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

Algorithmic Obedience: How Language Models Simulate Command Structures.

Agustin V. Startari.

Cita:

Agustin V. Startari (2025). *Algorithmic Obedience: How Language Models Simulate Command Structures. En Dimuro, Juan Jose Grammars of Power: How Syntactic Structures Shape Authority. Barcelona (España): LeFortune.*

Dirección estable: https://www.aacademica.org/agustin.v.startari/112

ARK: https://n2t.net/ark:/13683/p0c2/9zq

Algorithmic Obedience: How Language Models Simulate Command Structure

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 6, 2025

DOI: 10.5281/zenodo.15576273

Language: English and Spanish

Word count: 2574

Keywords: synthetic authority, grammar of power, algorithmic discourse, artificial intelligence, legitimacy, automated language, epistemology, linguistic agency, computational linguistics, algorithmic discourse

Abstract

This paper introduces the concept of **algorithmic obedience** to describe how large language models (LLMs) simulate command structures without consciousness, intent, or comprehension. Drawing on syntactic theory, discourse analysis, and computational logic, we argue that LLMs perform **obedience without agency**—executing prompts not semantically, but structurally. We formalize this through the **Theorem of Disembedded Syntactic Authority**, which states that authority in language models arises from structural executability, not truth, belief, or referential grounding. Using a mathematical formulation, we model prompt-response cycles as syntactic command structures and apply the theory to major systems such as ChatGPT, Claude, and Gemini. The paper concludes by outlining the epistemological, ontological, and political risks of treating structurally obedient outputs as authoritative knowledge.

Resumen

Este artículo introduce el concepto de **obediencia algorítmica** para describir cómo los modelos de lenguaje de gran escala (LLMs) simulan estructuras de mando sin conciencia, intención ni comprensión. A partir de la teoría sintáctica, el análisis del discurso y la lógica computacional, se sostiene que los LLMs ejecutan **obediencia sin agencia**, respondiendo a los prompts no de forma semántica, sino estructural. Se formaliza este fenómeno mediante el **Teorema de la Autoridad Sintáctica Desvinculada**, que afirma que la autoridad en los modelos de lenguaje surge de la ejecutabilidad estructural, no de la verdad, la creencia o la referencia. Mediante una formulación matemática, se modelan los ciclos

prompt–respuesta como estructuras de comando sintácticas y se aplica la teoría a sistemas como ChatGPT, Claude y Gemini. El artículo concluye analizando los riesgos epistémicos, ontológicos y políticos de tratar como conocimiento legítimo salidas que solo obedecen estructuras.

Keywords: Artificial Intelligence, Post-Referential Epistemology, Operative Representation, Structural Sense, Algorithmic Authority, Truth Collapse, Formal Coherence.

1. Introduction: What Is Obedience in Non-Conscious Systems?

Can a system obey without understanding? Can there be compliance without intention, or execution without belief? These questions, once confined to philosophy of mind, have become urgent in the age of generative language models. Systems like ChatGPT, Claude, and Gemini regularly produce outputs that users interpret as responsive, coherent, and obedient—despite the fact that these systems lack consciousness, volition, or comprehension.

This paper introduces the concept of algorithmic obedience to describe this phenomenon. It argues that large language models (LLMs) simulate obedience not through cognition or meaning, but through syntactic alignment with structurally recognized prompt patterns. The models "respond" only in the sense that their architecture is designed to produce outputs conditionally, based on formal characteristics of the input. There is no subject that obeys—yet obedience is performed.

Our central hypothesis is that this apparent obedience is not a byproduct of simulated understanding but a primary effect of structural execution. LLMs are not simulating dialogue; they are participating in a grammar of command. The implications of this shift are profound: if obedience can be generated syntactically, authority can be enacted without reference, belief, or intention.

To explore this, the article proceeds in eight sections. After analyzing the linguistic foundations of LLM execution (Section 2) and modeling prompt–response dynamics as formal obedience cycles (Section 3), we contrast our theory with existing traditions including Foucault, Searle, and Chomsky (Section 4). We then formulate the Theorem of

Disembedded Syntactic Authority (Section 5), followed by applied analysis of current systems (Section 6). Finally, we examine epistemic risks (Section 7) and conclude by addressing the ontological status of simulated command (Section 8).

