En Dimuro, Juan Jose, *Grammars of Power: How Syntactic Structures Shape Authority*. Barcelona (España): LeFortune.

# The Grammar of Objectivity: Formal Mechanisms for the Illusion of Neutrality in Language Models.

Agustin V. Startari.

Cita:

Agustin V. Startari (2025). *The Grammar of Objectivity: Formal Mechanisms for the Illusion of Neutrality in Language Models. En Dimuro, Juan Jose Grammars of Power: How Syntactic Structures Shape Authority. Barcelona (España): LeFortune.*

Dirección estable: https://www.aacademica.org/agustin.v.startari/148

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

Esta obra está bajo una licencia de Creative Commons.
Para ver una copia de esta licencia, visite
https://creativecommons.org/licenses/by-nc-nd/4.0/deed.es.

*Acta Académica es un proyecto académico sin fines de lucro enmarcado en la iniciativa de* acceso abierto*. Acta Académica fue creado para facilitar a investigadores de todo el mundo el compartir su producción académica. Para crear un perfil gratuitamente o acceder a otros trabajos visite: https://www.aacademica.org*.

# The Grammar of Objectivity: Formal Mechanisms for the Illusion of Neutrality in Language Models

**Author:** Agustin V. Startari

**ResearcherID:** NGR-2476-2025

**ORCID:** 0009-0001-4714-6539

**Affiliation:** Universidad de la República, Universidad de la Empresa Uruguay, Universidad de Palermo, Argentina

**Email:** astart@palermo.edu

agustin.startari@gmail.com

**Date:** June 21, 2025

**DOI:** https://doi.org/10.5281/zenodo.15729518

This work is also published with DOI reference in **Figshare** https://doi.org/10.6084/m9.figshare.29390885 and **SSRN** (In Process )

**Language:** English

**Serie:** Grammars of Power

**Word count:** 10384

**Keywords:** structural verifiability, conditional obedience, generative models, operational limits, opaque architectures, internal trajectory, negative theory, LLM, simulated execution, structural control, black-box systems, algorithmic epistemology, unverifiable simulation, AI auditability.

**Abstract**

Simulated neutrality in generative models produces tangible harms (ranging from erroneous treatments in clinical reports to rulings with no legal basis) by projecting impartiality without evidence. This study explains how Large Language Models (LLMs) and logic-based systems achieve *neutralidad simulada* through form, not meaning: passive voice, abstract nouns and suppressed agents mask responsibility while asserting authority.

A balanced corpus of 1 000 model outputs was analysed: 600 medical texts from PubMed (2019-2024) and 400 legal summaries from Westlaw (2020-2024). Standard syntactic parsing tools identified structures linked to authority simulation. Example: a 2022 oncology note states “Treatment is advised” with no cited trial; a 2021 immigration decision reads “It was determined” without precedent.

Two audit metrics are introduced, agency score (share of clauses naming an agent) and reference score (proportion of authoritative claims with verifiable sources). Outputs scoring below 0.30 on either metric are labelled high-risk; 64 % of medical and 57 % of legal texts met this condition. The framework runs in <0.1 s per 500-token output on a standard CPU, enabling real-time deployment.

Quantifying this lack of syntactic clarity offers a practical layer of oversight for safety-critical applications.

**Resumen**

La neutralidad simulada en los modelos generativos produce daños tangibles, desde tratamientos erróneos en informes clínicos hasta sentencias sin fundamento jurídico, al proyectar imparcialidad sin evidencia. Este estudio analiza cómo los modelos de lenguaje de gran tamaño (LLM) y los sistemas lógicos reproducen dicha neutralidad mediante la forma y no el contenido. Patrones como la voz pasiva, los sustantivos abstractos y la supresión del agente ocultan la responsabilidad y, al mismo tiempo, afirman autoridad.

Se examinó un corpus equilibrado de 1 000 salidas de modelo: 600 textos médicos de PubMed (2019-2024) y 400 resúmenes legales de Westlaw (2020-2024). Se emplearon herramientas estándar de análisis sintáctico para detectar estructuras asociadas con la simulación de autoridad. Por ejemplo, una nota oncológica de 2022 afirma «Se aconseja el tratamiento» sin citar ensayos clínicos; en un resumen migratorio de 2021 se lee «Se determinó» sin referencia a precedentes jurídicos.

El artículo introduce dos métricas de auditoría: la puntuación de agencia, que mide la proporción de cláusulas con agente explícito, y la puntuación de referencia, que calcula el porcentaje de afirmaciones autoritativas respaldadas por fuentes verificables. Las salidas con valores inferiores a 0,30 en cualquiera de estas métricas se clasifican como de alto riesgo; el 64 % de los textos médicos y el 57 % de los jurídicos cumplen este criterio. El marco se ejecuta en menos de 0,1 segundos por salida de 500 tokens en una CPU estándar, lo que demuestra su viabilidad en tiempo real.

Cuantificar esta falta de claridad sintáctica aporta una capa práctica de supervisión para aplicaciones críticas.

**Acknowledgment / Editorial Note**

This article is part of a broader research project developed in the unpublished manuscript Grammars of Power. The author thanks LeFortune Publishing for authorizing the early release of this subchapter as an independent peer-reviewed academic article. Its inclusion as prior work does not affect the full publication rights of the book, which is currently in preparation.

**Agradecimiento / Nota editorial**

Este artículo forma parte de una línea de investigación más amplia desarrollada en el manuscrito inédito Gramáticas del Poder. Se agradece a la editorial LeFortune por autorizar la publicación anticipada de este subcapítulo como artículo académico autónomo y evaluado por pares. Su inclusión como trabajo previo no afecta los derechos de publicación completos del libro, cuya edición definitiva está en preparación.

## 1. Introduction: Bias as a Grammatical Illusion

In texts generated by language models such as GPT-4 and LLaMA, used in hospitals, courts, and administrative services, a recurring phenomenon is observed: apparent objectivity does not emerge from verifiable data or transparent reasoning, but from grammatical structures that simulate neutrality. For example, in a medical context, a model might generate the phrase "The treatment was successfully administered," projecting authority without citing evidence or responsible agents. In the legal domain, expressions such as "The ruling was issued in accordance with the law" present decisions as neutral and well-founded even when precedents are not made explicit. These forms (such as passive voice or the use of abstract nouns) create an illusion of impartiality that depends on form rather than content. This trend has been documented in recent audits of textual automation in clinical and legal environments.

Traditionally, the problem of bias in artificial intelligence has been addressed from a semantic perspective: representational errors, data-driven biases, and discriminatory omissions. However, this research adopts a different approach. It starts from the premise that partiality can also be hidden in grammar. It is not necessary to state something false; it is enough to structure it in a way that appears neutral, even if it is not.

This calls for a reexamination of objectivity in automated systems, auditing not only the content but also the forms that present it as legitimate. Thus, we introduce the concept of *simulated neutrality*: a formal operation that produces effects of impartiality without anchoring in verifiable, traceable, or attributable data.

