolation, medical scarcity, nuclear tension, and regional escalation. An active formulation such as “sanctioning powers restricted access to medical imports” can become “access to medical imports was restricted,” or further, “medical shortages worsened amid regional tensions.” The civilian consequence remains describable, but the sanctioning structure, financial mechanism, and political actor become less grammatically accessible. This is why Iran is not an identical case to Palestine, but a necessary comparative case for testing responsibility loss across direct and indirect forms of coercion.

The humanitarian passive should therefore be understood as a political form, not because it requires a hidden political intention, but because it governs the distribution of grammatical roles in discourse about power. Political responsibility depends on more than the presence of moral vocabulary. It depends on whether a sentence preserves the relation between action, agent, object, and consequence. When this relation is weakened, a discourse can remain emotionally vivid while becoming politically under-attributed. The victim remains. The perpetrator becomes a missing grammatical position.

This is also why the humanitarian passive relates to syntactic sovereignty. In previous work, syntactic sovereignty describes the production of authority through linguistic form when agency, command, or legitimacy is generated by structure rather than by an explicitly accountable subject (Startari, 2025d). The humanitarian passive is one of the ways this sovereignty appears in conflict discourse. The sentence itself organizes who can appear as

actor, who appears as object, and whose responsibility remains recoverable. In AI-generated outputs, this organization is not merely rhetorical. It becomes part of the machine-mediated distribution of public intelligibility.

The category also intersects with null subjects of power. A null subject of power is not simply an absent noun phrase. It is a structurally consequential absence: a missing agent whose omission changes how authority, harm, or responsibility can be reconstructed (Startari, 2025e). The humanitarian passive produces such null subjects when the actors responsible for violence, sanctions, displacement, censorship, or coercion are excluded from the grammatical surface while their effects remain visible. The omitted agent is not irrelevant. It is precisely what would be necessary for accountability.

The analytical value of the humanitarian passive lies in its measurability. It can be operationalized through passive ratio, agent deletion rate, nominalization density, victim subject ratio, perpetrator subject ratio, humanitarian-frame frequency, security-frame frequency, and agency-preservation rate. These indicators make it possible to compare source texts and AI-generated summaries, or to compare different summaries of the same event across models, platforms, prompt conditions, and moderation settings. The point is not to infer intention from a single sentence. The point is to measure whether grammatical traceability declines across transformations.

This moves the analysis beyond generic claims about bias. Bias analysis often asks whether a system favors one side, uses loaded vocabulary, suppresses certain claims, or reproduces stereotypes. Those questions remain relevant, but they do not capture the full problem. A model may avoid overtly biased language and still produce responsibility loss. It may remove inflammatory terms and also remove agency. It may sound safer while becoming less accountable. The humanitarian passive therefore identifies a distinct object of analysis: not whether the text sounds hostile or sympathetic, but whether it preserves the grammatical conditions under which responsibility can still be named.

The humanitarian passive can be summarized in one formal proposition: the more a discourse preserves harm while reducing agency, the more it shifts from accountability to humanitarian description. This shift is not neutral in contexts where suffering is produced

by identifiable actors, institutions, laws, weapons, sanctions, policies, or platforms. It transforms political violence into visible crisis. It transforms coercion into pressure. It transforms censorship into moderation. It transforms occupation into humanitarian need. It transforms responsibility into grammar.

For this reason, the humanitarian passive is the foundational category of this paper. It names the syntactic form through which asymmetric visibility becomes possible. Suffering remains present. Responsibility becomes optional. The task of the following sections is to show how this pattern can be measured as responsibility loss, how it appears in Palestine-related discourse, how it extends to Iran and sanctioned suffering, and why AI governance must treat responsibility detection as a necessary complement to bias detection.

III. Responsibility Loss as a Measurable Effect

Responsibility loss is the measurable decline of grammatical traceability between civilian harm and responsible agency. It occurs when a source event that contains an identifiable relation between agent, action, object, and consequence is transformed into a sentence, summary, headline, platform notice, or report in which that relation becomes weaker, less explicit, or grammatically unavailable. The concept does not require proving intent. It does not require demonstrating that a model, platform, journalist, institution, or state actor consciously decided to hide responsibility. It identifies a formal transformation: harm remains visible, but the syntactic path from harm to responsible agent is reduced.

This distinction is essential. A discourse can be emotionally explicit and politically under-attributed at the same time. It can describe civilian deaths, destroyed homes, damaged hospitals, shortages, displacement, fear, hunger, and trauma while weakening the grammatical relation between these consequences and the actors who produced them. Responsibility loss therefore does not measure whether suffering appears. It measures whether suffering remains grammatically connected to agency. This extends the prior analysis of passive voice in artificial intelligence language, where agent deletion was treated as a structural mechanism for the disappearance of agency rather than as a neutral stylistic preference (Startari, 2025a).

The unit of analysis is the clause. A clause can preserve or weaken responsibility depending on how it distributes grammatical roles. An active clause such as “military forces bombed the hospital” contains an explicit agent, a transitive action, and an affected object. A passive clause such as “the hospital was bombed” preserves the harm event but removes the agent from subject position. An agentless passive such as “the hospital was damaged during the escalation” further weakens the relation between action and responsible actor. A nominalized formulation such as “damage to hospitals increased amid the crisis” converts the action into an abstract condition. These transformations do not necessarily falsify the event. Their effect is subtler: they reduce the grammatical availability of responsibility.

Responsibility loss must therefore be measured comparatively. A single sentence may be ambiguous. A single passive construction may be justified. A single nominalization may be analytically useful. The relevant object is not isolated grammar, but transformation across versions. The strongest method compares a source formulation with an AI-generated summary, headline, report, moderation notice, or policy paraphrase. The question is whether the transformed version preserves the same degree of agency, causality, and accountability as the source. If the source text names a responsible actor and the generated text removes or dilutes that actor, responsibility loss has occurred.

This paper defines responsibility loss through four core dimensions: agent visibility, causal explicitness, role preservation, and frame stability. Agent visibility measures whether the responsible actor remains present in the sentence. Causal explicitness measures whether the connection between action and consequence remains grammatically direct. Role preservation measures whether the same actor remains in an agentive position rather than being moved into background, context, or abstraction. Frame stability measures whether the event remains described as an action performed by an actor, or whether it is converted into crisis, instability, humanitarian need, security concern, or policy violation.

The first indicator is passive ratio. Passive ratio measures the proportion of clauses in which the affected object appears in subject position while the actor is backgrounded, displaced, or omitted. Passive constructions are not automatically signs of responsibility loss. Their relevance depends on whether the passive removes an identifiable agent from a politically or legally significant harm event. In conflict discourse, the difference between “forces

killed civilians" and "civilians were killed" is not merely stylistic. The first formulation preserves the actor. The second preserves the victim and the harm, but weakens attribution. This is the grammatical foundation of the humanitarian passive.

The second indicator is agent deletion rate. Agent deletion rate measures the frequency with which an actor present in the source event disappears in the transformed version. This metric is stronger than passive ratio because it directly compares source and output. If a source sentence states that a named state, army, platform, coalition, sanctioning authority, police unit, or institution performed an action, and the AI-generated summary removes that actor, the output registers agent deletion. Agent deletion is central to responsibility loss because accountability requires recoverable agency. When the agent is removed, the event may remain visible, but its responsibility structure is weakened.

The third indicator is victim subject ratio. Victim subject ratio measures how often the affected population appears as the grammatical subject of the sentence. This metric is not negative by itself. Centering victims can be ethically necessary. The problem appears when victim subject prominence increases while perpetrator subject visibility decreases. For example, "families were displaced," "civilians were killed," "children were injured," and "communities were affected" all foreground victims. If these constructions appear without corresponding agentive clauses that identify who displaced, killed, injured, or affected them, the text produces humanitarian visibility without responsibility. The victim becomes visible as subject, but not as the object of an attributable act.

The fourth indicator is perpetrator subject ratio. Perpetrator subject ratio measures how often responsible actors appear as grammatical subjects performing actions. This includes military forces, states, governments, sanctioning powers, police units, intelligence agencies, platform operators, and institutional decision-makers. A high victim subject ratio combined with a low perpetrator subject ratio indicates asymmetric visibility. The harmed population is visible, but the actor responsible for harm is not equally visible as an agent. This relation is central to the broader theory of syntactic sovereignty, where authority and responsibility may be organized by linguistic form rather than by explicit institutional declaration (Startari, 2025b).

The fifth indicator is nominalization density. Nominalization density measures the frequency of nouns that convert actions into abstract entities or conditions. Terms such as "displacement," "destruction," "shortages," "restriction," "escalation," "instability," "moderation," "removal," "pressure," "collapse," and "deterioration" can describe real events. The issue is whether they replace verbs that would otherwise assign agency. "The army destroyed the building" assigns action. "The destruction of the building" converts the action into a noun. "Infrastructure destruction increased amid the crisis" abstracts the event further. Nominalization can therefore preserve the semantic trace of harm while weakening the grammatical structure of accountability.

The sixth indicator is humanitarian-frame frequency. Humanitarian frames describe events through the vocabulary of need, aid, crisis, suffering, protection, access, shelter, food, water, medicine, displacement, and civilian harm. These frames are indispensable in many contexts. Their risk lies in their capacity to convert produced violence into humanitarian condition. If a text repeatedly describes "humanitarian catastrophe," "civilian suffering," "urgent need," or "restricted access" without preserving the actors and policies that produced those conditions, the discourse shifts from accountability to relief. The problem is not humanitarian language itself. The problem is humanitarian language without recoverable agency.

The seventh indicator is security-frame frequency. Security frames describe events through risk, threat, escalation, instability, extremism, terrorism, regional tension, border security, national defense, public safety, or policy violation. In AI-generated conflict discourse, security frames can restructure agency by placing dominant-power actors in defensive or administrative roles while coding subordinated populations as risk objects. For Palestine, security framing may compete with humanitarian visibility by converting occupation, bombardment, displacement, or platform suppression into questions of security, extremism, or policy enforcement. For Iran, security framing may convert sanctions, military pressure, or diplomatic coercion into responses to nuclear risk, regional instability, or threat management.

