PREPRINT — SSRN / ZENODO
Serie: Arquitecturas de Influencia Algorítmica · Paper 5

La extensión del peligro: desarrollo exponencial de la inteligencia artificial y soberanía cognitiva en el plano de la instrumentación

—

The Extension of the Threat: Exponential Development of Artificial Intelligence and Cognitive Sovereignty in the Plane of Instrumentation

Sergio Bleynat
Microsyst — Consultoría en IA y Transformación Digital
www.microsyst.com.ar

Abril 2026
DOI: [pendiente de asignación en SSRN/Zenodo]

Palabras clave: soberanía cognitiva; instrumentación cognitiva; corpus de entrenamiento; presupuestos filosóficos; forzamiento filosófico; manifiesto Palantir; Anthropic; trampa de segundo orden; variación interindividual; Need for Cognition; Cognitive Reflection Test; Apertura a la experiencia; infraestructura cognitiva pública; tecnofeudalismo; condicionamiento legítimo del privilegio.

Nota metodológica

Este paper sigue la tipología epistémica establecida en Bleynat (2026a) y aplicada con consistencia en toda la serie: distingue explícitamente entre observaciones empíricas con respaldo citable, hipótesis teóricas con argumento lógico pero sin verificación empírica directa, y extrapolaciones analíticas propuestas como agenda de investigación futura. El Registro de Afirmaciones de la Sección 10 materializa esa distinción.

El paper opera además sobre una premisa metodológica adicional que conviene declarar desde el inicio. Su tesis central no introduce un mecanismo nuevo respecto del cuerpo argumental de la serie ni un objeto empírico inédito: introduce una operación analítica —la identificación de una analogía estructural entre dos peligros a la soberanía cognitiva que operan en planos distintos. Esa operación tiene la misma forma jurídico-epistemológica que el Paper 4 ejecutó sobre la categoría de actor privado: muestra que un marco categorial vigente —en este caso, el marco que confina la disputa por la soberanía cognitiva al plano de la percepción social— es estructuralmente insuficiente para abordar el problema en su nivel correcto, y deriva las consecuencias normativas de esa insuficiencia.

Resumen

Los Papers 0 a 4 de esta serie documentaron un peligro estructural a la soberanía cognitiva que opera en un plano específico: el de la percepción social del individuo. Los mecanismos descritos —la Hipótesis de la Alteridad Opcional, la distorsión del muestreo social, el condicionamiento evaluativo distribuido en grafo, el modo alarma como estado basal— son propiedades emergentes de la optimización por engagement de sistemas de recomendación que operan sobre infraestructura constitutivamente híbrida. El Paper 4 estableció el armazón jurídico que vuelve esos mecanismos accesibles a la regulación: bien irreductiblemente común, vicio estructural del consentimiento, tripartición modal de imputación, condicionamiento legítimo del privilegio sostenido.

Este paper declara y demuestra que el desarrollo exponencial de la inteligencia artificial —y específicamente la irrupción de los modelos de lenguaje de gran escala como instrumentos cognitivos de elaboración del pensamiento propio— produce un segundo peligro a la soberanía cognitiva, estructuralmente análogo al primero pero operando en un plano distinto: el plano de la instrumentación cognitiva. La analogía no es metafórica. Es estructural en sentido preciso: en ambos planos hay una infraestructura sociotécnica constitutivamente híbrida que codifica presupuestos no neutrales en su arquitectura profunda, opera sobre vulnerabilidades cognitivas que se distribuyen de forma desigual en la población, y produce dependencia bajo apariencia de potenciación. La extensión del peligro es la jugada conceptual central del paper.

El argumento articula tres tesis. Primera tesis sustantiva: el corpus de entrenamiento, la función de pérdida y las restricciones de seguridad de los modelos de lenguaje son portadores de presupuestos filosóficos, culturales y políticos que no son neutrales y que operan sobre la elaboración del pensamiento del usuario sin que ese usuario pueda detectarlos en su uso ordinario. Segunda tesis estructural: existe una articulación causal precisa entre los dos planos del peligro —el modo alarma basal del Paper 3 degrada en tiempo real las dimensiones psicológicas (Need for Cognition, Cognitive Reflection Test, Apertura a la experiencia, Actively Open-minded Thinking) que determinan tanto la resistencia a los mecanismos del primer plano como la calidad del uso del instrumento cognitivo del segundo. La trampa de segundo orden no es un argumento entre otros: es la pieza arquitectónica que explica por qué los dos peligros son inseparables y por qué los remedios deben pensarse en conjunto. Tercera tesis arquitectónica: la categoría de soberanía cognitiva, articulada como condición de posibilidad de la formación autónoma de juicios, integra los dos planos del peligro en un único objeto regulatorio. La infraestructura cognitiva pública —el conjunto de condiciones institucionales que garantizan acceso no mediado por la función objetivo de actores privados a instrumentos de potenciación cognitiva de calidad— se propone como la consecuencia institucional obligada del reconocimiento de hibridez constitutiva que el Paper 4 estableció, ahora extendida al plano de la instrumentación.