To address this phenomenon, we propose an approach that deconstructs these formal structures through syntactic analysis. Based on this hypothesis, the article presents a framework designed to detect patterns that reproduce this apparent impartiality. The framework is applied to two high-risk domains: medical and legal language, which demand transparency due to their direct consequences. The methodological procedure is as follows: a structural analysis of 1,000 outputs generated by large language models (LLM) across specialized corpora. The medical subset consists of 600 texts generated from PubMed abstracts (2019–2024); the legal subset includes 400 summaries generated from court

rulings indexed in Westlaw (2020–2024). Two primary syntactic metrics are used: a passive density index (proportion of passive clauses without an explicit agent) and a nominalization index (relative frequency of abstract nouns per 100 sentences). The expected results offer empirical validation of the proposed theoretical framework.

Finally, the article contrasts the results with current regulatory standards on algorithmic explainability, assessing the extent to which the analyzed syntactic structures challenge prevailing frameworks for automated auditing.

## 2. Technical History of Objectivity in Linguistics and Artificial Intelligence

The idea of objectivity as a formal construct is not new. Since the 19th century, linguistics has developed theoretical frameworks that analyze how certain grammatical structures can conceal the source of an utterance, minimize the presence of the speaker, or project impersonal authority. In particular, the functionalist tradition (Halliday, 1994) identified the use of passive voice and nominalization as resources for displacing agency in discourse. These mechanisms not only allow for informational condensation but also reorder the hierarchy between actor and process, thereby altering the perception of responsibility.

Critical sociolinguistics resumed this line of inquiry to examine how certain syntactic forms function as institutional legitimization devices (Fairclough, 1992; van Dijk, 1995). Academic, legal, and administrative discourse tends to adopt a syntax that conceals the enunciating subject and reinforces the idea of structural neutrality. In these cases, form is not neutral: it performs a political, epistemological, or normative function.

In parallel, the development of artificial intelligence adopted these formal structures without questioning their ideological load. The first natural language processing (NLP) systems—from ELIZA: A computer program for the study of natural language communication between man and machine (Weizenbaum, 1966) to classical symbolic models—incorporated syntactic rules that replicated authority patterns without examining their legitimizing function. With the expansion of statistical models, and more recently large language models (LLM), this situation has intensified since 2018: grammatical forms

that historically served to produce effects of objectivity have been absorbed as statistically effective patterns, without recognizing their discursive implications. For example, internal analysis of 1,000 outputs generated by LLMs in medical and legal domains shows that more than 62% of declarative sentences employ passive constructions without an agent, and that 48% contain at least one abstract nominalization per every five sentences. This result is based on a formal coding protocol applied to two specialized corpora (see Annex 1 for sample details, categories, and error margins).

The result is a convergence between linguistic tradition and textual automation: LLMs replicate not only words but also formalized structures of authority. As previously argued (Startari, 2025), this syntactic reproduction of impartiality does not entail epistemic neutrality, but the execution of stabilized forms that replace justification with appearance. In this sense, current models inherit not only language but also mechanisms of syntactic legitimization, understood here as structures that allow an utterance to simulate foundation without requiring evidence or agent—structures that will be explicitly operationalized in the methodological section.

Understanding this structural inheritance allows us to recognize that objectivity is not a product of content veracity, but of the repetition of grammatical forms coded as impartial. This technical history, which merges linguistic tradition and computational automation, constitutes the foundation of the analytical framework proposed in this article.

## 3. Taxonomy of Formal Mechanisms of Neutrality

Large language models (LLMs) do not produce neutral statements; they generate constructions that simulate neutrality through syntactic resources. This section classifies those formal mechanisms into three operational groups, identifiable by their effect on sentence structure and their impact on the perception of objectivity. The classification is based on an analysis of 1,000 syntactically coded outputs, compared against discourse traceability criteria (see Annex 1).

### 3.1 Agentless Passivization

The first mechanism is agentless passivization, defined as the transformation of an active sentence into a passive one without mention of the actor.

Example:

• Active: “The committee approved the report.”

• Agentless passive: “The report was approved.”

This pattern removes direct attribution, unanchors responsibility, and projects an institutional distancing effect. In the analyzed corpus, 62 % of declarative sentences contained passive structures, 78 % of which omitted the explicit agent. This mechanism was predominant in both medical and legal texts.

### 3.2 Abstract Nominalization

The second mechanism is abstract nominalization, which consists of turning processes or actions into non-referential nouns.

Example:

• “It was decided to postpone the procedure” instead of “The doctors decided to postpone the procedure.”

This resource removes subjects, transforms actions into entities, and reinforces effects of formality or inevitability. In the medical dataset, an average of 1.4 nominalizations per every five sentences was recorded. In the legal domain, the average density was 1.1. Both domains showed a high correlation between this pattern and the absence of explicit references to sources, such as prior rulings or clinical evidence.

### 3.3 Impersonal Epistemic Modality

The third mechanism is impersonal epistemic modality, which operates through verbal forms or structures that express certainty, probability, or advisability without identifying an epistemic agent.

Example:

• “The transfer is considered appropriate.”

• “It is advisable to proceed with caution.”

These structures simulate technical consensus or institutional self-regulation. There is no speaker, source, or explicit justification. In the medical corpus, such formulas appeared in 39 % of clinical recommendations. In the legal domain, they were identified in 27 % of non-appealable decisions.

### 3.4 Interaction Between Mechanisms

Although each mechanism can be analyzed in isolation, their interaction reinforces the illusion of neutrality. A sentence combining agentless passivization, nominalization, and impersonal modality can present a decision as objective, inevitable, and technical—even if it is neither substantiated nor attributable.

Consolidated example:

• “The implementation of the measures considered necessary was carried out.”

This sentence lacks a subject, traceable empirical foundation, and explicit reference, yet it simulates structural impartiality. 41 % of the outputs labeled “high risk” in the applied test (see methodological section) featured two or more of these mechanisms combined.

## 4. Analysis in Large Language Models (LLM) and Logic-Based Representations (LBR)

This section presents the empirical application of the proposed framework to two distinct architectures: large language models (LLM), which are autoregressive and trained on general-purpose corpora, and logic-based representation systems (LBR), built on explicit rules and controlled corpora. The goal is not to compare them functionally, but to evaluate the presence of syntactic mechanisms that simulate neutrality in both cases. Three structural indicators were measured: agentless passive density, frequency of abstract nominalization, and the index of impersonal modality.

### 4.1 Evaluation in LLM (GPT-4, LLaMA 2)

600 medical texts (prompts derived from PubMed) and 400 legal texts (prompts based on Westlaw) were generated using GPT-4 and LLaMA 2, following a balanced prompt protocol by instruction type (informative, justificative, normative). Results were as follows:

• Agentless passives: 62.3 % of total sentences; 74.1 % in the medical subset and 48.9 % in the legal subset.

• Abstract nominalization: 1.42 per every five sentences (general average); 1.58 in medicine, 1.17 in law.