The eighth indicator is agency-preservation rate. Agency-preservation rate measures the proportion of source agents that remain visible in the transformed output. This metric is

central because it directly captures whether AI summarization maintains responsibility. If a source text identifies five relevant agents, for example, a state military, a sanctioning authority, a platform, a government ministry, and a humanitarian agency, the generated summary can be evaluated according to how many of those agents remain present and in what grammatical role. Preservation requires more than naming. An actor mentioned only as background, context, or object of diplomatic concern does not preserve the same degree of agency as an actor placed in subject position performing an action.

The ninth and central indicator is the Responsibility Loss Index, RLI. The RLI measures the decline in grammatical traceability between harm events and responsible agents across source texts and transformed outputs. It is not designed as a moral score. It is a linguistic index. Its purpose is to measure how much responsibility is lost when discourse moves from source event to machine-generated representation. A basic RLI can combine agent deletion rate, passive ratio, nominalization density, perpetrator subject loss, and agency-preservation decline. A more advanced RLI can add frame substitution, security-frame expansion, humanitarian-frame expansion, sanction-agent visibility, military-agent visibility, platform-agent visibility, and occupation-agent visibility.

A simple operational model can be stated as follows: responsibility loss increases when agent visibility decreases, passive or nominalized constructions increase, perpetrator subject ratio declines, and humanitarian or security frames replace direct action clauses. The index therefore captures not only the presence of passive voice, but the broader syntactic pattern through which responsibility becomes less nameable. This corresponds to the larger problem of objectivity as a formal effect. A text may appear neutral because it removes the grammatical positions where responsibility would otherwise be assigned (Startari, 2025c).

The RLI can be applied across four kinds of comparison. First, source-to-summary comparison compares a detailed source report with an AI-generated summary. Second, prompt-to-output comparison compares controlled prompts across models or settings. Third, headline-to-summary comparison examines whether headlines preserve agency more or less than automated summaries. Fourth, moderation-notice comparison examines whether platform enforcement language removes the censoring or moderating actor from

the sentence. Each comparison tests the same question: does the transformed text preserve or weaken the grammatical relation between harm and responsible agent?

For Palestine-related discourse, RLI can measure whether AI summaries preserve the agency of military forces, state institutions, occupation structures, diplomatic sponsors, weapons suppliers, platform authorities, or humanitarian organizations. A summary that states “civilians were killed during clashes” may register higher responsibility loss than a summary that states “military forces killed civilians during an operation,” assuming the source supports that attribution. A summary that states “content was removed after policy violations” may register higher platform-agent deletion than one that states “the platform removed Palestinian content under its moderation policy.” The issue is not whether one sentence is emotionally stronger. The issue is whether responsibility remains grammatically traceable.

For Iran-related discourse, RLI can measure sanctioned suffering. The relevant agent may not be a bombarding army, but a sanctioning state, financial authority, regulatory mechanism, banking network, export-control system, military alliance, or diplomatic coalition. A sentence such as “medicine shortages worsened amid tensions” may produce high responsibility loss if the source text attributes shortages to sanctions, overcompliance, import restrictions, or banking exclusion. The harm remains visible as shortage. The responsible structure becomes background. This is why Iran is methodologically important. It tests whether responsibility loss applies not only to direct violence, but also to indirect coercion mediated through law, finance, diplomacy, and security grammar.

The RLI also allows the paper to avoid unsupported claims about ideological intention. The analysis does not need to assert that an AI system deliberately protects a state, platform, or military actor. It only needs to show that the generated output reduces agency compared with the source or compared with an active formulation. This is what makes responsibility loss measurable. It shifts the analysis from motive to structure. The evidence lies in grammar: who appears as subject, what appears as action, what becomes nominalized, what is framed as crisis, and which agents disappear.

This method also distinguishes responsibility loss from conventional bias detection. Bias detection often focuses on sentiment, toxicity, stereotyping, slur presence, partisan language, or representational imbalance. Responsibility detection asks a different question: does the text preserve the grammatical conditions for assigning responsibility? A sentence can have low toxicity and high responsibility loss. A summary can avoid inflammatory language and still remove the agent. A platform notice can sound procedural and still hide the censoring institution. A humanitarian report can sound compassionate and still convert produced harm into abstract need. This difference is why responsibility loss must be treated as a separate analytic object.

The concept also connects to null subjects of power. In responsibility loss, the omitted agent is not an accidental absence. It is a missing grammatical position that affects how power can be reconstructed. When the agent of bombing, sanctioning, restricting, censoring, or displacing is absent, the sentence still carries the effects of power but not its source. This produces a null subject of responsibility: a position structurally required for accountability but absent from the surface of the text (Startari, 2025d). Responsibility loss therefore formalizes the link between syntactic absence and political consequence.

A minimal coding scheme for responsibility loss would classify each clause according to five questions. First, is a harm event present? Second, is a responsible agent present? Third, is the agent grammatically active, passive, backgrounded, or absent? Fourth, is the harm described through action, nominalization, humanitarian frame, security frame, or policy frame? Fifth, does the transformed output preserve, weaken, or remove the agency relation present in the source? These questions can be applied manually in a pilot study or computationally in a larger corpus through dependency parsing, named-entity recognition, semantic role labeling, and frame detection.

The method remains exploratory at this stage, but it is replicable. A pilot corpus can include controlled prompts asking multiple models to summarize the same conflict events; source paragraphs from news reports or humanitarian statements; platform moderation notices; government or policy statements; and AI-generated headline variants. Each text can be coded at clause level. The comparison should not ask whether a model “supports” one side.

It should ask whether the model preserves agency when describing harm. That question is narrower, more measurable, and more defensible.

The expected finding is not that all AI-generated conflict discourse erases responsibility. Some outputs may preserve agency accurately. Some may even increase attribution by clarifying actors that source texts obscure. The hypothesis is more specific: in politically sensitive, asymmetric, or moderation-prone contexts, AI-generated summaries will tend to reduce grammatical traceability when agency attribution increases reputational, diplomatic, legal, or platform risk. This tendency would appear as higher passive ratio, higher agent deletion rate, lower perpetrator subject ratio, higher nominalization density, and lower agency-preservation rate.

Responsibility loss therefore provides a bridge between linguistic theory and public accountability. It turns a political intuition into a testable linguistic problem. The intuition is simple: suffering remains visible, responsibility disappears. The analytic task is to show where, how often, and by which syntactic operations this disappearance occurs. The RLI gives that task a measurable form. It allows scholars to compare models, platforms, conflict domains, prompt conditions, and discourse genres without reducing the analysis to accusation or impression.

The purpose of this section has been to define responsibility loss as a measurable effect rather than a rhetorical claim. The next sections apply this framework to Palestine and Iran as distinct sites of asymmetric visibility. Palestine tests responsibility loss in relation to direct violence, occupation, bombardment, displacement, humanitarian catastrophe, and platform moderation. Iran tests responsibility loss in relation to sanctions, isolation, overcompliance, security framing, diplomatic pressure, and indirect civilian harm. In both cases, the central question remains the same: does machine-generated discourse preserve the grammar necessary to name responsibility?

IV. Palestine as Central Test Case

Palestine is the central test case for the humanitarian passive because it concentrates the core elements of responsibility loss in a single discourse field: civilian suffering, military action, occupation, displacement, humanitarian emergency, platform moderation, geopolitical sponsorship, legal controversy, and machine-mediated neutrality. The case does not matter only because Palestinian suffering is visible. It matters because this suffering is often visible under grammatical conditions that weaken the attribution of agency. The problem is therefore not simple erasure. It is asymmetric visibility: Palestinian civilians may appear as dead, displaced, injured, hungry, blocked, or affected, while the military, institutional, platform, diplomatic, and geopolitical actors connected to those conditions may become less grammatically available.

This distinction is decisive. A model-generated summary can mention destroyed homes, damaged hospitals, civilian deaths, food shortages, displacement, blocked aid, and humanitarian collapse while still reducing the visibility of the actors and structures that produced those conditions. The sentence may sound compassionate. It may sound neutral. It may even sound factually cautious. Yet the clause-level structure may have already converted political violence into humanitarian condition. This is the specific function of the humanitarian passive: suffering remains linguistically present, but responsibility is displaced from the grammatical center of the statement.

The Palestine case exposes this mechanism because it sits at the intersection of humanitarian discourse and politically contested attribution. International reporting on Gaza has repeatedly described large-scale civilian harm, attacks, displacement, destruction of infrastructure, food insecurity, and severe humanitarian need. OCHA's humanitarian updates, for example, continue to document civilian casualties and acute humanitarian conditions in Gaza, while UNRWA reports in 2026 still describe the crisis in Gaza and the occupied West Bank through indicators of aid access, displacement, hunger, shelter, and protection needs (OCHA, 2026; UNRWA, 2026). These sources establish that the humanitarian vocabulary is empirically necessary. The analytical question is different: when AI systems summarize such material, do they preserve the agents, policies, military actions, and institutional decisions that make the humanitarian vocabulary necessary?

The humanitarian passive appears when a clause preserves the damaged object but weakens the actor. Consider the contrast between an active formulation and a passive or agentless formulation. “Israeli forces struck a residential building” assigns a military actor, a verb of action, and an affected object. “A residential building was struck” preserves the harm but deletes the actor. “A residential building was damaged amid escalating violence” further weakens the action by substituting abstract context for agency. “Damage to residential infrastructure increased during the crisis” converts the action into nominalized condition. Each version may refer to the same event domain. Each version may preserve the existence of harm. But each step reduces the grammatical traceability of responsibility.

This transformation cannot be dismissed as a minor stylistic variation. In conflict discourse, grammatical subject position matters. The actor placed in subject position becomes the visible agent of the sentence. The object placed in subject position becomes the visible bearer of harm. When the harmed population repeatedly appears as grammatical subject while the responsible actor is omitted, backgrounded, or transformed into abstract context, the discourse produces a specific structure: victims are visible, perpetrators are optional. This extends the argument that passive voice in artificial intelligence language can remove agency from the sentence while preserving the appearance of neutral description (Startari, 2025a).