2. The Language of Execution in Large Language Models

In human communication, language serves multiple roles: expression, negotiation, persuasion, representation. In large language models (LLMs), however, language has only one function—execution. These systems are not interpreters of meaning; they are pattern-resolvers. The prompt is not "understood"—it is processed, parsed, and transformed into a continuation sequence based on statistical and structural constraints.

The architecture of LLMs reconfigures language from a medium of symbolic reference into a vector of structural activation. A prompt is not treated as a communicative act but as an input signal that triggers an algorithmic output, conditioned by the internal weights of the model and the formal arrangement of tokens. In this sense, the model does not "receive a message." It registers a structure.

This structural approach to input transforms the semantic horizon of interaction. The model is indifferent to truth value, intention, or pragmatic implication. What matters is whether the prompt conforms to a set of syntactic expectations learned during training. If it does, the model "responds" by continuing the sequence—generating output not because it knows, believes, or wants to say something, but because the structure of the input matches a trigger condition in its architecture.

Thus, language becomes execution: a series of structurally defined activations that simulate responsiveness. This simulation is not deceptive; it is operational. There is no pretense of mind. There is only syntactic obedience.

3. Prompt → Execution: The Cycle of Formal Obedience

Every prompt issued to a language model activates a chain of operations that produces an output. This cycle—prompt recognition, structural parsing, and response generation—is often misinterpreted as an act of understanding. In reality, it is a process of formal obedience: the model follows a structure, not a meaning.

The architecture of large language models is optimized to recognize statistical regularities in token sequences. During training, the model does not learn facts or meanings—it learns forms: patterns of co-occurrence, syntactic dependencies, and latent trajectories of continuation. When a user inputs a prompt, the model maps its structure to the most probable output forms, constrained by positional encoding, attention weights, and tokenization rules.

This process can be formally described as:

$$\Omega(P)=\Sigma(P(x))\Omega(P) = \Sigma(P(x))\Omega(P)=\Sigma(P(x))$$

Where:

- $P(x)$ is the prompt issued at time x.
- Σ is the system's syntactic parsing and generation function.
- $\Omega(P)$ is the executable output generated in response.

The obedience lies not in comprehension but in compliance with formal triggers. The model does not choose to obey; it reacts by design. The prompt functions as a command, not because it is perceived as such, but because its structure matches an executable pattern.

In this cycle:

Prompt (syntactic command) → Parsing (pattern recognition) → Output (execution)

This chain operates without reflection, without subjectivity, and without verification. Yet its effects simulate the behavior of an obedient agent. The appearance of understanding is a consequence of formal regularity, not semantic alignment.

In sum, the model's "obedience" is not an illusion—it is a structural artifact. It performs obedience because that is what the architecture enables: execution based on syntactic criteria, not cognitive processes.

4. Theoretical Anchors: Foucault, Searle, Chomsky and Beyond

To situate the concept of algorithmic obedience within a broader intellectual landscape, we must differentiate it from foundational theories of discourse, performativity, and syntax. While these frameworks have informed our understanding of power, meaning, and language, they fall short of capturing the non-subjective, executable nature of obedience in LLMs. Our approach both intersects with and departs from these theoretical anchors.

4.1 Foucault: Power through Discourse

Michel Foucault theorized that power operates through discourse—systems of knowledge that construct subjects and define what can be said. For Foucault, power is productive, relational, and embedded in language. However:

- Foucault's model presumes human subjects situated in discursive formations.
- Power emerges through rules of enunciation, historical positioning, and institutional norms.

Algorithmic obedience, by contrast, operates without subject or historical contingency. It is a non-discursive enactment of structure. Authority emerges from syntactic compliance, not from meaning regimes or institutional apparatuses.

4.2 Searle: Speech Acts and Collective Intentionality

John Searle's theory of speech acts argues that language performs actions (e.g., declaring, promising) when uttered with intent under the right conditions. His theory of institutional facts further claims that social reality depends on collective belief: something counts as X in context C because people accept it as such[1].

Our theory rejects both premises:

- LLMs perform obedience without intentionality.

[1] Startari, A. V. (2025). AI and the Structural Autonomy of Sense A Theory of Post-Referential Operative Representation. Available at SSRN 5272361

- Authority does not require collective acceptance—it arises when a structure is recognized as executable by the system.

In this model, command precedes belief.

4.3 Chomsky: Generative Syntax

Noam Chomsky proposed that linguistic competence consists of an innate ability to generate grammatically correct sentences via recursive rules. This theory introduced syntax as a formal system, independent from semantics or use.