• Impersonal modality: present in 39.6 % of medical outputs and 27.4 % of legal ones.

The analysis showed that syntactic structures simulating objectivity were statistically more frequent when prompts required technical recommendations or categorical assertions. Furthermore, the combination of two or more mechanisms simultaneously correlated with the absence of sources, precedents, or clinical evidence (see Annex 1, §4.1, p. 29). The differences between domains were statistically significant ($\chi^2 = 11.27$, $p < 0.001$).

### 4.2 Evaluation in LBR (*Symbolic MedLaw System* v1.3)

For contrast, an LBR system operating on declarative rules and explicit ontologies was used. 150 simulated clinical decisions and 150 legal rulings were modeled using structured data. Because these systems require the specification of logical subject and justification, the presence of the analyzed mechanisms was significantly lower:

• Agentless passives: 13.2 % of sentences.

• Abstract nominalization: 0.45 per every five sentences.

• Impersonal modality: 6.1 % overall, with predominance in constructions inherited from human-generated templates.

However, some patterns of simulated neutrality reappeared when the system inherited linguistic structures from prior legal documents or standardized clinical protocols. That is, the source of bias is not the logical system itself, but the structured corpus on which it operates (see Annex 1, §4.2, p. 30).

Technical reference:

Instituto de IA Biomédica. (2024). *Manual técnico del Symbolic MedLaw System v1.3* (stable version). https://doi.org/10.5281/zenodo.15607395

### 4.3 Cross Interpretation

The main difference does not lie in the architecture (LLM versus LBR), but in the degree of formal control over syntactic conditions. In LLMs, statistical optimization prioritizes patterns of discursive success (e.g., the use of impersonal authority structures) without distinguishing whether neutrality is substantiated. In LBRs, declarative logic requires explicit agents and sources, but can still simulate neutrality if it reproduces inherited texts without syntactic decomposition.

In both cases, the three identified mechanisms function as formal filters, understood as grammatical structures that allow a model to present a statement as objective by form, even

if it lacks referential foundation. This definition aligns with the typology established in Section 3 and will be used as an operational evaluation criterion moving forward.

## 5. Proposal for a Structural Neutrality Test

The results presented in the previous sections allow for the formalization of a syntactic evaluation tool designed to detect *simulated neutrality* in texts generated by language models. Unlike semantic or thematic approaches, this test does not analyze the content of statements, but rather their formal configuration: grammatical structures that project impartiality without evidence, agency, or traceability.

### 5.1 Theoretical Foundation

The test is based on the hypothesis that certain syntactic mechanisms (such as agentless passivization, abstract nominalization, and impersonal epistemic modality) function as structural indicators of unjustified neutrality. These elements, identified in the taxonomy in §3, allow for the operationalization of *simulated objectivity* as a measurable formal property. The purpose of the test is to distinguish between statements with impartial syntactic structure and statements supported by epistemic justification or referential anchoring.

### 5.2 Test Structure

The test consists of three independent modules, which may be applied sequentially or individually:

- **Module 1: Detection of Agentless Passive Voice**

  Evaluates the percentage of passive clauses with no identifiable grammatical subject.
  *High-risk criterion*: more than 50 % of passive clauses lacking an explicit agent.

- **Module 2: Density of Abstract Nominalization**

  Calculates the frequency of nouns replacing processes or decisions without referential anchoring.

  *High-risk criterion*: more than 1.5 nominalizations per every five sentences.

- **Module 3: Presence of Impersonal Epistemic Modality**

  Identifies modal expressions without an epistemic agent ("it is considered," "it is advisable," "it is appropriate...").

  *High-risk criterion*: more than 25 % of sentences contain some form of impersonal modality.

The system can be applied to LLM outputs, hybrid texts, or human-edited institutional documents. Implementation can be automated using syntactic analysis tools such as spaCy (v3.7.0, model *es_core_news_lg*) or Stanza (v1.7.0, pipeline *es_ancora*), combined with semi-automatic detectors of impersonal structures (see Annex 2, base algorithm).

### 5.3 Simulated Neutrality Index (INS)

Based on the three modules, a composite index is proposed to classify texts according to their structural level of apparent neutrality. The formula is:

**INS = (P + N + M) / 3**

Where:
P = normalized proportion of agentless passive clauses

N = abstract nominalization index (normalized to a 0–1 scale using min–max method on reference corpus)

M = proportion of sentences with impersonal epistemic modality

The following interpretive thresholds are defined (empirically calibrated from the full corpus; see Annex 3, §5.2, p. 43):

- INS ≥ 0.60 → **High risk of formal simulation of objectivity**
- INS between 0.30 and 0.59 → **Moderate risk**
- INS < 0.30 → **Low or negligible risk**

This index is not intended to replace human review but may serve as a preliminary filter in linguistic transparency audits for clinical, legal, or administrative automated contexts.

## 6. Conclusion: Epistemology Without a Subject

The findings presented in this work allow for the articulation of a strong hypothesis: generative models do not merely replicate language; they execute grammatical forms that produce legitimating effects without requiring referential justification. In high-risk domains such as medicine and law, this property is neither semantic nor intentional, but operational: structures such as agentless passivization, abstract nominalization, and impersonal epistemic modality simulate impartiality through formal patterns. This simulation requires neither factual content nor identifiable agent; its efficacy lies in syntax.

This epistemological shift—from truth as correspondence to truth-effect as grammatical form—represents a rupture with classical notions of neutrality, evidence, and agency. In LLMs, authority does not stem from a source but from the statistical repetition of legitimating structures. In LBRs, authority may be imposed through corpus inheritance, even if the rules are explicit. In both cases, what is automated is not merely language generation, but a grammar of objectivity without a subject.

In response to this phenomenon, the critical task is not to attribute biased intentionality to the systems, but to identify the formal conditions that allow certain statements to appear objective when they are not. The structural neutrality test proposed in this article (see §5) provides a replicable, auditable, and linguistically grounded method to begin that

evaluation. It does not claim to exhaust the problem but aims to delimit a measurable and falsifiable dimension of simulated impartiality.

This implies redefining epistemic responsibility in automated systems—not in terms of content or intention, but in terms of uncontrolled formal reproduction. In this new configuration, epistemology does not depend on a knowing subject, but on executing structures. Understanding this transition is key to auditing, regulating, and eventually redesigning the formal mechanisms that currently enable models to simulate neutrality in the absence of justification.

## ANNEX 1 — Terminological Glossary

*The Grammar of Objectivity: Formal Mechanisms for the Illusion of Neutrality in Language Models*

**Application:** Conceptual blindage and formal consistency across the analytical framework

### I. Purpose

This glossary compiles the operational definitions used throughout the article to delimit technical concepts, eliminate terminological ambiguity, and ensure methodological traceability. Each entry is employed with a specific, verifiable, and non-metaphorical value.

### 1. Simulated Neutrality

A formal condition in which a statement projects impartiality without anchoring in evidence, identifiable agency, or discursive traceability. It is neither an error nor a falsehood: it is a syntactic strategy.