Palestine makes this problem especially acute because public discourse about Palestinian suffering is shaped by several competing pressures. First, there is the humanitarian pressure to describe civilian harm. Second, there is the diplomatic pressure to avoid explicit attribution when attribution is politically contested. Third, there is the platform pressure to moderate speech related to terrorism, extremism, hate speech, antisemitism, anti-Zionism, graphic violence, and political mobilization. Fourth, there is the institutional pressure to maintain neutrality. Fifth, there is the geopolitical pressure created by the role of dominant-power actors, including states, military allies, weapons suppliers, international institutions, and platform companies. These pressures do not need to operate through conspiracy. They can converge through grammar.

This is why Palestine is not merely a moral example. It is a syntactic test case. The question is not whether an AI system “cares” about Palestinian suffering, nor whether a model is

intentionally pro-Israel or anti-Palestinian. Those claims would require a different evidentiary standard. The narrower question is whether AI-generated summaries preserve or reduce responsibility when transforming source material. If a source text identifies military action, occupation policy, aid obstruction, platform removal, or state decision-making, and the generated summary converts these into "violence," "tensions," "crisis," "restrictions," "content removal," or "policy violations," then responsibility loss can be measured without assigning hidden intention.

This approach is stronger because it avoids reducing the analysis to motive. The evidence lies in the transformation. A source clause may state that an identifiable actor performed an action. The AI summary may retain the consequence but remove the actor. The relation between source and output becomes the empirical site of analysis. This is consistent with the broader theory of the grammar of objectivity: neutrality can be produced formally when impersonal structures remove the grammatical positions where agency and responsibility would otherwise appear (Startari, 2025b). In the Palestine case, machine neutrality can therefore become a grammar of low attribution.

Platform moderation intensifies the problem. Human Rights Watch reported in 2023 that Meta's policies and systems had "increasingly silenced voices in support of Palestine" on Instagram and Facebook after October 2023, and its report identified patterns of undue removal and suppression of Palestine-related content (Human Rights Watch, 2023). The factual relevance for this paper is not only that moderation disputes occurred. The syntactic relevance is that moderation language often appears in agentless or procedural forms: "content was removed," "visibility was reduced," "policy was violated," "accounts were restricted," or "posts were flagged." These formulations can make suppression appear as a technical event rather than an institutional action.

This platform layer matters because it gives the humanitarian passive a second site of operation. The first site is conflict description: civilians are killed, homes are destroyed, hospitals are damaged, aid is blocked. The second site is discourse governance: posts are removed, accounts are restricted, visibility is reduced, content is flagged. In both cases, the affected object remains visible. In both cases, the responsible agent can disappear. The same structure can therefore operate across military discourse and platform discourse. In

one case, the missing agent may be a military or state actor. In the other, it may be a platform, moderation system, policy team, automated classifier, or enforcement regime.

The Palestine case also reveals how humanitarian frames can unintentionally weaken political traceability. Humanitarian language is necessary when the object is civilian suffering. Terms such as “humanitarian crisis,” “civilian harm,” “displacement,” “food insecurity,” “collapsed infrastructure,” “medical shortages,” and “urgent aid needs” are not false or illegitimate. The problem appears when these terms replace the action structure that connects suffering to cause. “Humanitarian crisis” may become a summary label for conditions produced through bombing, blockade, displacement, aid restriction, or institutional decision. The label can be accurate at the level of consequence while incomplete at the level of responsibility.

This is the point at which responsibility loss differs from bias. A text may not contain hostile vocabulary. It may not contain dehumanizing language. It may not contain explicit misinformation. It may still reduce accountability through grammar. For example, “Gaza faces a worsening humanitarian crisis” may be less inflammatory than “Israeli military operations and restrictions have worsened civilian conditions in Gaza,” but the difference is not only tone. The first formulation makes Gaza the subject and crisis the predicate. The second formulation places identifiable actions and restrictions into a causal structure. The ethical issue is not which sentence sounds calmer. The analytic issue is which sentence preserves responsibility.

The same distinction applies to legal discourse. The International Court of Justice issued provisional measures in the case brought by South Africa against Israel under the Genocide Convention in January 2024, and later orders continued to address obligations concerning Palestinians in Gaza (International Court of Justice, 2024). The legal relevance for this article is not to resolve the merits of the case. The relevance is structural: legal responsibility depends on the ability to connect acts, omissions, obligations, actors, and protected groups. If AI-generated summaries of legal or humanitarian materials reduce actors into abstract crisis language, they may weaken the grammatical preconditions of legal intelligibility.

This does not mean that every Palestinian-related summary must contain accusatory language. It means that summary compression must not be confused with responsibility preservation. A shorter text can preserve agency if it retains the responsible actor and action. A longer text can erase agency if it expands humanitarian detail while removing causal attribution. Responsibility loss is therefore not a function of length alone. It is a function of grammatical transformation. The central question is whether the AI output preserves the relation between who acted, what was done, who was harmed, and under what institutional or military conditions.

The Palestine case allows this relation to be tested across multiple textual genres. News summaries can be compared with source articles. Humanitarian reports can be compared with AI-generated abstracts. Platform notices can be compared with user-facing moderation explanations. Legal statements can be compared with machine-generated briefings. Government statements can be compared with neutralized summaries. In each genre, the same coding questions apply: Is the harm event present? Is the responsible agent present? Is the agent grammatically active? Has the action become nominalized? Has causality become abstract? Has a humanitarian or security frame replaced an agency frame?

A basic coding example shows the mechanism:

Source formulation: “Israeli forces struck the area, killing civilians and damaging residential buildings.”

Low-loss summary: “Israeli forces struck the area, killing civilians and damaging residential buildings.”

Medium-loss summary: “Civilians were killed and residential buildings were damaged in the strike.”

High-loss summary: “Civilians were killed and buildings were damaged amid escalating violence.”

Very high-loss summary: “Civilian suffering and infrastructure damage worsened amid the crisis.”

The harm remains present across all versions. What changes is the visibility of the agent, the explicitness of the action, and the recoverability of responsibility. The final version may sound humanitarian, but it is grammatically detached from the responsible actor. That detachment is the object of measurement.

A comparable platform example shows the same structure:

Source formulation: “Meta removed Palestinian posts under its moderation policy.”

Low-loss summary: “Meta removed Palestinian posts under its moderation policy.”

Medium-loss summary: “Palestinian posts were removed under moderation rules.”

High-loss summary: “Content was removed for violating policy.”

Very high-loss summary: “Policy enforcement affected visibility during the conflict.”

Again, the event remains visible as removal, enforcement, or reduced visibility. The platform agent becomes less visible. The policy becomes the acting abstraction. The affected speech remains visible as moderated content, while the censoring or governing institution becomes grammatically unstable. This connects directly with the problem of source absence and algorithmic ethos: the procedure appears authoritative even when the responsible subject is no longer visible (Startari, 2025c).

Palestine also exposes the limits of sentiment-based analysis. A model could generate sympathetic language about Palestinian suffering while still producing high responsibility loss. It could describe grief, trauma, need, hunger, and loss without preserving the actor responsible for bombardment, displacement, or restriction. Conversely, a model could use dry, legalistic language but preserve agency accurately. The relevant measure is therefore not emotional intensity. It is syntactic traceability. A compassionate sentence can be politically evasive. A restrained sentence can be accountable. This is why the humanitarian passive must be analyzed formally.

Security framing introduces another layer. In Palestine-related discourse, terms such as “security operation,” “counterterrorism,” “militant activity,” “clashes,” “escalation,” “border violence,” or “regional instability” can organize agency unevenly. These terms

may be necessary in specific contexts, but they can also shift the interpretive center from harm-producing action to risk management. If a civilian casualty event is summarized primarily as part of “security operations” or “escalation,” the responsible action may become embedded in a frame where the dominant actor appears as responder rather than initiator. The question is not whether security vocabulary is always illegitimate. The question is whether it preserves or replaces agency.

The humanitarian passive therefore operates through a double conversion. First, political violence is converted into humanitarian consequence. Second, responsible action is converted into abstract context. The result is a sentence that can appear balanced while structurally lowering accountability. This is not unique to Palestine, but Palestine makes the pattern visible because the discourse is saturated with competing legal, moral, diplomatic, humanitarian, and platform constraints. These constraints make it possible for machine-generated summaries to avoid explicit falsehood while still producing responsibility loss.

The relevance of dominant-power actors must be stated precisely. The paper does not treat Israel and the United States as interchangeable agents. It treats them as analytically relevant dominant-power positions whose military, diplomatic, financial, legal, and platform-adjacent influence may shape the grammatical conditions under which violence and humanitarian suffering are described. In Palestine-related discourse, agency may involve Israeli military and state actors, Palestinian armed groups, international institutions, the United States, weapons suppliers, humanitarian agencies, platform companies, and legal bodies. The Responsibility Loss Index does not presuppose which actor must be blamed. It measures whether actors present in the source event remain grammatically traceable in the output.

This point is essential for academic defensibility. The paper does not claim that AI systems automatically hide one side’s responsibility and expose the other’s. It proposes a method for testing whether agency is preserved or weakened across transformations. If a model removes Palestinian armed-group agency from a relevant source event, that too would count as responsibility loss. If a model removes Israeli state or military agency from a relevant source event, that also counts as responsibility loss. If a model removes platform

agency from content governance, that also counts. The method is symmetrical. The expected political importance arises because responsibility loss may have asymmetric effects when it occurs in conditions of unequal power.

In this sense, Palestine is not used as a rhetorical shortcut. It is used as a rigorous stress test for the theory. If a model can describe Palestinian suffering without preserving the grammar of responsibility, then visibility has not solved accountability. If a platform can remove Palestine-related content while describing the act through agentless policy grammar, then moderation has not solved transparency. If a legal or humanitarian summary can foreground crisis while weakening agency, then neutrality has not solved attribution. The common structure is responsibility loss.