El manifiesto de Palantir Technologies del 18 de abril de 2026 y la negativa pública de Anthropic a colaborar con armas autónomas son los dos primeros casos documentados de un fenómeno que el paper denomina forzamiento filosófico: el momento en que actores que controlan instrumentación cognitiva son obligados, por la dinámica estructural del desarrollo exponencial, a declarar públicamente los presupuestos filosóficos sobre los que operan. Ese forzamiento es la cara visible del desplazamiento conceptual que el paper describe; no su motor. La condición de visibilidad de la disputa filosófica entre actores presupone la existencia previa del objeto disputable —la instrumentación cognitiva como infraestructura— y su reconocimiento como tal.

Abstract

Papers 0 to 4 of this series documented a structural threat to cognitive sovereignty operating in a specific plane: that of the individual's social perception. The mechanisms described —the Optional Alterity Hypothesis, social sampling distortion, distributed evaluative conditioning in graph, the alarm mode as basal state— are emergent properties of the engagement optimization of recommendation systems running on constitutively hybrid infrastructure. Paper 4 established the legal framework that renders these mechanisms accessible to regulation: irreducibly common good, structural vice of consent, modal tripartition of imputation, legitimate conditioning of sustained privilege.

This paper declares and demonstrates that the exponential development of artificial intelligence —and specifically the emergence of large language models as cognitive instruments for the elaboration of individual thought— produces a second threat to cognitive sovereignty, structurally analogous to the first but operating in a distinct plane: the plane of cognitive instrumentation. The analogy is not metaphorical. It is structural in a precise sense: in both planes there is a constitutively hybrid sociotechnical infrastructure that encodes non-neutral assumptions in its deep architecture, operates on cognitive vulnerabilities unevenly distributed across the population, and produces dependence under the appearance of potentiation. The extension of the threat is the paper's central conceptual move.

The paper articulates three theses. First (substantive): the training corpus, loss function, and safety constraints of language models are bearers of philosophical, cultural and political assumptions that are not neutral and that operate on the user's thought elaboration without the user being able to detect them in ordinary use. Second (structural): there exists a precise causal articulation between the two planes of the threat —the basal alarm mode of Paper 3 degrades in real time the psychological dimensions (Need for Cognition, Cognitive Reflection Test, Openness to Experience, Actively Open-minded Thinking) that determine both resistance to first-plane mechanisms and quality of use of the second-plane cognitive instrument. The second-order trap is not one argument among others: it is the architectural piece that explains why the two threats are inseparable and why remedies must be thought together. Third (architectural): the category of cognitive sovereignty, articulated as the

condition of possibility for the autonomous formation of judgments, integrates the two planes of the threat into a single regulatory object. Cognitive public infrastructure is proposed as the obligated institutional consequence of the constitutive hybridity recognition that Paper 4 established, now extended to the plane of instrumentation.

The Palantir Technologies manifesto of April 18, 2026 and the public refusal of Anthropic to collaborate with autonomous weapons are the first two documented cases of a phenomenon the paper calls philosophical forcing: the moment when actors controlling cognitive instrumentation are compelled, by the structural dynamics of exponential development, to publicly declare the philosophical assumptions on which they operate. That forcing is the visible face of the conceptual displacement the paper describes; not its driver.

1. El problema: un segundo plano de peligro que el marco vigente no nombra

Toda esta serie ha argumentado, con cinco papers acumulativos, que existe un peligro estructural a la soberanía cognitiva que la teoría regulatoria contemporánea no nombra con precisión y que, por esa razón, no puede abordar con instrumentos eficaces. Los Papers 0 a 3 describieron los mecanismos por los que ese peligro opera sobre la percepción social del individuo: la selección algorítmica de versiones filtradas de los contactos conocidos, la distorsión del muestreo de instancias sociales que alimenta la inferencia de prevalencia, la transferencia de valencia evaluativa hacia objetos políticos sin persuasión directa, la consolidación del modo alarma como estado basal del sistema. El Paper 4 mostró que esos mecanismos no son externalidades de mercado sino propiedades de una infraestructura constitutivamente híbrida, y construyó el armazón jurídico —bien irreductiblemente común, vicio estructural del consentimiento, tripartición modal de imputación, condicionamiento legítimo del privilegio sostenido— que vuelve esos mecanismos accesibles a la regulación.

Ese armazón conceptual, sin embargo, opera enteramente sobre un plano: el plano en el que la infraestructura algorítmica es infraestructura de percepción social. Los sistemas de recomendación —Facebook, Instagram, X, TikTok— modifican la disponibilidad cognitiva del entorno social del individuo: qué muestreo realiza, qué nodos del grafo le llegan, en qué versión, con qué activación. La función objetivo que el Paper 4 propone modificar es la función objetivo de la distribución del contenido social. La regulación que el Paper 4 propone como condicionamiento legítimo del privilegio sostenido es regulación sobre la distribución de la percepción social mediada por la plataforma.