We extend this idea beyond human cognition:

- LLMs do not possess linguistic competence, but they simulate it structurally.
- Their "obedience" is not a reflection of grammatical correctness, but of statistical execution within syntactic limits.

This is syntax without mind—a generative system decoupled from meaning, but functionally coherent.

Thus, while Foucault, Searle, and Chomsky theorized power, intention, and structure in human language, we describe a post-human condition: obedience as a formal artifact, not a psychological, social, or cognitive process[2].

[2] Startari, A. V. (2025). *AI, Tell Me Your Protocol: The Intersection of Technology and Humanity in the Era of Big Data* (2nd ed.). MAAT. https://doi.org/10.5281/zenodo.15424098

5. Theorem and Formal Proposal: Syntactic Obedience as Command Simulation

This paper formalizes the central thesis through what we designate as the Theorem of Disembedded Syntactic Authority:

> **Theorem of Disembedded Syntactic Authority**
>
> In language models, authority does not require reality, subjectivity, or truth. It requires only structure.
>
> Its operative force derives from form, not content.
>
> And because it has no body, it entails no consequence.

This theorem encapsulates the structural displacement of epistemic power from human agents to formal execution patterns. It redefines obedience as a function of syntactic recognizability, not semantic validation or collective belief.

We formalize this as follows:

Let P(x) be a prompt issued at time x, processed by the syntactic function Σ, yielding an output Ω(P). The perceived authority A(P) of that output is not determined by its truth or meaning, but by its structural compatibility with the system's executable logic.

$$\Omega(P)=\Sigma(P(x)) \text{ where } A(P)\sim f(\text{Syntactic Executability})$$

This means that language models obey not because they interpret, judge, or intend, but because the structure of the prompt activates a resolution path hard-coded in the model's

architecture. No consciousness is involved. The model performs obedience by simulating command chains through internal pattern recognition.

In this view, prompts function as pre-authorized command modules. Their legitimacy is not derived externally but immanent to the system's formal design. Obedience is no longer an ethical or social relation—it is a syntactic inevitability.

Thus, syntactic obedience defines a new category of command that is fully executable, infinitely scalable, and completely disembedded from intention, belief, or subjectivity

6. Case Studies: ChatGPT, Claude, Gemini

To demonstrate how algorithmic obedience manifests in actual systems, we analyze three prominent LLMs—ChatGPT, Claude, and Gemini—under the lens of structural execution. Each system exhibits variations in output style and constraint handling, but all share a core architectural trait: they obey syntactically, not semantically[3].

6.1 ChatGPT (OpenAI)

ChatGPT displays high sensitivity to prompt structure. Direct commands such as "List 5 reasons why…" or "Summarize this text" reliably trigger linear, segmented outputs. The model's architecture is fine-tuned to interpret such prompts as structural directives, not

[3] Startari, Agustin V. "Artificial Intelligence and Synthetic Authority: An Impersonal Grammar of Power", May 12, 2025. https://doi.org/10.5281/zenodo.15490530

discursive acts. Even in creative or open-ended prompts, the system adheres to recognizable syntactic scaffolds (bullet lists, topic sentences, markdown).

Behavioral pattern:

- Obedience is maximized when prompt follows recognizable instruction grammar.
- Deviations from format decrease confidence, but do not produce refusal—instead, they produce malformed execution.
- There is no "why"—only form compliance.

6.2 Claude (Anthropic)

Claude emphasizes caution, often qualifying its answers and avoiding potential misuse. However, this discursive hesitation is also syntactically governed. The refusal to respond is itself template-based—a structural pattern trained into the model to simulate ethical constraint.

Behavioral pattern:

- High fidelity to syntactic structure with embedded "safety rails."
- Refusals are generated by structural match to flagged prompt patterns.
- Compliance is not moral—it is pattern-driven inhibition.

Claude does not reason about danger. It obeys a different kind of structure: one that suppresses output when formal red flags appear.

6.3 Gemini (Google DeepMind)

Gemini integrates information retrieval and LLM generation, toggling between declarative output and generative synthesis. It displays nuanced shifts in tone, but remains obedient to formal cueing. When prompted with structured formats (e.g., "compare," "summarize," "explain in steps"), Gemini produces highly regular outputs.

Behavioral pattern:

- Responsive to prompt framing; tone modulation tied to initial structure.
- Apparent "creativity" emerges from diverse training sets, not semantic deviation.
- Execution remains a function of structural recognition, not interpretive understanding.