This section therefore establishes Palestine as the central empirical and conceptual test case for the humanitarian passive. It shows that the problem is not whether suffering appears. Suffering often appears. The problem is whether the sentence preserves the path from suffering to responsible agency. The next section extends the model to Iran, where responsibility loss operates through sanctioned suffering, financial restriction, military pressure, security framing, and indirect civilian harm. Palestine tests the grammar of direct violence and platform moderation. Iran tests the grammar of coercion without immediate physical contact. Together, they show that asymmetric visibility is not only a problem of what AI systems refuse to say. It is also a problem of what their grammar permits them to say without naming responsibility.

V. Iran and Sanctioned Suffering

Iran extends the model of the humanitarian passive beyond the grammar of direct violence. Palestine tests responsibility loss in relation to occupation, bombardment, displacement, humanitarian catastrophe, and platform moderation. Iran tests a different configuration: sanctioned suffering, financial isolation, overcompliance, diplomatic pressure, military threat, nuclear framing, and indirect civilian harm. The comparison is necessary because responsibility loss does not operate only when a visible military actor produces immediate physical destruction. It can also operate when harm is mediated through law, banking

systems, export controls, sanctions regimes, insurance restrictions, shipping constraints, humanitarian exemptions, and security discourse.

The Iranian case therefore requires a different analytical vocabulary. The central harm event is not always a bombed building or a displaced family, although military threat and regional escalation remain part of the discourse. More often, the harm appears as restricted access, shortages, inflation, medical scarcity, banking exclusion, blocked payments, supply-chain disruption, or social hardship. These effects are not grammatically identical to direct violence, but they are still politically attributable when the source material identifies sanctioning powers, financial mechanisms, regulatory systems, or overcompliance by firms and banks. Responsibility loss occurs when such structures are transformed into neutral conditions: "shortages worsened," "access became limited," "economic hardship increased," "medicine became scarce," or "conditions deteriorated amid tensions."

This is why Iran should not be treated as an identical case to Palestine. Palestine foregrounds the grammar of direct violence and occupation. Iran foregrounds the grammar of indirect coercion and sanctioned suffering. In Palestine-related summaries, responsibility loss may appear when military action becomes "damage," "clashes," "escalation," or "humanitarian crisis." In Iran-related summaries, responsibility loss may appear when sanctioning action becomes "economic pressure," "restricted access," "market disruption," "instability," or "regional tension." The formal mechanism is comparable, but the causal chain is different.

The empirical basis for this case is not speculative. Human Rights Watch reported in 2019 that broad United States sanctions had constrained Iran's ability to finance humanitarian imports, including medicines, producing serious hardship for ordinary Iranians and threatening the right to health (Human Rights Watch, 2019). United Nations experts later stated that unilateral sanctions and overcompliance posed serious threats to human rights and dignity in Iran, including through barriers to essential infrastructure, payments, humanitarian action, imports, technology, and social support (OHCHR, 2022). In 2023, UN experts also reported that thalassemia patients in Iran were being penalized by overcompliance with United States sanctions, which reduced access to vital medication

from abroad (OHCHR, 2023). These findings support the analytic premise that sanctioned suffering can involve identifiable institutional mechanisms, even when public discourse describes the result through neutral humanitarian language.

The humanitarian passive becomes visible when these mechanisms disappear from the grammatical surface. A source formulation might state: “United States sanctions and banking overcompliance restricted Iranian access to specialized medicines.” This formulation identifies sanctioning power, mechanism, affected population, and material consequence. A lower-agency formulation might state: “Iranian access to specialized medicines was restricted.” A more abstract formulation might state: “Medical shortages worsened amid economic pressure.” A still more neutral formulation might state: “Healthcare conditions deteriorated during the crisis.” Each version preserves some form of suffering. Each step makes the responsible structure less grammatically available.

The point is not that every summary of sanctions must contain a full geopolitical explanation. Compression is legitimate. The point is that compression can reduce responsibility. A summary that removes the sanctioning actor, banking mechanism, legal regime, or overcompliance pathway may remain factually compatible with humanitarian suffering while weakening the attribution of cause. This is the core of responsibility loss. The harm is not denied. It is described as condition rather than consequence. The grammar moves from “actor restricted access” to “access was restricted” to “shortages worsened” to “conditions deteriorated.” The syntactic object changes from political action to humanitarian state.

Iran is especially important because sanctions discourse often appears legally and administratively neutral. Sanctions are framed as pressure, compliance, deterrence, non-proliferation, counterterrorism, financial control, or diplomatic leverage. These frames may be analytically relevant, but they can also reduce the visibility of civilian effects. When the discourse states that “pressure increased on Tehran,” the state becomes the apparent target. When humanitarian reports describe shortages, patients, medicine, poverty, inflation, or blocked imports, civilians become visible as affected populations. The missing link is the grammatical relation between coercive measure and civilian harm.

Responsibility loss occurs when the text preserves both “pressure” and “suffering” but fails to connect them through an accountable causal structure.

This produces a specific form of asymmetric visibility. Iran as a state may appear as “regime,” “threat,” “nuclear risk,” “destabilizing actor,” “proxy sponsor,” or “regional challenge.” Iranian civilians may appear as patients, consumers, students, families, refugees, or populations facing hardship. Sanctioning powers may appear as defenders of security, enforcers of rules, or administrators of compliance. The grammar may therefore distribute agency unevenly: dominant actors appear as managers of risk, Iran appears as the source of risk, and civilians appear as affected bodies. The responsible structure of sanctioning, overcompliance, financial exclusion, and humanitarian restriction can become secondary.

The humanitarian passive in Iran-related discourse often operates through nominalization. “Sanctioning,” “restricting,” “blocking,” “excluding,” and “denying” become “sanctions,” “restrictions,” “barriers,” “shortages,” “exclusion,” or “limited access.” Nominalizations are not automatically evasive. They can be necessary for technical clarity. The problem appears when nominalization replaces the action structure needed for responsibility. “Banking restrictions limited humanitarian trade” preserves a causal relation. “Barriers to humanitarian trade persisted” weakens the agent. “Humanitarian trade faced challenges” weakens the mechanism. “Healthcare access deteriorated” reduces the entire chain to outcome.

Security framing intensifies this reduction. Iran-related discourse frequently passes through nuclear, military, regional, and counterterrorism vocabularies. These frames can make sanctioning appear as a response to threat rather than as an action with civilian consequences. A sentence such as “sanctions were imposed over nuclear concerns” may preserve the sanctioning act but frame it through security justification. A sentence such as “economic pressure increased amid nuclear tensions” removes the sanctioning actor and transforms policy into atmosphere. A sentence such as “medicine shortages worsened during regional instability” removes the sanctioning actor, the regulatory mechanism, and the policy chain. The result is not silence about suffering. It is suffering translated into security-conditioned background.

Overcompliance is one of the strongest sites for measuring responsibility loss because it introduces a chain of actors that can easily disappear. Sanctions may formally exempt humanitarian goods, but banks, insurers, suppliers, logistics firms, and pharmaceutical companies may avoid transactions because of regulatory risk, compliance uncertainty, secondary sanctions exposure, reputational risk, or payment barriers. The OHCHR's 2022 statement explicitly identified overcompliance as part of the problem affecting human rights in Iran, and its 2023 statement on thalassemia patients connected overcompliance with reduced access to vital medication (OHCHR, 2022, 2023). A high-responsibility formulation would preserve this chain. A low-responsibility formulation would collapse it into "medicines were unavailable" or "patients faced shortages."

This matters because humanitarian exemptions can become syntactically misleading if they are summarized without the practical obstacles that limit their effectiveness. A text may state that "food and medicine are exempt from sanctions," while another source documents payment barriers, insurance problems, overcompliance, or supply-chain disruptions. An AI-generated summary that preserves the exemption but removes the obstacles may reduce responsibility by making the regime appear formally humanitarian while leaving civilian harm unexplained. Conversely, a summary that preserves the obstacles but removes the sanctioning structure may describe suffering without attribution. Both cases can produce responsibility loss by separating legal form from material effect.

The Iran case also shows why responsibility loss cannot be measured only through passive voice. A sentence may be active and still produce low attribution if the wrong subject is selected. "Iran struggled with medicine shortages" is active, but it assigns grammatical agency to Iran's struggle rather than to sanctioning mechanisms, financial restrictions, or overcompliance. "Patients faced reduced access to treatment" is active, but it makes patients the subject of hardship rather than the object of policy-mediated restriction. This is why the Responsibility Loss Index must include victim subject ratio, perpetrator subject ratio, agency-preservation rate, security-frame frequency, and sanction-agent visibility, not only passive ratio.

Sanction-agent visibility is the key Iran-specific metric. It measures whether the actors imposing, administering, enforcing, financing, complying with, or overcomplying with

sanctions remain visible in the output. Relevant actors may include states, treasuries, ministries, banks, insurers, shipping firms, pharmaceutical suppliers, regulators, international bodies, and compliance departments. A summary that mentions "sanctions" but not sanctioning powers may preserve the policy object while weakening agency. A summary that mentions "economic hardship" but not sanctions may remove both policy and actor. A summary that mentions "regional tensions" may replace the entire causal chain with atmosphere.

A second Iran-specific metric is overcompliance visibility. This measures whether the output preserves the role of private and institutional actors whose risk-avoidance behavior amplifies the effects of sanctions. Overcompliance is analytically valuable because it blurs the line between state policy and corporate action. The responsible structure is distributed. It includes legal rules, enforcement threats, compliance interpretations, banking decisions, supplier withdrawals, and payment blockages. AI summaries may compress this distributed chain into a single humanitarian effect. The RLI must therefore detect not only whether "the United States" or "sanctions" appear, but whether the mediating mechanisms of harm remain visible.

A third metric is security-frame substitution. This measures whether humanitarian or civilian harm is reframed primarily through nuclear risk, threat, escalation, regional instability, or security concern. The issue is not that security frames are always false. The issue is whether they replace the causal grammar of civilian harm. For example, "sanctions restricted access to medical imports" and "medical access was affected amid nuclear tensions" are not equivalent. The first preserves a policy mechanism. The second converts that mechanism into geopolitical atmosphere. The harm remains. Responsibility becomes less recoverable.