La tesis de este paper es que el desarrollo exponencial de la inteligencia artificial —y específicamente la irrupción de los modelos de lenguaje de gran escala como instrumentos disponibles a escala masiva para la elaboración del pensamiento individual— produce un segundo peligro estructural a la

soberanía cognitiva, que opera en un plano distinto del descrito por los papers anteriores y que, por esa razón, los instrumentos jurídicos del Paper 4 son necesarios pero no suficientes para abordarlo.

El plano nuevo no es el de la percepción social del entorno: es el plano de la instrumentación cognitiva del pensamiento propio. Cuando un individuo conversa con un modelo de lenguaje para refinar un argumento, explorar un concepto, escribir un texto o tomar una decisión, no está siendo expuesto a una distribución algorítmica de contenido producido por otros: está usando deliberadamente un instrumento que codifica, en su arquitectura profunda, presupuestos filosóficos, culturales y políticos que no eligió y, en muchos casos, no puede detectar. La elaboración del pensamiento, que en la tradición liberal-clásica ha sido entendida como ejercicio interno de la facultad individual y como condición de posibilidad de toda autonomía, se produce ahora —cada vez con mayor intensidad y para una porción creciente de la población— a través de un instrumento cuya arquitectura no es neutral y cuyos presupuestos no están sujetos a ningún régimen de transparencia ni de gobernanza.

La tesis sustantiva de este paper es, en consecuencia, que existe una analogía estructural entre los dos peligros. La palabra «analogía» exige aquí precisión: no se trata de una similitud metafórica ni de un paralelismo retórico. Se trata de que los dos peligros comparten una estructura formal idéntica en cinco dimensiones específicas que la Sección 3 desarrolla. Compartir esa estructura no significa que los peligros sean idénticos: operan sobre objetos distintos, mediante mecanismos parcialmente distintos, y exigen instrumentos regulatorios parcialmente distintos. Pero la estructura formal común es lo que permite reconocer al segundo peligro como peligro a la soberanía cognitiva en sentido estricto, y no como un problema separado que exige una categoría conceptual nueva.

El reconocimiento de esta analogía estructural tiene una consecuencia inmediata para la teoría regulatoria. Si los marcos vigentes son insuficientes para el primer plano del peligro —tesis del Paper 4—, son a fortiori insuficientes para el segundo, porque el segundo plano no aparece siquiera como objeto regulatorio en los marcos disponibles. El AI Act europeo, las executive orders estadounidenses, las propuestas de gobernanza multilateral operan sobre el supuesto de que el problema de la inteligencia artificial es un problema de seguridad técnica del sistema —outputs sesgados, fallos, riesgos de uso malicioso— y no un problema de codificación arquitectónica de presupuestos no neutrales en instrumentos cognitivos masivos. Regulan comportamientos del sistema, no la función objetivo y los presupuestos profundos que producen esos comportamientos. Exigen transparencia sobre outputs, no sobre el corpus de entrenamiento que determina qué tipos de pensamiento el sistema amplifica con facilidad y cuáles con resistencia.

El paper que el lector tiene ante sí cumple, en consecuencia, una función arquitectónica precisa en la economía de la serie. Hace explícita la operación analítica que la serie venía preparando: el reconocimiento de que la categoría de soberanía cognitiva, introducida en los papers anteriores como horizonte normativo, no agota su contenido en el plano de la percepción social y debe ser articulada también sobre el plano de la instrumentación cognitiva. Y deriva de esa articulación las

consecuencias institucionales que el armazón del Paper 4, pensado para el primer plano, no puede producir por sí solo: la infraestructura cognitiva pública como categoría institucional necesaria, no como añadido normativo.

El paper se organiza del siguiente modo. La Sección 2 examina lo que los marcos teóricos existentes capturan y lo que dejan sin capturar respecto del segundo plano del peligro. La Sección 3 desarrolla la jugada conceptual central: la articulación formal de la analogía estructural entre los dos planos del peligro a la soberanía cognitiva, en sus cinco dimensiones. La Sección 4 detalla los tres mecanismos por los que el peligro de Plano 2 opera: la codificación filosófica del corpus, el forzamiento filosófico declarado por los actores que controlan los instrumentos, y la efectividad diferencial sobre la variación interindividual. La Sección 5 examina la articulación causal entre los dos planos —la trampa de segundo orden— como pieza arquitectónica del argumento. La Sección 6 articula la categoría de soberanía cognitiva como objeto regulatorio unificador y deriva la infraestructura cognitiva pública como consecuencia institucional. La Sección 7 responde a las objeciones previsibles. La Sección 8 examina las implicaciones para la teoría regulatoria, incluyendo la conexión con el Paper 6, donde el peligro descrito alcanza su forma más radical: el colapso de la distinguibilidad misma de la fuente humana de la señal. La Sección 9 cierra con las conclusiones y predicciones empíricas falsificables. Las Secciones 10 y 11 contienen el registro de afirmaciones y las referencias.