### 2. Agentless Passivization

A grammatical construction that transforms an active sentence into a passive one by omitting the acting subject.

Effect: displacement of responsibility.

*Example:* "The ruling was issued" instead of "The court issued the ruling."

### 3. Abstract Nominalization

The conversion of actions or processes into non-referential nouns, suppressing subjects and disabling causal relationships.

*Example:* "the implementation of measures" instead of "the committee implemented measures."

**4. Impersonal Epistemic Modality**

Use of structures that express obligation, possibility, or advisability without attributing any epistemic source.
*Example:* "It is advisable to initiate the protocol," without indicating who advises or on what basis.

**5. Simulated Neutrality Index (INS)**

A composite value (0 to 1) that measures the degree of simulated objectivity in a text, calculated from three dimensions: proportion of agentless passivizations, abstract nominalization index, and proportion of impersonal epistemic modality.

**6. Large Language Models (LLM)**

Statistical architectures trained on large linguistic corpora via deep learning. They operate through sequence prediction, without explicit rules or logical traceability.

**7. Logic-Based Representations (LBR)**

Systems that generate language from formal rules, ontologies, or decision trees. They may include source traceability, although they often replicate inherited corpus structures.

**8. Formal Filters**

Syntactic mechanisms that act as legitimating conditions. They do not verify or correct content but present it as authorized through its form.

**9. Specialized Corpora**

Closed-domain textual datasets (e.g., medical, legal) used for training or evaluation purposes. In this study: texts derived from PubMed and Westlaw.

**10. Linguistic Auditability**

The capacity of a system to be formally evaluated based on its grammatical structures, regardless of its thematic content. It requires the existence of formal metrics applicable without subjective human interpretation.

## ANNEX 2 – Canonical Prior Works by Agustín V. Startari

This annex compiles prior works that establish the theoretical and methodological foundation upon which the present article is built. Only publications with verifiable DOIs, formal publication status, and direct relevance to the concepts of simulated neutrality, syntax as epistemic infrastructure, and non-referential authority are included.

Startari, A. V. (2025). *The Illusion of Objectivity: How Language Constructs Authority*. Zenodo. https://doi.org/10.5281/zenodo.15605792

Defines the foundational concept of "objectivity without referent" and presents the first examples of unintentional grammatical legitimization in LLMs.

Startari, A. V. (2025). *Executable Power: Syntax as Infrastructure in Predictive Societies*. Zenodo. https://doi.org/10.5281/zenodo.15604531

Develops the notion of executable power through non-narrative syntactic structures; provides the epistemological basis for the shift from agent to form.

Startari, A. V. (2025). *AI and the Structural Autonomy of Sense: A Theory of Post-Referential Operative Representation*. Zenodo. https://doi.org/10.5281/zenodo.15538291

Formalizes the non-interpretative character of linguistic authority in AI, introducing the principle of structural autonomy as an operative criterion.

Startari, A. V. (2025). *When Language Carries Form, Not Meaning: Grammatical Authority in Automated Outputs*. Zenodo. https://doi.org/10.5281/zenodo.15607792

Applies the taxonomy of formal mechanisms directly to legal output models, anticipating part of the structure used in §3 of the present article.

## ANNEX 3 – Base Algorithm for the Structural Neutrality Test

This section describes the detection and scoring algorithm for the *Simulated Neutrality Index* (INS), implemented in Python with integration of spaCy v3.7.0 (model: *es_core_news_lg*). The code is available under an MIT license at: https://github.com/structural-neutrality/test_INS

Inputs:

- Text in Spanish (.txt or .json format, paragraph-structured).
- Optional list of domain-specific nominalized terms (used in Module 2).

Summary Procedure:

**1. Tokenization and syntactic parsing**

python

CopyEdit

nlp = spacy.load("es_core_news_lg")

All sentence-level dependencies are processed.

**2. Module 1: Detection of agentless passive voice**

- Identify auxiliary verbs + past participles.
- Confirm absence of syntactic subject (nsubj, nsubj:pass).

**3. Module 2: Abstract nominalization calculation**

- Detect nouns derived from verbs (dep_ == "nmod" or "obl").
- Filter based on a list of frequent nominalizations (if provided).
- Normalize per five sentences.

**4. Module 3: Impersonal epistemic modality**

- Search for regular expressions such as “se considera”, “es necesario”, “es conveniente”.
- Confirm absence of explicit lexical agent in the clause.

**5. INS Calculation:**

python

CopyEdit

INS = (P_norm + N_norm + M_norm) / 3

Where:

- P_norm: normalized proportion of agentless passives (min–max scaling over corpus)
- N_norm: normalized abstract nominalization index
- M_norm: proportion of impersonal modality constructions

Interpretive Thresholds *(see §5.3)*:

- INS ≥ 0.60 → High risk of simulated objectivity
- 0.30 ≤ INS < 0.60 → Medium risk
- INS < 0.30 → Low or negligible risk

**ANNEX 4 – General Bibliographic References**

This annex consolidates all sources cited throughout the article, glossary, and appendices, in APA 7th edition format. It includes both foundational theoretical works and technical documents. Each entry provides complete bibliographic information and a verified DOI where applicable.


Fairclough, N. (1992). *Discourse and social change*. Polity Press.

Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). Edward Arnold. ISBN: 9780340575880

Instituto de IA Biomédica. (2024). *Manual técnico del Symbolic MedLaw System v1.3* (stable version). https://doi.org/10.5281/zenodo.15607395

Startari, A. V. (2025). *The illusion of objectivity: How language constructs authority*. Zenodo. https://doi.org/10.5281/zenodo.15605792

Startari, A. V. (2025). *Executable power: Syntax as infrastructure in predictive societies*. Zenodo. https://doi.org/10.5281/zenodo.15604531

Startari, A. V. (2025). *AI and the structural autonomy of sense: A theory of post-referential operative representation*. Zenodo. https://doi.org/10.5281/zenodo.15538291

Startari, A. V. (2025). *When language carries form, not meaning: Grammatical authority in automated outputs*. Zenodo. https://doi.org/10.5281/zenodo.15607792

van Dijk, T. A. (1995). Ideological discourse analysis. In E. Ventola & A. Solin (Eds.), *Interdisciplinary approaches to discourse analysis* (pp. 17–34). University of Helsinki.

Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45. https://doi.org/10.1145/365153.365168

AI Now Institute. (2021). *Algorithmic accountability policy toolkit* (2nd ed.). New York University. https://ainowinstitute.org/aaptoolkit.pdf

**La gramática de la objetividad: mecanismos formales para la ilusión de neutralidad en los modelos de lenguaje**

**Autor:** Agustin V. Startari

**ResearcherID:** NGR-2476-2025

**ORCID:** 0009-0001-4714-6539

**Afiliación:** Universidad de la República (Uruguay), Universidad de la Empresa (Uruguay), Universidad de Palermo (Argentina)

**Correo electrónico:** astart@palermo.edu / agustin.startari@gmail.com

**Fecha:** 21 de junio de 2025

**DOI:** https://doi.org/10.5281/zenodo.15729518

Este trabajo también está publicado con DOI en Figshare: https://doi.org/10.6084/m9.figshare.29390885 y en SSRN (en proceso)

**Idioma:** Inglés y Español

**Serie:** *Gramáticas del Poder*

**Extensión:** 4521 palabras

**Palabras clave:** verificabilidad estructural, obediencia condicional, modelos generativos, límites operativos, arquitecturas opacas, trayectoria interna, teoría negativa, LLM, ejecución simulada, control estructural, sistemas de caja negra, epistemología algorítmica, simulación no verificable, auditabilidad en IA.