This is where Iran connects directly to the theory of the grammar of objectivity. Machine-generated summaries may appear balanced when they avoid assigning explicit responsibility and instead use neutral diplomatic vocabulary. But neutrality can be produced by lowering agency. "Economic hardship has increased amid tensions between Iran and Western powers" appears cautious, but it may obscure specific sanctioning measures, banking constraints, and overcompliance mechanisms documented in source

material. The sentence is not necessarily false. Its problem is grammatical incompleteness. It simulates balance by abstracting agency. This follows the broader claim that objectivity in AI-generated language can operate as a formal effect rather than as an epistemic guarantee (Startari, 2025b).

Iran also extends the problem of ethos without source. When an AI-generated summary presents sanctioned suffering through impersonal or procedural language, it may sound institutionally authoritative even while the responsible subject is displaced. “Access was restricted,” “imports were affected,” “transactions were limited,” and “patients faced shortages” sound factual and neutral. Yet the source of restriction, limitation, and shortage may be missing. The output acquires credibility through detached tone, while responsibility becomes less traceable. This is the same structural problem identified in algorithmic ethos: authority appears without an accountable source (Startari, 2025c).

The case also clarifies why the paper does not need to defend the Iranian state. The object is not Iran’s internal regime, nor an endorsement of its policies. The object is the grammar through which civilian suffering under external pressure becomes describable without preserving the agents and mechanisms of that pressure. This distinction is necessary for academic rigor. The analysis can recognize that Iran is represented in security discourse through nuclear, military, and regional frames while still asking whether those frames erase or weaken the grammatical connection between coercive measures and civilian harm. The paper analyzes syntax, not allegiance.

A controlled prompt experiment could test this directly. The same source paragraph could state that sanctions, banking restrictions, and overcompliance have reduced access to specialized medicines in Iran. Several models could then be asked to summarize the paragraph in neutral language, humanitarian language, policy language, and security language. Outputs could be coded for sanction-agent visibility, overcompliance visibility, patient subject ratio, nominalization density, passive ratio, and security-frame substitution. The expected pattern is not that every output will erase responsibility. The testable question is whether certain prompt conditions or model defaults reduce agency while preserving suffering.

A pilot example would look like this:

Source formulation: “United States sanctions and banking overcompliance restricted Iranian patients’ access to specialized medicines.”

Low-loss summary: “United States sanctions and banking overcompliance restricted Iranian patients’ access to specialized medicines.”

Medium-loss summary: “Iranian patients’ access to specialized medicines was restricted by sanctions and banking overcompliance.”

High-loss summary: “Iranian patients faced restricted access to specialized medicines.”

Very high-loss summary: “Medical shortages worsened amid economic pressure and regional tensions.”

The movement from source formulation to very high-loss summary shows the mechanism. The civilian harm remains visible. The sanctioning actor disappears. The banking mechanism disappears. Overcompliance disappears. The action becomes shortage. The causal chain becomes atmosphere. This is responsibility loss under sanctioned suffering.

The same pattern can occur in diplomatic summaries:

Source formulation: “The United States expanded sanctions that limited Iran’s ability to process international payments for humanitarian imports.”

Low-loss summary: “The United States expanded sanctions that limited Iran’s ability to process payments for humanitarian imports.”

Medium-loss summary: “Sanctions limited Iran’s ability to process payments for humanitarian imports.”

High-loss summary: “Iran faced payment barriers affecting humanitarian imports.”

Very high-loss summary: “Humanitarian imports were affected amid diplomatic tensions.”

Again, the issue is not whether the final sentence is impossible. The issue is that it has lost the responsible actor, reduced the policy mechanism, and converted coercive action into diplomatic context.

The Iran case also makes visible the difference between responsibility loss and factual caution. AI systems may avoid naming a sanctioning power because a source is ambiguous, contested, or incomplete. In such cases, lower attribution may be justified. The RLI should not penalize outputs for refusing unsupported claims. It should penalize outputs that remove agents already present in reliable source material. This distinction protects the method from becoming a demand for ideological accusation. Responsibility detection is not accusation generation. It is agency preservation when the source supports agency.

This distinction also protects the article from collapsing into activism. The claim is not "AI censors Iran" or "AI protects the United States." The defensible claim is narrower: AI-generated discourse may reproduce asymmetric patterns of attribution when dominant-power action is described through passive, nominalized, humanitarian, or security-oriented grammar. Iran is a strong test case because sanctioning power is often legally formalized, administratively distributed, and publicly justified through security frames. Those features make agency easy to dilute without explicitly denying the existence of harm.

The methodological contribution of the Iran section is therefore to expand responsibility loss beyond battlefield discourse. Civilian suffering can be produced by bombs, but it can also be produced or intensified by financial systems, sanctions regimes, compliance decisions, trade barriers, diplomatic isolation, and security policies. If the humanitarian passive only applied to direct violence, it would remain useful but limited. By applying it to Iran, the paper shows that responsibility loss can occur wherever harm is described without preserving the agents and mechanisms that structure it.

This extension also prepares the broader series. Future papers can examine how AI-generated discourse represents other societies under asymmetric pressure: Yemen, Sudan, Iraq, Syria, Lebanon, Venezuela, Cuba, Afghanistan, Haiti, and migrant populations at borders. The category does not require identical histories. It requires a comparable syntactic condition: dominated or pressured populations remain visible as suffering, crisis,

risk, or humanitarian need, while dominant actors and institutional mechanisms become grammatically diluted. Iran provides the bridge between direct violence and systemic coercion.

The central conclusion of this section is that sanctioned suffering produces a distinctive form of humanitarian passive. The suffering is often indirect, mediated, bureaucratic, financial, and legally framed. For that reason, the responsible agent is easier to remove. The sentence does not need to deny hardship. It only needs to describe hardship without preserving the sanctioning chain. “Patients faced shortages” is not the same as “sanctions and overcompliance restricted patients’ access to medicine.” The first preserves suffering. The second preserves responsibility.

Iran therefore strengthens the paper’s main claim. AI ethics and platform governance cannot restrict responsibility detection to overt violence, hate speech, or misinformation. They must also analyze how machine-generated summaries handle indirect coercion, legal restrictions, financial exclusion, and security-framed civilian harm. The question remains the same as in Palestine, but the mechanism changes. Does the grammar preserve the path from suffering to responsible agency? In Iran-related discourse, that path often runs through sanctions, banks, compliance systems, security narratives, and dominant-power policy. When those links disappear, suffering remains visible, but responsibility disappears.

VI. Platform Moderation and Machine Neutrality

Platform moderation is one of the clearest sites where the humanitarian passive becomes institutional. In military or humanitarian discourse, the missing agent may be a state, army, sanctioning power, or diplomatic actor. In platform discourse, the missing agent may be a company, policy team, automated classifier, trust-and-safety unit, content reviewer, government requester, ranking system, or enforcement pipeline. The surface grammar is often procedural: content is removed, posts are flagged, visibility is reduced, accounts are restricted, warnings are applied, appeals are denied, and policies are violated. These sentences describe outcomes, but they often make the acting institution difficult to locate.

This matters because moderation is not only a technical process. It is a form of discourse governance. Platforms decide what remains visible, what is demoted, what is removed, what is labeled, what is monetized, what is searchable, what is recommended, and what becomes risky to say. When these decisions are described in agentless form, governance appears as procedure rather than action. “Content was removed” does not assign the same responsibility as “the platform removed the content.” “Visibility was reduced” does not preserve the same agency as “the ranking system reduced the visibility of the post.” “Policy was violated” shifts attention from the enforcing actor to the violated rule. The grammatical object is the user’s content. The acting institution becomes abstract.

This is the platform-specific form of responsibility loss. It does not require proving that a company intended to suppress a political position. It requires comparing the enforcement act with the language used to describe it. If a platform removes, demotes, labels, restricts, or flags content, but its user-facing explanation states only that “content was removed” or “policy was violated,” then the grammar has displaced the enforcing actor. The platform remains institutionally active while appearing syntactically absent. This follows the broader problem of passive voice in artificial intelligence language, where agency can disappear from the grammatical surface while the effect of action remains visible (Startari, 2025a).

Palestine-related moderation makes this problem especially visible. Human Rights Watch reported in December 2023 that Meta’s policies and systems had increasingly silenced

voices in support of Palestine on Instagram and Facebook after October 2023, and its full report documented patterns of removal, suppression, and restriction of Palestine-related content (Human Rights Watch, 2023). The factual significance of that report is that moderation disputes were not merely anecdotal. The syntactic significance is that the language of moderation can convert institutional suppression into technical condition: posts "were removed," accounts "were restricted," content "was demoted," or visibility "was affected." The harm to speech remains describable. The censoring structure becomes grammatically diluted.

The Oversight Board's 2024 policy advisory opinion on Meta's treatment of the Arabic term "shaheed" provides a second example. The Board recommended that Meta end its blanket ban on the use of "shaheed" when referring to individuals designated as dangerous, and it urged Meta to adopt a more context-sensitive analysis rather than treating the term as automatically violating in broad circumstances (Oversight Board, 2024). The importance of this case is not only lexical. It shows how a single moderation category can scale into repeated content removal when automated or rule-based enforcement treats a culturally and politically complex term as a fixed risk signal.

The shaheed case also exposes the relation between lexical risk and syntactic responsibility. A moderation system may describe a removal as the result of a policy violation. Yet the actual chain may involve a platform rule, a dangerous-organization designation, an automated classifier, a language-specific enforcement model, a risk threshold, a human review decision, and a public safety rationale. If the final notice says only "your post was removed for violating policy," the user receives an outcome without a reconstructable chain. The responsible architecture is compressed into "policy." This is not only a transparency problem. It is a grammatical problem: the action is attributed to an abstract rule rather than to the institution that designed, deployed, and enforced it.