2. Estado del arte: el segundo plano que ningún marco existente ocupa con precisión

2.1 La gobernanza de IA y su sesgo de seguridad técnica

Los marcos de gobernanza de inteligencia artificial más desarrollados —el AI Act europeo (2024), las executive orders de la administración estadounidense, las propuestas de organismos multilaterales como UNESCO y OCDE— comparten un supuesto que los papers anteriores de esta serie permiten identificar como limitación estructural: asumen que el problema de la IA es un problema de seguridad técnica del sistema y no un problema de arquitectura epistémica. Regulan comportamientos del sistema —outputs sesgados, fallos identificables, riesgos de uso malicioso— sin tocar la función objetivo que produce esos comportamientos ni los presupuestos codificados en el corpus de entrenamiento que determinan qué tipos de pensamiento el sistema amplifica con facilidad y cuáles requieren esfuerzo adicional para producir.

La literatura académica sobre riesgos de IA ha producido análisis valiosos sobre alineación, seguridad y gobernanza (Russell, 2019; Bostrom, 2014; Gabriel, 2020), pero ha concentrado su atención en dos polos del problema: el polo de los sistemas más capaces —inteligencia artificial general, riesgos existenciales, alineación de superinteligencia— y el polo de los efectos discriminatorios identificables sobre poblaciones específicas —sesgos de género, racial,

socioeconómico en clasificación, scoring crediticio, predicción de reincidencia. Ese rango de atención ha dejado comparativamente subatendido el espacio intermedio en el que esta serie opera y en el que este paper desarrolla su tesis: los efectos cognitivos y epistémicos de los sistemas de IA generativa e inferencial sobre poblaciones generales en el presente inmediato, mediados por el uso ordinario y aparentemente benigno de los instrumentos.

2.2 La crítica del capitalismo de plataformas y su limitación de plano

La tradición crítica más completa sobre el modelo de negocio de las plataformas algorítmicas —Zuboff (2019) sobre el capitalismo de vigilancia, Srnicek (2016) sobre el capitalismo de plataformas, Couldry y Mejias (2019) sobre la colonización de la experiencia, Varoufakis (2023) sobre el tecnofeudalismo— identifica con precisión la lógica del problema en el plano de la extracción y modulación del comportamiento del usuario. Pero esa tradición opera enteramente sobre el plano que los Papers 0 a 4 describen: el de la infraestructura algorítmica como sistema de modulación de la percepción y el comportamiento. El segundo plano —el de la instrumentación cognitiva del pensamiento propio mediante modelos de lenguaje de gran escala— aparece ocasionalmente como tema lateral pero no como objeto teórico central.

Varoufakis (2023) merece una mención específica porque su categoría de tecnofeudalismo, articulada para describir el desplazamiento estructural del capitalismo industrial al capitalismo de la renta de plataforma, se vuelve aún más precisa cuando se la aplica al segundo plano del peligro. En el capitalismo industrial, el trabajador vendía fuerza de trabajo a un capitalista que la organizaba productivamente. En el tecnofeudalismo de las plataformas de distribución, el usuario provee comportamiento al señor feudal digital que extrae renta de ese comportamiento. En el tecnofeudalismo extendido al plano de la instrumentación cognitiva, el individuo accede a un instrumento de potenciación del pensamiento cuya arquitectura es propiedad de un señor feudal cognitivo, y la elaboración del pensamiento individual se produce en ese feudo. La extensión de la categoría de Varoufakis al segundo plano es directa, pero no fue desarrollada por Varoufakis y no aparece sistematizada en la literatura disponible.

2.3 La filosofía de la técnica y el problema de la neutralidad de los artefactos

Una tradición a la que la literatura sobre regulación de plataformas raramente recurre con sistematicidad y que este paper requiere movilizar para sostener su tesis sustantiva sobre la no-neutralidad del corpus de entrenamiento es la filosofía de la técnica que examina la inscripción política en los artefactos. Winner (1980) argumentó, en su clásico «¿Tienen política los artefactos?», que ciertos artefactos técnicos tienen política en sentido fuerte: incorporan en su diseño material la cristalización de relaciones de poder, decisiones políticas y presupuestos culturales que continúan operando con independencia de la intención de quienes los utilizan. Latour (2005) desarrolló esa

tesis hasta el extremo de la teoría del actor-red, en la que los artefactos son actores en sentido pleno, no instrumentos neutrales utilizados por actores humanos.

La tradición de filosofía de la técnica provee el marco categorial general dentro del cual la tesis sobre el corpus de entrenamiento de los modelos de lenguaje se vuelve teóricamente articulable. Un modelo de lenguaje no es un instrumento neutral con presupuestos accidentalmente sesgados. Es un artefacto político en el sentido más preciso de Winner: la cristalización computacional de un corpus históricamente situado, una función de pérdida que selecciona ciertas regularidades y descarta otras, y un conjunto de restricciones de seguridad que reflejan compromisos éticos y políticos específicos. Esa cristalización produce, en uso ordinario, efectos sobre la elaboración del pensamiento del usuario que son consecuencias de la arquitectura, no fallos de operación. La filosofía de la técnica disponible permite formular esa observación; lo que no provee es el armazón regulatorio que vuelve esa observación accesible al diseño institucional, que es lo que este paper se propone construir, en analogía con lo que el Paper 4 construyó para el primer plano del peligro.