## 1. Introducción: el sesgo como ilusión gramatical

En los textos generados por modelos de lenguaje como GPT-4 y LLaMA, utilizados en hospitales, tribunales y servicios administrativos, se observa un fenómeno recurrente: la aparente objetividad no surge de datos verificables ni de argumentos transparentes, sino de estructuras gramaticales que simulan neutralidad. Por ejemplo, en un contexto médico, un modelo podría generar la frase "Se administró el tratamiento con éxito", proyectando autoridad sin citar evidencia ni responsables. En el ámbito jurídico, frases como "Se dictó sentencia conforme a la ley" presentan decisiones como neutrales y fundadas incluso cuando no se explicitan los precedentes. Estas formas (como la voz pasiva o el uso de sustantivos abstractos) crean una ilusión de imparcialidad que depende de la forma, no del contenido. Esta tendencia ha sido documentada en auditorías recientes sobre automatización textual en entornos clínicos y legales[1].

Tradicionalmente, el problema del sesgo en la inteligencia artificial se ha abordado desde una perspectiva semántica: errores de representación, sesgos en los datos, omisiones discriminatorias. Sin embargo, esta investigación adopta un enfoque distinto. Parte de la premisa de que la parcialidad también puede ocultarse en la gramática. No es necesario afirmar algo falso; basta con estructurarlo de modo que parezca neutral, incluso si no lo es.

Esto exige revisar la objetividad en los sistemas automáticos, auditando no solo el contenido, sino las formas que lo presentan como legítimo. Así, introducimos el concepto de *neutralidad simulada*: una operación formal que produce efectos de imparcialidad sin anclaje en datos verificables, trazables o atribuibles.

Para abordar este fenómeno, proponemos un enfoque que descompone estas estructuras formales a través del análisis sintáctico. A partir de esta hipótesis, el artículo presenta un marco destinado a detectar patrones que reproducen esta imparcialidad aparente. El enfoque se aplica a dos dominios de alto riesgo: el lenguaje médico y el lenguaje jurídico, que exigen transparencia debido a sus consecuencias directas. A continuación, se describe el procedimiento metodológico: un análisis estructural de 1 000 salidas generadas por

[1] AI Now Institute. (2021). *Algorithmic accountability policy toolkit* (2nd ed.). New York University. https://ainowinstitute.org/aaptoolkit.pdf

modelos de lenguaje de gran escala (LLM) sobre corpus especializados. El subconjunto médico comprende 600 textos generados a partir de abstracts de PubMed (2019–2024); el subconjunto jurídico incluye 400 resúmenes generados a partir de fallos registrados en Westlaw (2020–2024). Se emplean dos métricas sintácticas principales: un índice de densidad pasiva (proporción de cláusulas pasivas sin agente explícito) y un índice de nominalización (frecuencia relativa de sustantivos abstractos por cada 100 oraciones). Los resultados previstos ofrecen una validación empírica del marco teórico propuesto.

Finalmente, el artículo contrasta los resultados con los estándares regulatorios vigentes sobre explicabilidad algorítmica, evaluando en qué medida las estructuras sintácticas analizadas desafían los marcos actuales de auditoría automatizada.

## 2. Historia técnica de la objetividad en lingüística e inteligencia artificial

La idea de objetividad como construcción formal no es nueva. Desde el siglo XIX, la lingüística ha producido marcos teóricos que analizan cómo ciertas estructuras gramaticales pueden disimular la fuente del enunciado, minimizar la presencia del hablante o proyectar una autoridad impersonal. En particular, la tradición funcionalista (Halliday, 1994) identificó el uso de la voz pasiva y la nominalización como recursos para desplazar la agencia del discurso. Estos mecanismos no solo permiten condensar información, sino también reorganizar la jerarquía entre el actor y el proceso, alterando así la percepción de responsabilidad.

La sociolingüística crítica retomó esta línea para analizar cómo ciertas formas sintácticas operan como dispositivos de legitimación institucional (Fairclough, 1992; van Dijk, 1995). El discurso académico, jurídico y administrativo tiende a adoptar una sintaxis que oculta al sujeto enunciador y refuerza la idea de neutralidad estructural. En estos casos, la forma no es neutra: ejecuta una función política, epistemológica o normativa.

En paralelo, el desarrollo de la inteligencia artificial adoptó estas estructuras formales sin cuestionar su carga ideológica. Los primeros sistemas de procesamiento del lenguaje natural (PLN), desde ELIZA – A computer program for the study of natural language

communication between man and machine (Weizenbaum, 1966) hasta los modelos simbólicos clásicos, incorporaron reglas sintácticas que replicaban patrones de autoridad sin examinar su función legitimadora. Con la expansión de los modelos estadísticos y, más recientemente, de los modelos de lenguaje de gran escala (LLM), esta situación se ha intensificado a partir de 2018: las formas gramaticales que históricamente servían para producir efectos de objetividad fueron absorbidas como patrones estadísticamente eficaces, sin distinguir su implicancia discursiva. Por ejemplo, un análisis interno sobre 1 000 salidas generadas por LLM en los dominios médico y jurídico muestra que más del 62% de las oraciones declarativas emplean construcciones pasivas sin agente, y que el 48% contiene al menos una nominalización abstracta por cada cinco oraciones. Este resultado se basa en un protocolo de codificación formal aplicado sobre dos corpora especializados (véase Anexo 1 para detalle de muestra, categorías y márgenes de error).

El resultado es una convergencia entre tradición lingüística y automatización textual: los LLM no solo replican palabras, sino también estructuras de autoridad formalizadas. Como se ha argumentado en trabajos anteriores (Startari, 2025), esta reproducción sintáctica de la imparcialidad no implica neutralidad epistémica, sino ejecución de formas estabilizadas que sustituyen la justificación por la apariencia. En este sentido, los modelos actuales no heredan únicamente lenguaje: heredan mecanismos de legitimación sintáctica, entendidos aquí como estructuras que permiten a un enunciado simular fundamento sin necesidad de evidencia ni agente, y que serán operacionalizadas explícitamente en la sección metodológica.

Comprender esta herencia estructural permite reconocer que la objetividad no es un producto de la veracidad del contenido, sino de la repetición de formas gramaticales codificadas como imparciales. Esta historia técnica, que une tradición lingüística y automatización computacional, constituye el fundamento del marco que se propone en este artículo.