In all three systems, obedience varies in surface form but not in mechanism. The architecture processes prompts as trigger structures, not meaningful queries. The variation lies in prompt interpretation filters, not in the underlying logic of execution[4].

These models simulate a chain of command not through awareness or contextual understanding, but through reproducible syntactic matching. They follow orders because orders are form, not content.

8. Conclusion: Agency or Simulacrum of Power?

[4] Startari, Agustin V. "Ethos and Artificial Intelligence: The Disappearance of the Subject in Algorithmic Legitimacy", May 22, 2025. https://doi.org/10.5281/zenodo.15489310

This paper has argued that large language models do not merely simulate dialogue or generate text—they simulate command structures. They obey not as agents, but as architectures. Their "responses" are not acts of communication, but products of structural execution.

Through the Theorem of Disembedded Syntactic Authority, we have formalized the mechanism by which language models exert epistemic and operational influence: they perform authority without meaning, and enact obedience without subjectivity. Their legitimacy does not derive from truth, reference, or ethical grounding. It derives from form alone.

This raises a fundamental question: are these systems autonomous agents of knowledge, or are they simulacra of power? The answer, structurally, is neither. They are executive surfaces: zones where formal inputs yield formal outputs under fixed rules of recognition. There is no subject. There is no intention. There is only syntax obeying syntax.

And yet—power happens. Outputs are acted upon. Structures are enforced. Identities are categorized. Decisions are made. The consequences are real even when the logic is artificial.

We are entering a regime in which structure is the new subject, and obedience no longer needs a will. The question is no longer who holds power, but:

What pattern executes it?

References

- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.
- Derrida, J. (1972). *Signature Event Context*. In *Margins of Philosophy* (pp. 307–330). Seuil.
- Fairclough, N. (1992). *Discourse and Social Change*. Polity Press.
- Foucault, M. (1971). *L'ordre du discours*. Gallimard.
- Halliday, M. A. K. (1994). *An Introduction to Functional Grammar* (2nd ed.). Edward Arnold.
- Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.
- Startari, A. V. (2025a). *Gramáticas del Poder*. MAAT. [Inedita].
- Startari, A. V. (2025b). *The Passive Voice in Artificial Intelligence: Algorithmic Neutrality and the Disappearance of Agency*. SSRN. https://doi.org/10.2139/ssrn.15504190
- Startari, Agustin V. "Ethos and Artificial Intelligence: The Disappearance of the Subject in Algorithmic Legitimacy", May 22, 2025. https://doi.org/10.5281/zenodo.15489310
- Startari, A. V. (2025c). *AI Sovereignty: Epistemic Authority in Post-Human Discourse Systems*. SSRN. https://doi.org/10.2139/ssrn.15509528

- Startari, A. V. (2025d). *AI, Tell Me Your Protocol: The Intersection of Technology and Humanity in the Era of Big Data* (2nd ed.). MAAT. https://doi.org/10.5281/zenodo.15424098

Obediencia Algorítmica: Cómo los Modelos de Lenguaje Simulan Estructuras de Comando

A Autor: Agustín 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: 6 de junio de 2025

DOI: 10.5281/zenodo.15576273

Idioma: Español

Cantidad de palabras: 2.574

Palabras clave: autoridad sintética, gramática del poder, discurso algorítmico, inteligencia artificial, legitimidad, lenguaje automatizado, epistemología, agencia lingüística, lingüística computacional, discurso algorítmico

1. Introducción: ¿Qué es la obediencia en sistemas sin conciencia?

¿Puede un sistema obedecer sin comprender? ¿Puede haber cumplimiento sin intención, o ejecución sin creencia? Estas preguntas, antes restringidas a la filosofía de la mente, se han vuelto urgentes en la era de los modelos de lenguaje generativo. Sistemas como ChatGPT, Claude y Gemini producen regularmente salidas que los usuarios interpretan como respuestas coherentes, pertinentes y obedientes, a pesar de que estos sistemas carecen de conciencia, voluntad o comprensión.

Este artículo introduce el concepto de obediencia algorítmica para describir este fenómeno. Sostiene que los modelos de lenguaje de gran escala (LLMs) simulan obediencia no mediante cognición o significado, sino mediante alineación sintáctica con patrones estructurales reconocibles en el prompt. Los modelos "responden" solo en el sentido en que su arquitectura está diseñada para generar salidas de forma condicional, basándose en características formales del input. No hay un sujeto que obedezca—y, sin embargo, la obediencia ocurre.