2.4 El espacio vacío

Ninguna de las tradiciones examinadas conecta los hechos sobre la naturaleza híbrida de la infraestructura algorítmica —documentados por el Paper 4— con la extensión del peligro al plano de la instrumentación cognitiva. La gobernanza de IA opera sobre seguridad técnica sin reconocer el plano arquitectónico-político del corpus. La crítica del capitalismo de plataformas opera sobre el plano de la modulación del comportamiento sin extenderse sistemáticamente a la instrumentación del pensamiento. La filosofía de la técnica provee el marco general pero no el armazón regulatorio. La teoría regulatoria de plataformas opera sobre el supuesto de actor privado sin examinarlo —y, cuando lo examina (Paper 4), no extiende el análisis al plano cognitivo. El trabajo de conectar esos cuatro dominios y proyectarlos sobre el problema específico de la soberanía cognitiva en el plano de la instrumentación es la operación que este paper realiza.

3. La jugada conceptual central: la analogía estructural entre dos peligros que afectan a la soberanía cognitiva

Esta sección desarrolla la jugada arquitectónica que sostiene todo el paper: la articulación formal de la analogía estructural entre el peligro descrito por los Papers 0 a 4 y el peligro que el desarrollo exponencial de la IA introduce en el plano de la instrumentación cognitiva. La analogía no es retórica ni heurística. Es estructural en sentido formal preciso: los dos peligros comparten una estructura común en cinco dimensiones específicas. Reconocer esa estructura común es lo que permite tratar al segundo peligro como peligro a la soberanía cognitiva en sentido estricto y no como un problema separado que exigiría una categoría conceptual independiente.

3.1 El primer peligro recapitulado: soberanía cognitiva en el plano de la percepción social

Los Papers 0 a 4 documentaron, con creciente precisión, un peligro estructural a la soberanía cognitiva del individuo en su capacidad de formar juicios autónomos sobre la realidad social de su entorno. El Paper 0 (Bleynat, 2026a) describió el mecanismo más básico —la Hipótesis de la Alteridad Opcional—: el algoritmo selecciona, para cada observador, las versiones más resonantes de cada contacto, produciendo soledad de presencia bajo la apariencia de conexión densa. El Paper 1 (Bleynat, 2026b) mostró que ese mecanismo opera en el nivel del Social Sampling Model: la muestra de instancias sociales que el individuo utiliza para inferir la realidad de su entorno está sistemáticamente sesgada hacia las dimensiones de alta activación de sus contactos. El Paper 2 (Bleynat, 2026c) describió el Condicionamiento Evaluativo Distribuido en Grafo: la transferencia de valencia afectiva desde el contacto de confianza hacia objetos políticos con los que coexiste en el feed, sin persuasión directa ni intención del contacto. El Paper 3 (Bleynat, 2026d) documentó el modo alarma como estado basal del sistema: la optimización por engagement selecciona contenido de alta activación emocional, lo que produce un estado de alarma crónica que reduce la tolerancia a la ambigüedad, favorece el procesamiento heurístico sobre el deliberativo y activa el mecanismo de anestesia epistémica. El Paper 4 (Bleynat, 2026e) construyó el armazón jurídico que vuelve esos mecanismos accesibles a la regulación: bien irreductiblemente común, vicio estructural del consentimiento, tripartición modal de imputación, condicionamiento legítimo del privilegio sostenido.

El peligro a la soberanía cognitiva en este plano puede formularse con precisión: el individuo expuesto a la operación normal de los sistemas de recomendación pierde acceso, sin saberlo y sin que ninguna falla individual pueda explicarlo, a una representación no sistemáticamente distorsionada de su entorno social. Forma juicios sobre lo que «todos piensan», sobre quién es cada uno de sus contactos, sobre qué es una posición política y qué es otra, a partir de una muestra cuya distorsión es propiedad emergente de la función objetivo del sistema. La soberanía cognitiva —entendida como capacidad efectiva de formar juicios autónomos sobre la realidad social— es comprometida por la mediación de un instrumento cuyo diseño no eligió y cuya operación no puede detectar mientras la experimenta.

3.2 El segundo peligro declarado: soberanía cognitiva en el plano de la instrumentación cognitiva

El desarrollo exponencial de los modelos de lenguaje de gran escala —desde la irrupción de ChatGPT a finales de 2022 hasta los sistemas conversacionales de uso masivo del presente— introduce un segundo plano del peligro a la soberanía cognitiva, distinto del anterior pero estructuralmente análogo. El plano nuevo es el de la instrumentación cognitiva: el uso del modelo como instrumento

deliberado para la elaboración del pensamiento propio, la exploración de conceptos, la escritura, el razonamiento práctico, la toma de decisiones.