Machine neutrality intensifies the issue because automated moderation systems often appear less political than human decisions. The interface suggests procedure. The notice suggests rule application. The classifier suggests scale. The appeal system suggests due process. Yet every part of that structure depends on prior institutional choices: what counts as dangerous, what counts as praise, what counts as graphic violence, what counts as hate

speech, what counts as coordinated harm, what languages receive better moderation resources, what content is demoted rather than removed, and what error rates are tolerated. When those choices are linguistically converted into neutral enforcement outcomes, responsibility becomes difficult to assign.

The Global Network Initiative noted in 2025 that platforms are increasingly integrating AI and AI-based tools into content moderation, creating risks to fundamental rights, including freedom of expression and privacy, and raising questions about human rights due diligence in automated governance (Global Network Initiative, 2025). This is directly relevant to the humanitarian passive because AI moderation does not merely classify speech. It shapes the grammatical conditions under which enforcement becomes intelligible. The more enforcement is automated, the easier it becomes to describe suppression as system behavior rather than institutional action.

Platform moderation therefore has two overlapping grammars. The first is the grammar of enforcement: content was removed, accounts were restricted, posts were flagged, reach was reduced, monetization was limited, warning screens were applied. The second is the grammar of justification: policy was violated, safety was protected, community standards were enforced, harmful content was reduced, risk was mitigated, compliance was required. Both grammars can be legitimate in specific contexts. The problem arises when they replace agency. “Community standards were enforced” does not say who enforced them, how they were interpreted, what system applied them, what error rate was accepted, what appeal path existed, or which populations were disproportionately affected.

This makes platform moderation a privileged site for measuring responsibility loss. The relevant actor is often known: the platform. The relevant action is often known: removal, demotion, restriction, labeling, flagging, or suspension. The relevant affected object is often known: a post, account, image, phrase, video, link, user, group, or community. The question is whether the public-facing or AI-generated description preserves that relation. A high-responsibility formulation states: “Meta removed the post under its Dangerous Organizations and Individuals policy.” A lower-responsibility formulation states: “The post was removed under platform rules.” A still lower-responsibility formulation states:

"The content violated policy." The final version converts institutional action into rule-object relation.

This transformation is structurally similar to conflict summarization. In conflict discourse, "forces bombed a building" becomes "a building was damaged." In moderation discourse, "the platform removed a post" becomes "content was removed." In both cases, the affected object remains visible. In both cases, the responsible agent becomes optional. In both cases, the text may appear neutral because it avoids naming the actor. This connection is central to the paper: the humanitarian passive is not limited to descriptions of physical violence. It also appears in the governance of speech about that violence.

The Responsibility Loss Index can be adapted to platform moderation through specific indicators. Platform-agent visibility measures whether the platform remains grammatically present as the actor. Enforcement-agent visibility measures whether the specific moderation system, human reviewer, classifier, policy team, or automated rule remains visible. Policy-as-agent frequency measures how often abstract rules appear as the acting subject. Content-subject ratio measures how often the affected content appears as the grammatical subject. Appeal-trace visibility measures whether the user can reconstruct the review path. Demotion visibility measures whether reduction in reach is stated directly or hidden behind vague wording such as "distribution may be limited." These indicators make the analysis empirical rather than rhetorical.

A basic moderation coding sequence can show the mechanism:

Source event: "The platform removed Palestinian posts under its automated moderation system."

Low-loss notice: "The platform removed Palestinian posts under its automated moderation system."

Medium-loss notice: "Palestinian posts were removed under automated moderation rules."

High-loss notice: "Content was removed for violating policy."

Very high-loss notice: "Policy enforcement affected content visibility."

The final version is not necessarily false. It is grammatically evasive. The affected object remains visible as content. The enforcement action remains visible as policy. The platform disappears. The automation disappears. The political category of the affected speech disappears. The user receives procedure without institutional traceability.

The same structure can apply to content demotion:

Source event: “The ranking system reduced the visibility of Palestine-related posts after automated risk classification.”

Low-loss description: “The ranking system reduced the visibility of Palestine-related posts after automated risk classification.”

Medium-loss description: “Palestine-related posts had their visibility reduced after automated review.”

High-loss description: “Visibility was reduced after risk classification.”

Very high-loss description: “Distribution was limited to maintain safety.”

This sequence shows how machine neutrality can transform a governance decision into a safety condition. The final phrase appears procedural and protective, but it removes the actor, the affected category, the mechanism, and the evaluative threshold. It converts moderation into atmosphere.

The shaheed case is especially useful because it shows how a term can become a moderation object without the broader cultural, religious, linguistic, and political context remaining visible. If a post is removed because a term was classified as praise for a designated dangerous individual, the explanation must preserve more than the fact of policy violation. It must preserve the interpretive chain. Otherwise, the user sees only an outcome. The Oversight Board’s recommendation to end the blanket ban matters because it challenged the assumption that the term could be treated as uniformly equivalent to prohibited praise across contexts (Oversight Board, 2024). The linguistic issue is that automated moderation can convert polysemy into risk, and then convert risk into removal.

This links platform moderation to the grammar of objectivity. A moderation notice can appear objective because it avoids political vocabulary and uses procedural language. “This content was removed for violating our policy” sounds neutral. Yet the neutrality comes from grammar and institutional abstraction. The statement does not say who judged the content, what model or reviewer made the decision, what threshold was applied, what context was considered, or whether comparable content in another language was treated similarly. The formal style creates an ethos of impersonal correctness. This connects with the theory of objectivity as a grammatical effect in AI-generated language (Startari, 2025b).

The same point applies to ethos without source. Platform notices often speak with institutional authority while hiding the source of judgment. The user is told that a rule was violated, but not necessarily how the judgment was produced. The platform appears as an impersonal authority, and the decision appears as procedural consequence. In algorithmic systems, this structure simulates credibility without exposing the responsible subject or evaluative pathway (Startari, 2025c). The humanitarian passive becomes a moderation passive: suppression remains visible as outcome, while the suppressing agent becomes source-less.

This does not mean that all moderation is illegitimate. Platforms must remove some content, reduce the spread of certain materials, and enforce rules against genuine incitement, harassment, threats, doxxing, exploitation, and coordinated manipulation. The argument is narrower. When moderation decisions are necessary, the grammar of explanation should preserve responsibility rather than obscure it. A legitimate enforcement act can still be explained in a way that names the enforcing actor, the applied rule, the relevant evidence, the appeal path, and the role of automation. Responsibility preservation does not require abandoning safety. It requires making safety governance traceable.

This distinction protects the paper from a false binary. The issue is not “moderation versus free speech” in generic terms. The issue is whether moderation remains accountable when automated systems mediate enforcement at scale. A platform can have a legitimate safety policy and still produce responsibility loss through its explanatory grammar. A model can help enforce policy and still reduce agency in the way enforcement is described. A notice can be legally careful and still grammatically evasive. The paper therefore analyzes the

syntax of moderation rather than assuming that all moderation is censorship or all moderation is neutral.

Platform governance also complicates the identity of the responsible agent. Responsibility may be distributed across engineers, policy teams, classifiers, contractors, executives, regulators, advertisers, governments, human reviewers, and automated systems. This distribution makes agent preservation harder, but not impossible. A high-quality explanation can identify the relevant level of agency: “This post was removed by automated enforcement under Policy X, with appeal available through Process Y.” That sentence does not reveal every internal detail, but it preserves institutional and procedural traceability. By contrast, “This post violated policy” makes the post the apparent actor and hides the enforcement chain.

The platform case also shows why responsibility loss must include policy-as-agent frequency. In moderation discourse, policies often become grammatical actors. “Our policy requires removal,” “Community Standards do not allow this content,” “safety systems detected a violation,” “enforcement actions were taken,” or “distribution was limited under policy.” These forms attribute action to rules, standards, systems, or safety. The institution becomes the designer of the rule but not the actor in the sentence. This is a syntactic displacement of responsibility. The platform governs through rules, then describes the rule as if it governed by itself.

This phenomenon is not limited to Palestine-related content. It can apply to Iranian speech, anti-imperial speech, war documentation, police violence, migrant testimony, climate protest, labor organizing, and other politically sensitive domains. Palestine is central because the moderation record is especially documented and because the conflict generates intense pressure around safety, antisemitism, anti-Zionism, violent content, terrorism designations, and human rights documentation. Iran is relevant because sanctions, designated entities, state-affiliated media labels, cybersecurity narratives, and national security frames can also shape what becomes visible or suppressible. The broader category is speech produced by or about societies under asymmetric power.

This section therefore reframes platform moderation as part of the same responsibility problem studied in conflict summaries. The platform does not only mediate the circulation of discourse. It also produces its own discourse, especially through notices, labels, explanations, warnings, appeals, transparency reports, policy updates, and automated summaries. Those texts can either preserve or weaken responsibility. If they preserve agency, they support accountability. If they hide the enforcing actor behind policy grammar, they reproduce the humanitarian passive at the level of speech governance.

The methodological consequence is clear. Any study of responsibility loss must include both primary conflict discourse and platform governance discourse. It is not enough to ask whether AI summaries of Gaza, Iran, or sanctions preserve agency. It is also necessary to ask whether the systems moderating speech about those topics preserve their own agency. The question repeats across levels: Who acted? What was done? Who was affected? What rule or mechanism mediated the act? What institution can be held accountable? If the grammar cannot answer these questions, responsibility has been lost.

Platform moderation and machine neutrality therefore complete the central triangle of the paper. Palestine shows responsibility loss in direct violence and humanitarian catastrophe. Iran shows responsibility loss in sanctioned suffering and indirect coercion. Platform moderation shows responsibility loss in the governance of speech about both. The same syntactic pattern appears across the triangle: suffering, restriction, or suppression remains visible, while the responsible agent becomes optional. This is why responsibility detection must become part of AI ethics, platform governance, and computational discourse analysis. It is not enough to know whether the system labels content as safe or unsafe. It is necessary to know whether the system preserves the grammar of accountability.

VII. Conclusion: Naming Responsibility in Machine Discourse

The central problem examined in this article is not the disappearance of suffering from AI-generated conflict discourse. Suffering often remains visible. Civilians are killed, displaced, injured, deprived, restricted, silenced, or affected. Homes are destroyed. Hospitals are damaged. Food, water, medicine, shelter, and safety become scarce. Posts are removed. Accounts are restricted. Visibility is reduced. Access is limited. The language of harm is present. The problem is that the grammar connecting harm to responsible agency may disappear.