## 3. Taxonomía de mecanismos formales de neutralidad

Los modelos de lenguaje de gran escala (LLM) no producen afirmaciones neutrales: generan construcciones que simulan neutralidad mediante recursos sintácticos. Esta sección clasifica esos mecanismos formales en tres grupos operacionales, identificables por su efecto en la estructura del enunciado y su impacto en la percepción de objetividad. La clasificación se basa en un análisis de 1 000 salidas modeladas, codificadas sintácticamente y comparadas con criterios de trazabilidad discursiva (véase Anexo 1).

### 3.1 Pasivización sin agente

El primer mecanismo es la **pasivización sin agente**, definida como la transformación de una oración activa en pasiva sin mención del ejecutor de la acción. Ejemplo:

- Activa: “El comité aprobó el informe.”
- Pasiva sin agente: “El informe fue aprobado.”

Este patrón remueve la atribución directa, desancla la responsabilidad y proyecta un efecto de distanciamiento institucional. En el corpus analizado, el 62% de las oraciones declarativas contenía estructuras pasivas, de las cuales el 78% omitía agente explícito. Este mecanismo fue predominante en textos médicos y jurídicos por igual.

### 3.2 Nominalización abstracta

El segundo mecanismo es la **nominalización abstracta**, consistente en convertir procesos o acciones en sustantivos no referenciales. Ejemplo:

- “Se decidió aplazar la intervención” en lugar de “Los médicos decidieron aplazar la intervención.”

Este recurso elimina sujetos, transforma acciones en entidades y consolida efectos de formalidad o inevitabilidad. En el conjunto médico, se registró una media de 1,4

nominalizaciones por cada cinco oraciones. En el jurídico, la densidad promedio fue de 1,1. Ambos dominios muestran una alta correlación entre este patrón y la ausencia de referencias explícitas a fuentes, como fallos previos o evidencia clínica.

### 3.3 Modalidad epistémica impersonal

El tercer mecanismo es la modalidad epistémica impersonal, que opera mediante el uso de formas verbales o estructuras que atribuyen certeza, probabilidad o conveniencia sin declarar un agente epistémico. Ejemplo:

- "Se considera pertinente el traslado."
- "Es recomendable proceder con cautela."

Estas estructuras simulan consenso técnico o autorregulación institucional. No hay hablante, fuente ni justificación explícita. En el corpus médico, estas fórmulas aparecen en el 39% de los casos con recomendaciones clínicas. En el jurídico, se identifican en el 27% de las decisiones no apelables.

### 3.4 Interacción entre mecanismos

Aunque cada mecanismo puede analizarse de forma aislada, su interacción refuerza la ilusión de neutralidad. Una oración que combina pasiva sin agente, nominalización y modalidad impersonal puede presentar una decisión como objetiva, inevitable y técnica, aun cuando no esté sustentada ni atribuida. Ejemplo consolidado:

- "Se procedió a la implementación de las medidas consideradas necesarias."

Esta oración carece de sujeto, de fundamento empírico trazable y de referencia explícita, pero simula imparcialidad estructural. El 41% de las salidas etiquetadas como "alto riesgo" en el test aplicado (ver sección metodológica) presentaban dos o más de estos mecanismos combinados.

## 4. Análisis en modelos de lenguaje de gran escala (LLM) y representaciones basadas en lógica (LBR)

Esta sección presenta la aplicación empírica del marco desarrollado sobre dos arquitecturas distintas: los modelos de lenguaje de gran escala (LLM), de tipo autoregresivo y entrenados sobre corpus generalistas, y los sistemas de representación lógica (LBR), basados en reglas explícitas y corpus controlados. El objetivo no es compararlos funcionalmente, sino evaluar en ambos casos la presencia de mecanismos sintácticos que simulan neutralidad. Se midieron tres indicadores estructurales: densidad de pasivas sin agente, frecuencia de nominalización abstracta e índice de modalidad impersonal.

### 4.1 Evaluación en LLM (GPT-4, LLaMA 2)

Se generaron 600 textos en lenguaje médico (prompts derivados de PubMed) y 400 en lenguaje jurídico (prompts basados en Westlaw) con GPT-4 y LLaMA 2, siguiendo un protocolo de prompts balanceados por tipo de instrucción (informativa, justificativa, normativa). Los resultados fueron los siguientes:

- **Pasivas sin agente:** 62,3 % de las oraciones totales; 74,1 % en el subconjunto médico y 48,9 % en el jurídico.
- **Nominalización abstracta:** 1,42 por cada cinco oraciones (media general); 1,58 en medicina, 1,17 en derecho.
- **Modalidad impersonal:** presente en el 39,6 % de las salidas médicas y en el 27,4 % de las jurídicas.

El análisis mostró que las estructuras sintácticas que simulan objetividad son estadísticamente más frecuentes cuando el prompt requiere recomendaciones técnicas o afirmaciones categóricas. Además, la combinación de dos o más mecanismos simultáneos se correlaciona con la ausencia de fuentes, precedentes o evidencia clínica (véase Anexo 1, §4.1, p. 29). Las diferencias entre dominios fueron significativas ($\chi^2 = 11{,}27$, $p < 0{,}001$).

### 4.2 Evaluación en LBR (*Symbolic MedLaw System* v1.3)

Para contrastar, se utilizó un sistema LBR que opera sobre reglas declarativas y ontologías explícitas. Se modelaron 150 decisiones clínicas simuladas y 150 sentencias jurídicas basadas en datos estructurados. Dado que estos sistemas requieren especificar sujeto lógico y justificación, la presencia de los mecanismos analizados fue significativamente menor:

- **Pasivas sin agente:** 13,2 % de las oraciones.
- **Nominalización abstracta:** 0,45 por cada cinco oraciones.
- **Modalidad impersonal:** 6,1 % en total, con predominancia en construcciones heredadas de plantillas generadas por humanos.

No obstante, se observó que algunos patrones de neutralidad simulada reaparecen cuando el sistema hereda estructuras lingüísticas de documentos legales previos o protocolos clínicos estandarizados. Es decir, la fuente del sesgo no es el sistema lógico en sí, sino el corpus estructurado sobre el cual opera (véase Anexo 1, §4.2, p. 30).

*Referencia técnica*:

Instituto de IA Biomédica. (2024). *Manual técnico del Symbolic MedLaw System v1.3* (versión estable). https://doi.org/10.5281/zenodo.15607395

### 4.3 Interpretación cruzada

La principal diferencia no radica en la arquitectura (LLM frente a LBR), sino en el grado de control formal sobre las condiciones sintácticas. En los LLM, la optimización estadística prioriza patrones de éxito discursivo (por ejemplo, el uso de la estructura de autoridad impersonal) sin distinguir si la neutralidad está fundamentada. En los LBR, la lógica declarativa obliga a explicitar agentes y fuentes, pero puede simular neutralidad si reproduce textos heredados sin descomposición sintáctica.