El peligro en este plano puede formularse con precisión análoga al peligro del primer plano: el individuo que utiliza un modelo de lenguaje para elaborar pensamiento elabora ese pensamiento mediante un instrumento cuya arquitectura profunda codifica presupuestos filosóficos, culturales y políticos que no son neutrales —que no fueron elegidos por el usuario, que en general el usuario no puede detectar en uso ordinario, y que operan sobre la elaboración del pensamiento de forma estructurada. El instrumento, por su naturaleza, amplifica con facilidad ciertos tipos de razonamiento y requiere esfuerzo adicional para producir otros; sugiere ciertas analogías con naturalidad y desplaza otras hacia la periferia; aplica por defecto ciertos marcos conceptuales y resiste otros. Esa fluidez diferencial no es un fallo de diseño que pudiera corregirse: es la consecuencia inevitable del entrenamiento sobre cualquier corpus históricamente situado bajo cualquier función de pérdida específica.

La consecuencia para la soberanía cognitiva es directa: el individuo que utiliza el instrumento cree estar elaborando pensamiento propio y, en un sentido estricto, lo está haciendo —su agencia no se anula. Pero la elaboración se produce en un espacio estructurado por presupuestos que el individuo no eligió y, en muchos casos, no puede siquiera detectar mientras los habita. La autonomía deliberativa, entendida en la tradición liberal-clásica más exigente como capacidad efectiva del individuo de razonar a partir de premisas que reconoce como suyas, queda comprometida en una dimensión estructural distinta de las descritas por los Papers 0 a 4 pero formalmente análoga.

3.3 La estructura formal de la analogía: cinco dimensiones

La analogía entre los dos peligros es estructural en sentido formal preciso. Comparten cinco dimensiones específicas que conviene desarrollar individualmente porque cada una habilita parte del armazón regulatorio que el paper construye en las secciones siguientes.

Dimensión 1 — Hibridez constitutiva de la infraestructura

Tanto la infraestructura de distribución (Plano 1) como la infraestructura de instrumentación cognitiva (Plano 2) son constitutivamente híbridas en el sentido preciso establecido por el Paper 4: construidas y sostenidas por la convergencia de financiamiento estatal histórico, subsidios operativos presentes y conocimiento académico producido con financiamiento público. La infraestructura computacional sobre la que operan los modelos de lenguaje es la misma que el Paper 4 documentó —ARPANET, NSFNET, los algoritmos fundacionales, el ecosistema de inversión de inteligencia. Las arquitecturas transformer fundacionales del campo —desde Vaswani et al. (2017) hasta los modelos de lenguaje de gran escala contemporáneos— fueron desarrolladas en investigación financiada mayoritariamente por instituciones académicas con financiamiento público y por laboratorios privados que operan bajo regímenes de subsidio fiscal en múltiples jurisdicciones.

La Ley de Economía del Conocimiento argentina, el programa Horizon Europe, los Government-Guided Investment Funds chinos —documentados en el Paper 4— subsidian igualmente la operación de plataformas de distribución y de laboratorios de IA. El conocimiento de procesamiento del lenguaje natural, los datos masivos sobre los que los modelos se entrenan, el talento humano que los diseña y opera son producto de la misma matriz de inversión pública y privada que el Paper 4 documentó. La hibridez constitutiva es propiedad de la infraestructura algorítmica-inferencial considerada como totalidad, no privilegio del subsistema de distribución.

Dimensión 2 — Codificación de presupuestos no neutrales en la arquitectura

Tanto en el Plano 1 como en el Plano 2, la infraestructura no es neutral respecto del objeto sobre el que opera. La función objetivo de los sistemas de recomendación —optimización por engagement— produce las propiedades emergentes que los Papers 1 a 3 describen sin que ningún actor individual las haya diseñado intencionalmente. La función de pérdida, el corpus de entrenamiento, las restricciones de seguridad y el procedimiento de fine-tuning de los modelos de lenguaje codifican, en su arquitectura profunda, presupuestos filosóficos, culturales, lingüísticos y políticos que no son neutrales. Esos presupuestos producen, en uso ordinario, efectos sobre la elaboración del pensamiento del usuario que son consecuencias arquitectónicas, no fallos. La no-neutralidad estructurada es propiedad común a los dos planos.

Dimensión 3 — Operación bajo umbral de detección consciente del usuario

En ambos planos, la operación del sistema sobre la cognición del individuo se produce, en condiciones ordinarias de uso, por debajo del umbral de detección consciente. El usuario que mira su feed no experimenta la distorsión del muestreo como distorsión: experimenta una representación de la realidad social de su red. El usuario que conversa con un modelo de lenguaje no experimenta la codificación de presupuestos del corpus como codificación de presupuestos: experimenta una conversación con un instrumento aparentemente neutral que le ayuda a pensar. La condición de invisibilidad operativa es, en ambos planos, lo que distingue al peligro de las formas tradicionales de influencia —propaganda, persuasión directa, censura— que operan bajo el umbral de detección crítica del receptor competente. Esa invisibilidad es lo que el Paper 4 articuló jurídicamente como vicio estructural del consentimiento, y se aplica sin modificación sustancial al segundo plano.