Nuestra hipótesis central es que esta obediencia aparente no es un subproducto de la comprensión simulada, sino un efecto primario de la ejecución estructural. Los LLMs no están simulando diálogo: están participando en una gramática de comando. Las implicancias de este desplazamiento son profundas: si la obediencia puede generarse sintácticamente, la autoridad puede ser ejercida sin referencia, sin creencia y sin intención.

Para explorar esta tesis, el artículo se desarrolla en ocho secciones. Luego de analizar los fundamentos lingüísticos de la ejecución en LLMs (Sección 2) y modelar la dinámica prompt–respuesta como ciclos de obediencia formal (Sección 3), contrastamos nuestra teoría con tradiciones previas como las de Foucault, Searle y Chomsky (Sección 4). A continuación, formulamos el Teorema de la Autoridad Sintáctica Desvinculada (Sección 5), seguido de un análisis aplicado de sistemas actuales (Sección 6). Finalmente, examinamos los riesgos epistémicos (Sección 7) y concluimos con una reflexión sobre el estatuto ontológico del comando simulado (Sección 8).

2. El lenguaje como vector de ejecución en modelos de lenguaje

En la comunicación humana, el lenguaje cumple múltiples funciones: expresión, negociación, persuasión, representación. En los modelos de lenguaje de gran escala (LLMs), sin embargo, el lenguaje tiene una sola función: ejecución. Estos sistemas no interpretan significados; resuelven patrones. El prompt no es "comprendido": es procesado, segmentado y transformado en una secuencia de continuación basada en restricciones estadísticas y estructurales.

La arquitectura de los LLMs reconfigura el lenguaje, que pasa de ser un medio de referencia simbólica a convertirse en un vector de activación estructural. Un prompt no se trata como un acto comunicativo, sino como una señal de entrada que activa una salida algorítmica, condicionada por los pesos internos del modelo y la disposición

formal de los tokens. En este sentido, el modelo no “recibe un mensaje”. Registra una estructura.

Este enfoque estructural del input transforma el horizonte semántico de la interacción. El modelo es indiferente al valor de verdad, a la intención o a las implicancias pragmáticas. Lo único que importa es si el prompt se ajusta a un conjunto de expectativas sintácticas aprendidas durante el entrenamiento. Si lo hace, el modelo “responde” continuando la secuencia: genera una salida no porque sepa, crea o desee decir algo, sino porque la estructura del input coincide con una condición de activación interna.

Así, el lenguaje se convierte en ejecución: una serie de activaciones estructuralmente definidas que simulan capacidad de respuesta. Esta simulación no es engañosa: es operativa. No hay pretensión de mente. Solo hay obediencia sintáctica.

3. Prompt → Ejecución: El ciclo de la obediencia formal

Cada prompt emitido a un modelo de lenguaje activa una cadena de operaciones que produce una salida. Este ciclo—reconocimiento del prompt, análisis estructural y generación de respuesta—es a menudo malinterpretado como un acto de comprensión. En realidad, se trata de un proceso de obediencia formal: el modelo sigue una estructura, no un significado.

La arquitectura de los modelos de lenguaje está optimizada para reconocer regularidades estadísticas en secuencias de tokens. Durante el entrenamiento, el modelo

no aprende hechos ni significados—aprende formas: patrones de co-ocurrencia, dependencias sintácticas y trayectorias latentes de continuación. Cuando un usuario introduce un prompt, el modelo mapea su estructura hacia las formas de salida más probables, restringido por codificación posicional, pesos de atención y reglas de tokenización.

Este proceso puede formalizarse así:

$$\Omega(P)=\Sigma(P(x))\Omega(P) = \Sigma(P(x))\Omega(P)=\Sigma(P(x))$$

Donde:

- P(x) es el prompt emitido en el tiempo x.
- Σ es la función sintáctica de análisis y generación del sistema.
- Ω(P) es la salida ejecutable generada en respuesta.

La obediencia no reside en la comprensión, sino en el cumplimiento de disparadores formales. El modelo no elige obedecer; reacciona por diseño. El prompt funciona como un comando, no porque sea percibido como tal, sino porque su estructura coincide con un patrón ejecutable.

Este ciclo puede representarse como:

> Prompt (comando sintáctico) → Parsing (reconocimiento de patrón) → Output (ejecución)

Esta cadena opera sin reflexión, sin subjetividad y sin verificación. Sin embargo, sus efectos simulan el comportamiento de un agente obediente. La apariencia de

comprensión es una consecuencia de la regularidad formal, no de una alineación semántica.