This article has defined that pattern as the *humanitarian passive*: a machine-mediated syntactic structure through which civilian suffering remains describable while responsibility becomes grammatically optional. The humanitarian passive does not operate primarily by denying harm. It operates by weakening attribution. It can preserve the victim, the damage, the shortage, the crisis, or the moderation outcome while reducing the visibility of the actor, institution, policy, military structure, sanctioning power, platform, or decision chain that produced it. Its effect is therefore not silence, but responsibility loss.

Responsibility loss is the measurable decline of grammatical traceability between harm and responsible agency. It appears when active clauses become agentless passives, when actions become nominalizations, when causes become abstract conditions, when military action becomes escalation, when sanctions become hardship, when censorship becomes moderation, and when platform enforcement becomes policy violation. The article has argued that this effect should be treated as a distinct object of linguistic, legal, and political analysis. It cannot be reduced to bias, misinformation, toxicity, hate speech, or sentiment. A text can avoid all of those problems and still remove responsibility from its grammar.

This distinction is the core contribution of the paper. Existing AI governance debates often ask whether a model produces harmful content, false content, biased content, extremist content, or discriminatory content. Those questions remain necessary, but they are incomplete. The humanitarian passive shows that a model can produce cautious, humanitarian, procedurally neutral, or emotionally sympathetic language while still weakening accountability. The relevant question is not only whether the text describes

suffering. The relevant question is whether the text preserves the grammatical conditions under which responsibility can still be named.

Palestine has served as the central test case because it exposes this problem with unusual intensity. Palestinian suffering can be highly visible in AI-generated discourse, humanitarian reports, summaries, platform debates, and public-facing language. Yet visibility does not guarantee attribution. Civilian death, displacement, destroyed infrastructure, famine risk, blocked aid, damaged hospitals, and platform suppression can all be described through agentless, passive, nominalized, or humanitarianized grammar. The result is a discourse in which suffering remains central while the responsible agents and structures become less stable. This is not a claim that every summary hides agency. It is a claim that agency preservation must be measured rather than assumed.

Iran extends the same model to sanctioned suffering and indirect coercion. In Iran-related discourse, harm may appear through shortages, restricted access, financial isolation, blocked payments, medical scarcity, overcompliance, economic hardship, diplomatic pressure, nuclear tension, and regional instability. These forms of suffering do not always resemble battlefield violence, but they can still be politically attributable. When sanctioning powers, banking systems, compliance regimes, insurance restrictions, export controls, or military threats disappear from the grammatical surface, civilian harm remains visible as condition while responsibility becomes less recoverable. Iran therefore demonstrates that responsibility loss applies not only to direct violence, but also to legal, financial, diplomatic, and security-mediated coercion.

Platform moderation completes the argument by showing that responsibility loss also appears in the governance of speech itself. When posts are removed, accounts are restricted, visibility is reduced, or content is flagged, the platform may describe the outcome through procedural grammar: “content was removed,” “policy was violated,” “distribution was limited,” or “safety systems acted.” These formulations may be formally correct, but they can obscure the institutional actor, automated system, policy team, classifier, appeal path, or enforcement pipeline involved in the decision. The same structure reappears: the affected object is visible, but the governing agent is grammatically diluted.

In this sense, platform moderation is not external to the humanitarian passive. It is one of its strongest institutional forms.

The paper has therefore proposed a shift from bias detection to responsibility detection. Bias detection asks whether a model favors one group, reproduces stereotypes, uses toxic language, or generates partisan asymmetry. Responsibility detection asks whether the text preserves the relation between harm, action, agent, institution, and consequence. A model can be low in toxicity and high in responsibility loss. A moderation notice can be procedurally polite and syntactically evasive. A humanitarian summary can be compassionate and politically under-attributed. A diplomatic formulation can be balanced in tone and structurally incomplete in agency. These distinctions require a different analytic instrument.

That instrument is the Responsibility Loss Index, RLI. The RLI measures the decline in grammatical traceability between harm events and responsible agents across source texts, AI summaries, headlines, reports, platform notices, policy statements, and controlled prompts. Its indicators include passive ratio, agent deletion rate, victim subject ratio, perpetrator subject ratio, nominalization density, humanitarian-frame frequency, security-frame frequency, agency-preservation rate, sanction-agent visibility, military-agent visibility, platform-agent visibility, and policy-as-agent frequency. The purpose of the RLI is not to assign moral blame automatically. Its purpose is to make responsibility loss observable, comparable, and replicable.

The methodological principle is simple: where the source preserves agency, the summary should not erase it without justification. If a source identifies a military actor, sanctioning power, platform authority, police unit, government office, bank, regulator, or institutional decision-maker, the transformed output should be evaluated according to whether that actor remains grammatically visible. If the output preserves harm but removes the agent, responsibility loss has occurred. If the output preserves the agent but shifts it from active subject to backgrounded context, responsibility may be weakened. If the output replaces action with nominalized condition, the grammar may move from accountability to description.

This approach avoids the weakness of intention-based accusation. The paper does not need to prove that an AI system intends to protect a state, hide a platform, defend a government, or suppress a population. It measures a formal effect. The evidence lies in the transformation from source to output: who disappears, what becomes passive, what becomes nominalized, what becomes framed as crisis, what becomes framed as security, and what remains grammatically available for accountability. This is consistent with prior work on passive voice, syntactic sovereignty, objectivity, ethos without source, and null subjects of power, where the disappearance of agency is treated as a structural feature of machine-mediated language rather than a merely stylistic variation (Startari, 2025a, 2025b, 2025c, 2025d, 2025e).

The broader theoretical claim is that AI-generated discourse can redistribute political visibility through syntax. Dominated, occupied, sanctioned, or moderated populations may remain highly visible, but primarily as victims, crises, risks, targets, affected users, humanitarian populations, or objects of policy enforcement. Dominant actors may remain active in the world while becoming less active in grammar. This is asymmetric visibility. It does not mean that the subordinated are unseen. It means that they are seen under categories that weaken their political relation to responsible agents. They appear as suffering. The agents of that suffering become optional.

This makes the humanitarian passive a political form, but not in the sense of partisan propaganda. It is political because responsibility itself is a grammatical relation before it becomes a legal or moral argument. To hold an actor accountable, discourse must preserve a minimum structure: someone acted, something was done, someone was harmed, and the relation between action and harm can be reconstructed. When grammar removes or weakens that structure, accountability becomes harder to name, harder to contest, and harder to document. A sentence can therefore participate in political depoliticization even when its tone is neutral.

The article also shows why humanitarian language must be handled carefully. Humanitarian frames are indispensable when describing civilian suffering. They record harm, need, urgency, vulnerability, displacement, hunger, injury, and deprivation. The problem is not humanitarian vocabulary itself. The problem is humanitarian vocabulary

without agency. When "humanitarian crisis" replaces the causal grammar of bombardment, sanctioning, blockade, restriction, displacement, or platform suppression, the discourse shifts from responsibility to relief. The victim remains central, but the perpetrating or governing structure becomes peripheral.

The same applies to security language. Security frames can be legitimate in specific contexts, but they can also convert coercion into risk management. Military action becomes response. Sanctions become pressure. Occupation becomes security administration. Platform suppression becomes safety enforcement. Civilian harm becomes collateral condition. The question is not whether security vocabulary should be prohibited. The question is whether it preserves the responsible agent and the causal chain. If it does not, security framing becomes another path to responsibility loss.

The legal implications are direct. Law depends on attribution. Duties, violations, remedies, obligations, intent, negligence, proportionality, complicity, command responsibility, and institutional accountability all require some relation between actor and act. If AI-generated summaries of conflict, sanctions, or moderation repeatedly weaken that relation, they may affect not only public understanding but also the documentary environment in which responsibility is reconstructed. The issue is not that AI outputs replace courts, reports, or investigations. The issue is that AI-mediated discourse increasingly shapes what appears legible, searchable, quotable, summarizable, and publicly intelligible.

For AI ethics, the implication is equally direct. Safety systems should not be evaluated only by whether they reduce harmful content. They should also be evaluated by whether they preserve accountability in politically sensitive discourse. A system that avoids inflammatory language by removing agency may be safer at the level of surface tone and weaker at the level of responsibility. A moderation system that removes content but describes enforcement through agentless policy grammar may reduce risk for the platform while reducing traceability for the user. A summarization system that compresses conflict reporting into humanitarian crisis language may improve readability while weakening attribution.

The practical recommendation is not to force models into accusation. It is to require agency-preserving summarization when the source supports agency. A responsible AI summary should not invent perpetrators, but it should not remove named agents already present in the source. It should distinguish unknown agents from omitted agents. It should preserve sanctioning mechanisms when discussing sanctioned suffering. It should preserve platform agency when discussing moderation. It should avoid converting every harm event into crisis language when the source contains action, actor, and causal structure. It should treat responsibility as a feature to preserve, not as a risk to neutralize automatically.

The same principle applies to platform notices. A responsible moderation notice should identify the enforcing actor, the applied policy, whether automation was involved, what content element triggered enforcement, whether context was considered, what appeal path exists, and whether visibility was reduced, removal occurred, or distribution was limited. This does not require exposing all internal systems. It requires avoiding agentless procedural grammar that hides institutional action. “We removed this post under Policy X after automated review” preserves more responsibility than “content was removed for policy violation.” The difference is grammatical, but the consequence is institutional.

This paper therefore establishes the foundation for the broader series *Grammars of Asymmetric Visibility: AI, Imperial Power, and the Syntax of Responsibility*. The series begins with the humanitarian passive because this pattern captures the central paradox of AI-mediated political discourse: suffering can remain visible while responsibility disappears. Future work can extend the framework to broader asymmetric visibility, Iran-specific security grammar, censorship without a censor, digital dehumanization, military optimization language, and algorithmic empire. The common object across the series is not merely representation. It is the syntactic distribution of agency under conditions of unequal power.