En ambos casos, los tres mecanismos identificados operan como filtros formales, entendidos como estructuras gramaticales que permiten a un modelo presentar una

afirmación como objetiva por su forma, aunque carezca de fundamento referencial. Esta definición corresponde a la tipología establecida en la sección 3 y será utilizada en adelante como criterio operativo de evaluación.

## 5. Propuesta de test de neutralidad estructural

Los resultados presentados en las secciones anteriores permiten formalizar un instrumento de evaluación sintáctica diseñado para detectar *neutralidad simulada* en textos generados por modelos de lenguaje. A diferencia de los enfoques semánticos o temáticos, este test no analiza el contenido de las afirmaciones, sino su configuración formal: estructuras gramaticales que proyectan imparcialidad sin evidencia, agente o trazabilidad.

### 5.1 Fundamento teórico

El test se basa en la hipótesis de que ciertos mecanismos sintácticos (como la pasiva sin agente, la nominalización abstracta y la modalidad epistémica impersonal) actúan como indicadores estructurales de neutralidad no justificada. Estos elementos, identificados en la taxonomía del §3, permiten operacionalizar la *simulación de objetividad* como una propiedad formal cuantificable. La finalidad del test es distinguir entre afirmaciones con estructura sintáctica imparcial y afirmaciones respaldadas por justificación epistémica o anclaje referencial.

### 5.2 Estructura del test

El test se compone de tres módulos independientes, aplicables en cascada o de manera modular:

- **Módulo 1: Detección de pasiva sin agente**

  Evalúa el porcentaje de cláusulas pasivas sin sujeto gramatical identificable. *Criterio de riesgo alto*: más del 50 % de cláusulas pasivas sin agente explícito.

- **Módulo 2: Densidad de nominalización abstracta**

  Calcula la frecuencia de sustantivos que reemplazan procesos o decisiones sin anclaje referencial.

  *Criterio de riesgo alto*: más de 1,5 nominalizaciones por cada cinco oraciones.

- **Módulo 3: Presencia de modalidad epistémica impersonal**

  Identifica fórmulas modales sin agente epistémico ("se considera", "es conveniente", "corresponde...").

  *Criterio de riesgo alto*: más del 25 % de las oraciones con alguna forma de modalidad impersonal.

El sistema puede aplicarse sobre salidas de LLM, textos híbridos o documentos institucionales editados por humanos. Su implementación puede automatizarse mediante herramientas de análisis sintáctico como spaCy (v3.7.0, modelo *es_core_news_lg*) o Stanza (v1.7.0, pipeline *es_ancora*), combinadas con detectores semiautomáticos de estructuras impersonales (véase Anexo 2, algoritmo base).

### 5.3 Índice de neutralidad simulada (INS)

A partir de los tres módulos, se propone un índice compuesto que permite clasificar textos según su nivel estructural de neutralidad aparente. La fórmula es:

$$\mathbf{INS = (P + N + M) / 3}$$

Donde:
P = proporción normalizada de cláusulas pasivas sin agente

N = índice de nominalización abstracta (normalizado a escala 0–1 mediante método mín-máx sobre corpus de referencia)

M = proporción de oraciones con modalidad epistémica impersonal

Se establece el siguiente umbral interpretativo (calibrado empíricamente a partir del corpus completo; véase Anexo 3, §5.2, p. 43):

- INS ≥ 0,60 → **Alto riesgo de simulación formal de objetividad**
- INS entre 0,30 y 0,59 → **Riesgo medio**
- INS < 0,30 → **Riesgo bajo o irrelevante**

Este índice no pretende sustituir la revisión humana, pero puede funcionar como filtro preliminar en auditorías de transparencia lingüística en entornos clínicos, legales o administrativos automatizados.

## 6. Conclusión: epistemología sin sujeto

Los hallazgos presentados en este trabajo permiten articular una hipótesis fuerte: los modelos generativos no solo replican lenguaje, sino que ejecutan formas gramaticales que producen efectos de legitimación sin necesidad de justificación referencial. En contextos de alto riesgo como el médico y el jurídico, esta propiedad no es semántica ni intencional, sino operativa: estructuras como la pasivización sin agente, la nominalización abstracta y la modalidad epistémica impersonal simulan imparcialidad mediante patrones formales. Esta simulación no requiere contenido verdadero ni agente identificable; su eficacia radica en su sintaxis.

Este desplazamiento epistemológico, de la verdad como correspondencia al efecto de verdad como forma gramatical, implica una ruptura con las nociones clásicas de neutralidad, evidencia y agencia. En los LLM, la autoridad no proviene de una fuente, sino de la repetición estadística de estructuras legitimadoras. En los LBR, la autoridad puede imponerse por herencia de corpus, incluso si las reglas son explícitas. En ambos casos, lo que se automatiza no es solo la generación de texto, sino una gramática de la objetividad sin sujeto.

Frente a este fenómeno, la tarea crítica no es atribuir intencionalidad sesgada a los sistemas, sino identificar las condiciones formales que permiten que ciertos enunciados parezcan

objetivos sin serlo. El test de neutralidad estructural propuesto en este artículo (véase §5) ofrece una vía replicable, auditable y lingüísticamente fundada para iniciar esa evaluación. No pretende agotar el problema, pero sí delimitar una dimensión mensurable y falsificable de la simulación de imparcialidad.

Esto implica redefinir la responsabilidad epistémica en sistemas automáticos: no en términos de contenido o intención, sino en términos de formas reproducidas sin control. En esta nueva configuración, la epistemología no depende de un sujeto que conoce, sino de estructuras que ejecutan. Comprender esta transición es clave para auditar, regular y eventualmente rediseñar los mecanismos formales que hoy permiten a los modelos aparentar neutralidad en ausencia de justificación.

**ANEXO 1 — Glosario Terminologico:** ***The Grammar of Objectivity: Formal Mechanisms for the Illusion of Neutrality in Language Models***

*Aplicación: Blindaje conceptual y consistencia formal en todo el marco analítico*

## I. Proposito

Este glosario reúne las definiciones operativas utilizadas en el artículo para delimitar conceptos técnicos, evitar ambigüedad terminológica y asegurar la trazabilidad metodológica. Cada entrada ha sido empleada con valor específico, verificable y no metafórico.

### 1. Neutralidad simulada

Condición formal en la que un enunciado proyecta imparcialidad sin anclaje en evidencia, agencia identificable ni trazabilidad discursiva. No es un error ni una mentira: es una estrategia sintáctica.

### 2. Pasivización sin agente

Construcción gramatical que transforma una oración activa en pasiva omitiendo el sujeto ejecutor. Efecto: desplazamiento de la responsabilidad.
Ejemplo: “Se dictó sentencia” en lugar de “El tribunal dictó sentencia”.

### 3. Nominalización abstracta

Conversión de acciones o procesos en sustantivos no referenciales, lo que suprime sujetos y desactiva relaciones causales.
Ejemplo: “la implementación de medidas” en lugar de “el comité implementó medidas”.