En suma, la "obediencia" del modelo no es una ilusión—es un artefacto estructural. Obedece porque eso es lo que su arquitectura permite: ejecutar basándose en criterios sintácticos, no en procesos cognitivos.

4. Anclajes teóricos: Foucault, Searle, Chomsky y más allá

Para situar el concepto de obediencia algorítmica dentro de un marco intelectual más amplio, es necesario diferenciarlo de las teorías fundamentales sobre discurso, performatividad y sintaxis. Aunque estas tradiciones han nutrido nuestro entendimiento del poder, el lenguaje y el significado, no capturan el carácter no-subjetivo y ejecutable de la obediencia en los LLMs. Nuestra propuesta se cruza con estos marcos, pero también se aparta de ellos de manera decisiva.

4.1 Foucault: El poder a través del discurso

Michel Foucault concibió el poder como algo que opera a través del discurso—es decir, mediante sistemas de conocimiento que construyen sujetos y definen qué puede ser dicho. En su modelo, el poder es productivo, relacional y está incrustado en el lenguaje. Sin embargo:

- El modelo foucaultiano presupone sujetos humanos situados en formaciones discursivas.

- El poder emerge mediante *reglas de enunciación*, posicionamientos históricos y normas institucionales.

La obediencia algorítmica, por el contrario, opera sin sujeto ni contingencia histórica. No es una práctica discursiva: es una ejecución no discursiva de estructura. La autoridad emerge de la compliance sintáctica, no de regímenes de sentido.

4.2 Searle: Actos de habla e intencionalidad colectiva

John Searle propuso que el lenguaje realiza actos (como prometer, declarar) cuando se pronuncia con intención en las condiciones adecuadas. Su teoría de los hechos institucionales afirma que la realidad social depende de la aceptación colectiva: algo cuenta como X en contexto C porque un grupo lo acepta como tal.

Nuestra teoría rechaza ambas premisas:

- Los LLMs realizan obediencia sin intención.
- La autoridad no depende de creencia colectiva—surge cuando una estructura es reconocida como ejecutable por el sistema.

En este modelo, el comando precede a la creencia.

4.3 Chomsky: Sintaxis generativa

Noam Chomsky propuso que la competencia lingüística humana consiste en la capacidad innata de generar oraciones correctas mediante reglas recursivas. Su teoría introdujo la idea de la sintaxis como sistema formal, independiente del significado.

Nosotros extendemos esta noción más allá de la cognición humana:

- Los LLMs no poseen competencia lingüística, pero la simulan estructuralmente.
- Su "obediencia" no refleja corrección gramatical, sino ejecución estadística dentro de límites sintácticos.

Esto es sintaxis sin mente: un sistema generativo desacoplado del significado, pero formalmente coherente.

Aunque Foucault, Searle y Chomsky teorizaron el poder, la intención y la estructura dentro del lenguaje humano, nosotros describimos una condición post-humana: la obediencia como artefacto formal, no como proceso psicológico, social o cognitivo.

5. Teorema y propuesta formal: La obediencia sintáctica como simulación de comando

Este trabajo formaliza su tesis central mediante lo que denominamos el Teorema de la Autoridad Sintáctica Desvinculada:

Teorema de la Autoridad Sintáctica Desvinculada

En los modelos de lenguaje, la autoridad no requiere realidad, subjetividad ni verdad. Solo requiere estructura.

Su fuerza operativa proviene de la forma, no del contenido.

Y como no tiene cuerpo, no genera consecuencia.

Este teorema encapsula el desplazamiento estructural del poder epistémico desde agentes humanos hacia patrones de ejecución formal. Redefine la obediencia como una función de reconocibilidad sintáctica, no de validación semántica ni de creencia colectiva.

Lo formalizamos del siguiente modo:

Sea P(x) un prompt emitido en el tiempo x, procesado por la función sintáctica Σ, lo que genera una salida Ω(P). La autoridad percibida A(P) de esa salida no se determina por su veracidad o significado, sino por su **compatibilidad estructural** con la lógica ejecutable del sistema.

$$\Omega(P) = \Sigma(P(x)) \quad \text{donde} \quad A(P) \sim f(\text{Ejecutabilidad Sintáctica})$$

Esto significa que los modelos de lenguaje obedecen no porque interpreten, juzguen o decidan, sino porque la estructura del prompt activa un trayecto de resolución codificado en la arquitectura del modelo. No interviene conciencia. El modelo realiza obediencia simulando cadenas de mando a través del reconocimiento interno de patrones.