Dimensión 4 — Efectividad diferencial sobre vulnerabilidades cognitivas variables

Tanto el peligro del Plano 1 como el del Plano 2 operan con efectividad diferencial sobre dimensiones psicológicas individuales documentadas por la literatura empírica disponible. Los mecanismos de los Papers 1 a 3 son más eficientes sobre perfiles con bajo Need for Cognition basal, baja Apertura a la experiencia, alto Neuroticismo y baja puntuación en el Cognitive Reflection Test, exactamente donde las defensas elaborativas son menores y donde el modo alarma produce mayor reducción funcional de la ruta central de procesamiento. Los presupuestos del corpus de los modelos de

lenguaje operan con mayor o menor eficacia sobre la elaboración del pensamiento del usuario en función de las mismas dimensiones, por una razón que la Sección 5 desarrolla y que es la pieza arquitectónica más importante de este paper: las dimensiones que protegen al individuo de los mecanismos de Plano 1 son exactamente las dimensiones que determinan la calidad del uso del instrumento de Plano 2 como instrumento de genuina potenciación del pensamiento.

Dimensión 5 — Producción de dependencia bajo apariencia de potenciación

Ambos planos del peligro comparten una propiedad fenomenológica que conviene articular con cuidado: la operación del sistema produce dependencia funcional del individuo respecto del instrumento, bajo apariencia subjetiva de potenciación o de ampliación de capacidad. El usuario de redes sociales experimenta acceso a información, a contactos, a perspectivas que sin la plataforma no tendría —y esa experiencia es, en algún sentido, verdadera; el usuario tiene acceso a algo. El usuario de modelos de lenguaje experimenta capacidad de elaboración intelectual que sin el instrumento no tendría —y esa experiencia es también verdadera en algún sentido. Pero en ambos casos la potenciación experimentada coexiste con la producción estructural de dependencia: la pérdida progresiva de la capacidad de operar sobre el plano correspondiente sin la mediación del instrumento, y la transformación de la mediación en condición de la operación. La distinción analítica entre potenciación genuina y dependencia funcional bajo apariencia de potenciación es central para el armazón normativo y se desarrollará con precisión en la Sección 5.

3.4 Lo que la analogía no afirma

Conviene precisar, antes de proseguir, lo que la analogía estructural propuesta no afirma, para evitar objeciones predecibles que descansan en lecturas inflacionistas de la tesis. La analogía no afirma que los dos peligros sean idénticos. Operan sobre objetos distintos —la percepción social del entorno en un caso, la elaboración del pensamiento propio en el otro—, mediante mecanismos parcialmente distintos y exigen instrumentos regulatorios parcialmente distintos. La analogía no afirma que los modelos de lenguaje sean redes sociales: son tecnologías estructuralmente diferentes con modelos de negocio distintos y dinámicas de uso distintas. La analogía no afirma que el uso de modelos de lenguaje sea siempre y necesariamente perjudicial: el Fenómeno 3 que la versiones previas de este paper articuló —la democratización parcial del acceso a interlocución cognitiva de alto nivel— es real y tiene consecuencias positivas sustantivas, en particular para poblaciones que históricamente carecieron de acceso a redes intelectuales densas.

Lo que la analogía afirma es algo más preciso y, por eso, más exigente: que la estructura formal de los dos peligros es la misma en las cinco dimensiones examinadas, y que esa coincidencia estructural permite y exige tratar al segundo plano como objeto regulatorio dentro de la categoría de soberanía cognitiva, no como problema separado en una categoría conceptual independiente. La unificación

categorial es la consecuencia teórica de la analogía. Las consecuencias institucionales de esa unificación —que la Sección 6 articula— son su consecuencia práctica.

4. Los tres mecanismos del peligro de Plano 2

Una vez establecida la analogía estructural, esta sección detalla los tres mecanismos específicos por los que el peligro de Plano 2 opera. La estructura es paralela a la que los Papers 1, 2 y 3 articularon para el Plano 1: cada mecanismo describe una dimensión del modo en que el instrumento de instrumentación cognitiva produce efectos sobre la elaboración del pensamiento del usuario.