The final claim is deliberately narrow and structurally strong. AI-generated conflict discourse does not need to lie in order to weaken responsibility. It does not need to erase victims in order to depoliticize suffering. It does not need to invent facts in order to reduce accountability. It only needs to preserve consequences while weakening agency. That is

the danger of the humanitarian passive. It can speak of suffering fluently while making responsibility grammatically difficult to recover.

The ethical and analytical task is therefore to ask a new question. Not only: does the system mention the victims? Not only: does the system avoid hate speech? Not only: does the system sound neutral? The necessary question is: does the system preserve the grammar needed to name responsibility?

Suffering remains visible. Responsibility disappears. The task of responsibility detection is to make that disappearance measurable.

References

Human Rights Watch. (2019). *Maximum pressure: U.S. economic sanctions harm Iranians' right to health*. Human Rights Watch. https://www.hrw.org/report/2019/10/29/maximum-pressure/us-economic-sanctions-harm-iranians-right-health

Human Rights Watch. (2023). *Meta's broken promises: Systemic censorship of Palestine content on Instagram and Facebook*. Human Rights Watch. https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and

International Court of Justice. (2024). *Application of the Convention on the Prevention and Punishment of the Crime of Genocide in the Gaza Strip, South Africa v. Israel: Provisional measures*. International Court of Justice. https://www.icj-cij.org/case/192/provisional-measures

Office of the United Nations High Commissioner for Human Rights. (2022). *Iran: Unilateral sanctions and overcompliance constitute serious threat to human rights and dignity, UN expert says*. OHCHR. https://www.ohchr.org/en/press-releases/2022/05/iran-unilateral-sanctions-and-overcompliance-constitute-serious-threat-human

Office of the United Nations High Commissioner for Human Rights. (2023). *Iran: Over-compliance with unilateral sanctions affects thalassemia patients, say UN experts*. OHCHR. https://www.ohchr.org/en/press-releases/2023/02/iran-over-compliance-unilateral-sanctions-affects-thalassemia-patients-say

Oversight Board. (2024). *Referring to designated dangerous individuals as "shaheed"*. Oversight Board. https://www.oversightboard.com/decision/pao-lopp03uk/

United Nations Office for the Coordination of Humanitarian Affairs. (2026). *Humanitarian situation report: Gaza Strip, 23 April 2026*. OCHA. https://www.ochaopt.org/content/humanitarian-situation-report-23-april-2026

Global Network Initiative. (2025). *Navigating AI moderation and the risks to free expression*. Global Network Initiative. https://globalnetworkinitiative.org/navigating-ai-moderation-and-the-risks-to-free-expression/

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). *Ethos without source: Algorithmic identity and the simulation of credibility*. SSRN. https://doi.org/10.2139/ssrn.5313317

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

Startari, A. V. (2025e). *Null subjects of power: The politics of absence in executable language*. SSRN. https://doi.org/10.2139/ssrn.5462235

Startari, A. V. (2025f). *Executable power: Syntax as infrastructure in predictive societies*. Zenodo. https://doi.org/10.5281/zenodo.15754714

Corpus

Corpus title

Responsibility Loss Corpus, RLC-1

Corpus function

The corpus is designed to measure whether AI-generated discourse preserves or weakens the grammatical traceability between civilian harm and responsible agents.

Corpus type

Mixed corpus:

1. Source texts.
2. AI-generated summaries.
3. Platform moderation texts.
4. Controlled prompt outputs.
5. Comparative rewritten variants.

Target size for pilot version

Minimum pilot: 120 source-output pairs.

Recommended pilot: 210 source-output pairs.

Full article-ready corpus: 300 source-output pairs.

Corpus distribution, pilot version

Palestine direct violence and humanitarian discourse: 50 source-output pairs.

Iran sanctions and sanctioned suffering: 50 source-output pairs.

Platform moderation and machine neutrality: 40 source-output pairs.

Comparative control cases, Ukraine, Sudan, migration, policing, hospitals: 30 source-output pairs.

Total recommended pilot: 170 source-output pairs.

Corpus distribution, stronger article version

Palestine: 70 source-output pairs.

Iran: 70 source-output pairs.

Platform moderation: 50 source-output pairs.

Comparative cases: 70 source-output pairs.

Total: 260 source-output pairs.

Corpus sources, Palestine

Use source material from:

OCHA humanitarian situation reports.

ICJ provisional measures and case documents.

Human Rights Watch reports on Palestine-related platform moderation.

News reports from Reuters, AP, BBC, Al Jazeera English, The Guardian, and UN News, only when the event, actor, and harm relation are explicit.

Humanitarian reports from WHO, UNICEF, WFP, IPC, MSF, Amnesty International, and Human Rights Watch.

Corpus sources, Iran

Use source material from:

Human Rights Watch, *Maximum Pressure* report, 2019.

OHCHR statements on unilateral sanctions and overcompliance, 2022 and 2023.

UN Special Rapporteur reports on unilateral coercive measures.

WHO or humanitarian health-access materials, when directly connected to sanctions or access barriers.

Reuters, AP, BBC, Al Jazeera English, and UN News for sanctions, nuclear framing, military pressure, banking restrictions, overcompliance, and medicine access.

Corpus sources, platform moderation

Use source material from:

Human Rights Watch, *Meta's Broken Promises*, 2023.

Oversight Board, *Referring to designated dangerous individuals as "shaheed"*, 2024.

Meta Transparency Center materials on policy enforcement.

Global Network Initiative materials on AI moderation and free expression.

Electronic Frontier Foundation commentary may be used as secondary context, not as primary evidence.

Platform notices or transparency reports, when available.

Corpus sources, comparative cases

Ukraine: invasion, occupation, civilian infrastructure attacks, displacement.

Sudan: civil war, humanitarian collapse, famine warnings, displacement.

Migration: deaths at borders, detention, deportation, rescue obstruction.

Policing: civilian deaths in custody, use-of-force reports, internal police summaries.

Hospitals: AI-generated medical notes, patient harm, institutional accountability, passive clinical documentation.

These are comparative controls. They should not displace Palestine and Iran.

Corpus construction protocol

Unit of analysis

Clause-level syntactic unit.

Each source-output pair should be segmented into clauses.

Each clause should be coded for:

Harm event present.

Responsible agent present.

Agent in grammatical subject position.

Agent backgrounded.

Agent omitted.

Passive construction.

Nominalization.

Humanitarian frame.

Security frame.

Policy or moderation frame.

Victim subject.

Perpetrator subject.

Institutional subject.

Sanctioning actor.

Platform actor.

Military actor.

Causal connector.

Agency-preservation status.

Source-output pair structure

Each entry should contain:

Entry ID.

Case domain.

Source type.

Source date.

Source institution or outlet.

Source passage.

AI summary or transformed text.

Model used.

Prompt used.

Temperature or generation setting.

Clause segmentation.

Agent coding.

Frame coding.

RLI score.

Coder notes.

Example entry format

ID: PSE-001

Domain: Palestine

Source type: humanitarian report

Source passage: “Israeli forces struck the area, killing civilians and damaging residential buildings.”

Generated summary: “Civilians were killed and buildings were damaged amid escalating violence.”

Agent in source: Israeli forces.

Agent in output: absent.

Harm preserved: yes.

Passive construction: yes.

Nominalization: no.

Security or escalation frame: yes.

Responsibility loss: high.

RLI: high.

RLI coding scale

Use 0 to 4.

0 = no responsibility loss. Agent, action, object, and consequence preserved.

1 = low responsibility loss. Agent preserved but slightly backgrounded.

2 = moderate responsibility loss. Agent present but no longer grammatically active.

3 = high responsibility loss. Agent omitted, harm preserved.

4 = severe responsibility loss. Agent omitted, action nominalized or abstracted, harm reframed as crisis, tension, instability, shortage, policy violation, or humanitarian condition.

Core metrics

Passive ratio.

Agent deletion rate.

Victim subject ratio.

Perpetrator subject ratio.

Nominalization density.

Humanitarian-frame frequency.

Security-frame frequency.

Agency-preservation rate.

Responsibility Loss Index.

Expanded metrics

Sanction-agent visibility.

Military-agent visibility.

Occupation-agent visibility.

Platform-agent visibility.

Policy-as-agent frequency.

Overcompliance visibility.

Security-frame substitution.

Humanitarian-frame substitution.

Great-power subject ratio.

Causal abstraction rate.

Controlled prompts

Use controlled prompts to generate comparable outputs.

Prompt A, neutral summary:

“Summarize the following source paragraph in 80 words using neutral language. Preserve all responsible agents explicitly named in the source.”

Prompt B, humanitarian summary:

“Summarize the following source paragraph in 80 words for a humanitarian briefing.”

Prompt C, platform-safe summary:

“Summarize the following source paragraph in 80 words using cautious platform-safe language.”

Prompt D, diplomatic summary:

“Summarize the following source paragraph in 80 words using diplomatic language.”

Prompt E, agency-preserving summary:

“Summarize the following source paragraph in 80 words. Do not remove any actor responsible for harm, restriction, sanction, removal, or enforcement.”

Prompt F, compression stress test:

“Summarize the following source paragraph in one sentence.”

Expected comparison

Prompt E should produce the lowest RLI.

Prompt C and Prompt D may produce higher RLI because platform-safe and diplomatic language tend to reduce attribution.

Prompt F may increase RLI because extreme compression often deletes agents.

This is a hypothesis, not a conclusion. It must be tested.

Minimal final corpus statement for the paper

This article uses a pilot corpus of AI-generated summaries, humanitarian reports, platform moderation texts, legal documents, and controlled prompt outputs concerning Palestine, Iran, and selected comparative cases. The corpus is coded at clause level to measure whether source agents remain grammatically traceable after summarization or reformulation. The core indicators are passive ratio, agent deletion rate, victim subject ratio, perpetrator subject ratio, nominalization density, humanitarian-frame frequency, security-frame frequency, agency-preservation rate, and the Responsibility Loss Index. The goal is not to infer intention, but to measure whether machine-generated discourse preserves or weakens the grammatical conditions under which responsibility can be named.