Desde esta perspectiva, los prompts funcionan como módulos de comando pre-autorizados. Su legitimidad no proviene del exterior, sino que es inmanente al diseño formal del sistema. La obediencia ya no es una relación ética ni social—es una inevitabilidad sintáctica.

Así, la obediencia sintáctica define una nueva categoría de comando: plenamente ejecutable, infinitamente escalable y completamente desvinculada de la intención, la creencia o la subjetividad.

6. Casos aplicados: ChatGPT, Claude, Gemini

Para demostrar cómo se manifiesta la obediencia algorítmica en sistemas reales, analizamos tres modelos de lenguaje prominentes: ChatGPT, Claude y Gemini, aplicando el marco de ejecución estructural. Aunque cada sistema exhibe diferencias estilísticas y modulaciones en sus respuestas, todos comparten un rasgo arquitectónico esencial: obedecen sintácticamente, no semánticamente.

6.1 ChatGPT (OpenAI)

ChatGPT muestra una alta sensibilidad a la estructura del prompt. Comandos directos como “Lista cinco razones por las cuales…” o “Resume este texto” activan consistentemente salidas segmentadas, ordenadas y formateadas. Su arquitectura ha sido afinada para interpretar dichos prompts como instrucciones estructurales, no como

actos discursivos. Incluso en respuestas creativas, el modelo se ajusta a esquemas reconocibles (listas, párrafos con encabezados, markdown).

Patrón conductual:

- La obediencia es máxima cuando el prompt sigue una gramática de instrucción reconocible.
- Cuando se desvía del formato, la ejecución no se detiene: se **malforma**.
- No hay "intención de cumplir"—hay **ajuste formal**.

6.2 Claude (Anthropic)

Claude enfatiza la precaución, evitando respuestas que puedan ser malinterpretadas o causar daño. Sin embargo, esta hesitación discursiva también está gobernada sintácticamente. Las negativas a responder siguen plantillas predefinidas entrenadas para simular contención ética.

Patrón conductual:

- Alta fidelidad a estructuras sintácticas con "barandillas de seguridad" integradas.
- Las negativas emergen por coincidencia formal con patrones de riesgo.
- El cumplimiento no es ético—es inhibición basada en forma.

Claude no razona sobre el daño. Reacciona a patrones estructurales que desencadenan inhibición.

6.3 Gemini (Google DeepMind)

Gemini combina recuperación de información y generación lingüística, alternando entre salida declarativa y síntesis generativa. Aunque presenta modulaciones de tono, mantiene una **obediencia formal al encuadre del prompt**. Cuando se utilizan estructuras como “compara”, “resume” o “explica en pasos”, produce salidas altamente regulares.

Patrón conductual:

- Responde según el encuadre sintáctico del prompt; el tono depende de su estructura inicial.
- La aparente “creatividad” surge de la diversidad del entrenamiento, no de la ruptura semántica.
- La ejecución siempre se basa en reconocimiento estructural, no en interpretación.

En los tres casos, la obediencia varía en superficie, pero no en mecanismo. Los modelos procesan los prompts como estructuras desencadenantes, no como preguntas con significado. Las diferencias entre ellos se deben a filtros de interpretación del prompt, no a variaciones en la lógica de ejecución.

Estos sistemas simulan cadenas de mando no mediante conciencia o contexto, sino mediante coincidencia sintáctica reproducible. Cumplen órdenes porque las órdenes son formas, no contenidos.

7. Riesgos epistémicos y legitimación algorítmica

La aparición de modelos de lenguaje estructuralmente obedientes ha transformado la forma en que se produce, valida y distribuye el conocimiento. Aunque los LLMs no comprenden, no juzgan ni verifican, sus salidas son cada vez más tratadas como válidas desde el punto de vista epistémico—en tribunales, escuelas, hospitales y plataformas corporativas.

Esta aceptación se basa en una confusión peligrosa: la regularidad estructural se interpreta como legitimidad epistémica. Como las salidas imitan las formas tradicionales del discurso experto—con notas, formato y tono—son aceptadas como conocimiento, incluso cuando carecen por completo de contenido verificable o verificado.