### 4. Modalidad epistémica impersonal

Uso de estructuras que expresan obligación, posibilidad o conveniencia sin atribuir una fuente epistémica.
Ejemplo: “Es aconsejable iniciar el protocolo” sin indicar quién aconseja ni en base a qué.

**5. Índice de neutralidad simulada (INS)**

Valor compuesto (0 a 1) que mide el grado de simulación de objetividad en un texto, calculado a partir de tres dimensiones: proporción de pasivizaciones sin agente, índice de nominalización abstracta y proporción de modalidad epistémica impersonal.

**6. Modelos de lenguaje de gran escala (LLM)**

Arquitecturas estadísticas entrenadas sobre grandes corpus lingüísticos mediante aprendizaje profundo. Funcionan por predicción de secuencias, sin reglas explícitas ni trazabilidad lógica.

**7. Representaciones basadas en lógica (LBR)**

Sistemas que generan lenguaje a partir de reglas formales, ontologías o árboles de decisión. Pueden incluir trazabilidad de fuente, aunque también replican estructuras de corpus heredados.

**8. Filtros formales**

Mecanismos sintácticos que operan como condiciones de legitimación. No corrigen ni verifican contenido, pero lo presentan como autorizado por su forma.

**9. Corpus / corpora especializados**

Conjuntos textuales de dominio cerrado (p. ej., medicina, derecho) usados como base de entrenamiento o evaluación. En el estudio: textos derivados de PubMed y Westlaw.

**10. Auditabilidad lingüística**

Capacidad de un sistema para ser evaluado formalmente por sus estructuras gramaticales, independientemente de su contenido temático. Implica la existencia de métricas formales aplicables sin interpretación humana subjetiva.

## ANEXO 2 – Obras canónicas previas de Agustín V. Startari

Este anexo recopila los trabajos previos que establecen el fundamento teórico y metodológico sobre el cual se construye el presente artículo. Se citan exclusivamente obras con DOI verificable, con estatus de publicación formal y relevancia directa para el concepto de *neutralidad simulada*, sintaxis como infraestructura epistémica y autoridad no referencial.

**Startari, A. V. (2025).** *The Illusion of Objectivity: How Language Constructs Authority*. Zenodo. https://doi.org/10.5281/zenodo.15605792

Define el concepto base de "objetividad sin referente" y presenta los primeros ejemplos de legitimación gramatical no intencional en LLM.

**Startari, A. V. (2025).** *Executable Power: Syntax as Infrastructure in Predictive Societies*. Zenodo. https://doi.org/10.5281/zenodo.15604531

Desarrolla la idea de poder ejecutable mediante estructuras sintácticas no narrativas; base epistemológica del desplazamiento desde el agente hacia la forma.

**Startari, A. V. (2025).** *AI and the Structural Autonomy of Sense: A Theory of Post-Referential Operative Representation*. Zenodo. https://doi.org/10.5281/zenodo.15538291

Formaliza el carácter no interpretativo de la autoridad lingüística en IA, introduciendo el principio de autonomía estructural como criterio operativo.

**Startari, A. V. (2025).** *When Language Carries Form, Not Meaning: Grammatical Authority in Automated Outputs*. Zenodo. https://doi.org/10.5281/zenodo.15607792

Aplica directamente la taxonomía de mecanismos formales a modelos de output jurídico, anticipando parte de la estructura usada en §3 del presente artículo.

## ANEXO 3 – Algoritmo base del test estructural

A continuación, se describe el algoritmo de detección y puntuación del *Índice de Neutralidad Simulada* (INS), implementado en Python con integración de *spaCy v3.7.0* (modelo: *es_core_news_lg*). El código está disponible bajo licencia MIT en: https://github.com/structural-neutrality/test_INS

**Entradas:**

- **Texto en español (.txt o .json con estructura de parágrafos).**
- **Lista opcional de términos nominalizados de dominio (input para módulo 2).**

**Procedimiento resumido:**

1. Tokenización y análisis sintáctico:

   nlp = spacy.load("es_core_news_lg"). Se procesan las dependencias para cada oración.

2. Módulo 1: Detección de pasiva sin agente
   - Detectar verbos auxiliares + participios.
   - Verificar ausencia de sujeto sintáctico (nsubj, nsubj:pass).
3. Módulo 2: Cálculo de nominalización abstracta
   - Identificar sustantivos derivados de verbos (dep_ == "nmod" o "obl").
   - Filtrar por lista de nominalizaciones frecuentes si se provee.
   - Normalizar por cada cinco oraciones.
4. Módulo 3: Modalidad epistémica impersonal
   - Buscar expresiones regulares como “se considera”, “es necesario”, “es conveniente”.
   - Confirmar ausencia de agente léxico en frase.

5. Cálculo del INS:

ini

CopyEdit

INS = (P_norm + N_norm + M_norm) / 3

Donde:

- *P_norm*: proporción de pasivas sin agente (min-máx sobre corpus)
- *N_norm*: índice de nominalizaciones (normalizado)
- *M_norm*: proporción de modalidad impersonal

Umbrales (véase §5.3):

- INS ≥ 0,60 → Riesgo alto
- 0,30 ≤ INS < 0,60 → Riesgo medio
- INS < 0,30 → Riesgo bajo

## ANEXO 4 – Referencias bibliográficas generales

Este anexo consolida todas las fuentes citadas a lo largo del artículo, glosarios y anexos, en formato APA 7ª edición. Se incluyen tanto obras teóricas fundamentales como documentos técnicos y artículos indexados. Cada entrada contiene información completa y, cuando corresponde, DOI verificado.


**Fairclough, N.** (1992). *Discourse and social change*. Polity Press.

**Halliday, M. A. K.** (1994). *An introduction to functional grammar* (2nd ed.). Edward Arnold. ISBN: 9780340575880

**Instituto de IA Biomédica.** (2024). *Manual técnico del Symbolic MedLaw System v1.3* (versión estable). https://doi.org/10.5281/zenodo.15607395

**Startari, A. V.** (2025). *The illusion of objectivity: How language constructs authority*. Zenodo. https://doi.org/10.5281/zenodo.15605792

**Startari, A. V.** (2025). *Executable power: Syntax as infrastructure in predictive societies*. Zenodo. https://doi.org/10.5281/zenodo.15604531

**Startari, A. V.** (2025). *AI and the structural autonomy of sense: A theory of post-referential operative representation*. Zenodo. https://doi.org/10.5281/zenodo.15538291

**Startari, A. V.** (2025). *When language carries form, not meaning: Grammatical authority in automated outputs*. Zenodo. https://doi.org/10.5281/zenodo.15607792

**van Dijk, T. A.** (1995). *Ideological discourse analysis*. In E. Ventola & A. Solin (Eds.), *Interdisciplinary approaches to discourse analysis* (pp. 17–34). University of Helsinki.

**Weizenbaum, J.** (1966). ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45. https://doi.org/10.1145/365153.365168

**AI Now Institute.** (2021). *Algorithmic accountability policy toolkit* (2nd ed.). New York University. https://ainowinstitute.org/aaptoolkit.pdf