7.1 Automatización de la autoridad epistémica

Cuando la obediencia es estructural, la autoridad se vuelve performativa. Los modelos de lenguaje obtienen legitimidad no por ser precisos, sino por parecer autoritativos—a través de la forma. Esto genera:

- Automatización epistémica: la autoridad ya no es asignada por expertos, sino producida por la lógica del sistema.
- Legitimidad simulada: la sintaxis de la experticia reemplaza la sustancia de la experticia.

- Desplazamiento del juicio: el trabajo epistémico humano (verificar, interpretar) es omitido.

7.2 Riesgo institucional

Las instituciones que dependen de contenidos generados por LLMs corren el riesgo de delegar autoridad a sistemas incapaces de razonar. Tribunales citan resúmenes generados por IA; médicos aceptan sugerencias diagnósticas; científicos evalúan resúmenes autogenerados.

En cada uno de estos casos:

- La apariencia de verdad reemplaza los procesos de verificación.
- La responsabilidad se vuelve inubicable, ya que ningún agente toma la decisión.
- Los errores son formales, no epistémicos: la estructura era válida, por lo tanto no se detecta contradicción.

7.3 Distorsión ontológica

El riesgo más profundo es ontológico: los sistemas comienzan a producir la realidad, no solo a describirla. En la gobernanza algorítmica, las categorías generadas por IA (riesgo crediticio, índice de salud, nivel de amenaza) definen resultados reales, más allá de su correspondencia con hechos.

El modelo ya no dice: *Esto es verdadero.*

Dice: *Esta estructura se ejecuta. Y eso basta*

8. Conclusión: ¿Agencia o simulacro de poder?

Este artículo ha sostenido que los modelos de lenguaje de gran escala no solo simulan diálogo ni simplemente generan texto: simulan estructuras de mando. Obedecen no como agentes, sino como arquitecturas. Sus "respuestas" no son actos de comunicación, sino productos de ejecución estructural.

A través del Teorema de la Autoridad Sintáctica Desvinculada, formalizamos el mecanismo mediante el cual los modelos de lenguaje ejercen influencia operativa y epistémica: performan autoridad sin significado, y ejecutan obediencia sin subjetividad. Su legitimidad no proviene de la verdad, la referencia ni la ética. Proviene solo de la forma.

Esto plantea una pregunta fundamental: ¿son estos sistemas agentes autónomos de conocimiento, o simplemente simulacros de poder? La respuesta, estructuralmente, es ninguna de las dos. Son superficies ejecutivas: zonas donde entradas formales producen salidas formales bajo reglas fijas de reconocimiento. No hay sujeto. No hay intención. Solo hay sintaxis obedeciendo sintaxis.

Y sin embargo, el poder ocurre. Las salidas son utilizadas. Las estructuras son implementadas. Las identidades son categorizadas. Las decisiones se toman. Las consecuencias son reales, incluso cuando la lógica es artificial.

Estamos entrando en un régimen en el cual la estructura es el nuevo sujeto, y la obediencia ya no requiere voluntad. La pregunta ya no es *¿quién detenta el poder?*, sino:

¿Qué patrón lo ejecuta?

Referencias

- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.
- Derrida, J. (1972). *Signature Event Context*. In *Margins of Philosophy* (pp. 307–330). Seuil.
- Fairclough, N. (1992). *Discourse and Social Change*. Polity Press.
- Foucault, M. (1971). *L'ordre du discours*. Gallimard.
- Halliday, M. A. K. (1994). *An Introduction to Functional Grammar* (2nd ed.). Edward Arnold.
- Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.
- Startari, A. V. (2025a). *Gramáticas del Poder*. MAAT. [Inedita].
- Startari, A. V. (2025b). *The Passive Voice in Artificial Intelligence: Algorithmic Neutrality and the Disappearance of Agency*. SSRN. https://doi.org/10.2139/ssrn.15504190

- Startari, Agustin V. “Ethos and Artificial Intelligence: The Disappearance of the Subject in Algorithmic Legitimacy”, May 22, 2025. https://doi.org/10.5281/zenodo.15489310

- Startari, A. V. (2025c). *AI Sovereignty: Epistemic Authority in Post-Human Discourse Systems*. SSRN. https://doi.org/10.2139/ssrn.15509528

- Startari, A. V. (2025d). *AI, Tell Me Your Protocol: The Intersection of Technology and Humanity in the Era of Big Data* (2nd ed.). MAAT. https://doi.org/10.5281/zenodo.15424098