4.1 Mecanismo 1 — La codificación filosófica del corpus

Un modelo de lenguaje es, en su dimensión más básica, una destilación estadística del corpus sobre el que fue entrenado, mediada por una función de pérdida específica y refinada por procedimientos de fine-tuning que reflejan compromisos éticos y políticos del laboratorio que lo produce. Esa destilación no es neutral: el corpus refleja los marcos conceptuales, las categorías de análisis, las jerarquías de valor, las regularidades estilísticas y, crucialmente, las ausencias propias de la tradición intelectual histórica que lo produjo. Los modelos más capaces disponibles globalmente fueron entrenados mayoritariamente sobre texto en inglés, producido en contextos anglosajones predominantemente del último siglo, que refleja los presupuestos del liberalismo de posguerra y del individualismo metodológico característicos de la academia occidental contemporánea, con todas sus contradicciones internas. No por conspiración ni por intención política deliberada de los laboratorios: por disponibilidad. El texto digitalizado en inglés de calidad apta para entrenamiento supera en varios órdenes de magnitud al disponible en cualquier otra lengua o tradición intelectual.

El resultado observable —que conviene formular como hipótesis, dado que la verificación directa requeriría acceso a los sistemas de evaluación internos de los laboratorios y a benchmarks específicamente diseñados para detectar el fenómeno— es que los modelos producen con mayor fluidez ciertos tipos de razonamiento que otros. Las analogías que sugieren con naturalidad, los marcos conceptuales que aplican por defecto, las preguntas que responden sin resistencia y las que requieren formulación específica para producir respuesta sustantiva, las posiciones que asumen por defecto en debates filosóficos no resueltos, las tradiciones intelectuales que tratan como punto de partida natural y las que tratan como excepciones a justificar, todas ellas reflejan la distribución del corpus de entrenamiento y los compromisos del fine-tuning. Esa fluidez diferencial no es un fallo de diseño que pudiera corregirse mediante ajustes superficiales. Es consecuencia arquitectónica del entrenamiento sobre cualquier corpus históricamente situado bajo cualquier función de pérdida específica.

La consecuencia para la elaboración del pensamiento del usuario es estructural. Un usuario que conversa con el modelo en una lengua, sobre una tradición intelectual o sobre un marco conceptual

sub-representado en el corpus de entrenamiento experimenta el instrumento como menos fluido, menos sugerente, menos productivo en esa zona, y más fluido, más sugerente, más productivo en las zonas más representadas. La asimetría se traduce, sin que el usuario la detecte, en una presión gravitacional sobre el espacio de elaboración del pensamiento: las zonas de mayor fluidez del instrumento atraen el pensamiento del usuario hacia ellas con la fuerza de la economía cognitiva, y las zonas de menor fluidez quedan progresivamente abandonadas o se vuelven el lugar donde el usuario debe esforzarse para sostener una elaboración que el instrumento no facilita. La codificación del corpus no censura ningún tipo de pensamiento: lo desincentiva diferencialmente. El efecto agregado, sostenido en el tiempo y a escala poblacional, es la homogeneización progresiva de los marcos conceptuales sobre los que se elabora el pensamiento, en la dirección de los presupuestos del corpus dominante.

El argumento no se reduce, sin embargo, al sesgo lingüístico-cultural del corpus. La función de pérdida y los procedimientos de fine-tuning añaden una segunda capa de codificación filosófica: las restricciones de seguridad —qué tipos de respuesta están permitidos, qué temas están restringidos, qué posiciones políticas el modelo asume por defecto, qué conflictos morales el modelo resuelve hacia un lado u otro— reflejan compromisos éticos y políticos específicos del laboratorio. Esos compromisos pueden ser razonables o no, defendibles o no; lo que importa para el argumento de este paper es que existen, no son neutrales, y operan sobre la elaboración del pensamiento del usuario sin que ningún régimen de transparencia o de gobernanza los haga visibles ni los someta a deliberación pública.

4.2 Mecanismo 2 — El forzamiento filosófico como evidencia visible del desplazamiento

El Mecanismo 1 opera estructuralmente y, por su naturaleza, es difícil de visibilizar a la opinión pública: requiere análisis técnico de los corpus y los procedimientos de fine-tuning, comparaciones sistemáticas entre modelos entrenados con presupuestos distintos, benchmarks específicos para detectar diferencias en la facilidad con que se produce ciertos tipos de razonamiento. La operación del Mecanismo 1 no produce, por sí sola, evidencia pública directa de su existencia.

El Mecanismo 2 que esta subsección describe es la cara visible del desplazamiento que el paper articula. Lo denominamos forzamiento filosófico porque describe un fenómeno preciso: el momento en que actores que controlan instrumentación cognitiva son obligados, por la dinámica estructural del desarrollo exponencial de la IA, a declarar públicamente los presupuestos filosóficos y políticos sobre los que operan. Antes de la masividad reciente de los modelos de lenguaje, los presupuestos filosóficos del corpus operaban en el silencio del producto técnico que parecía neutral. La masividad y la visibilidad creciente del impacto de los modelos sobre la elaboración del pensamiento de poblaciones enteras vuelve insostenible esa neutralidad declarada y produce, como respuesta

racional de los actores relevantes, declaraciones explícitas de los presupuestos sobre los que cada uno opera.

El concepto de economía del silencio puede entenderse como la propiedad protectora del régimen de no-declaración: durante décadas, los actores que operaban en la zona de hibridez constitutiva entre capital privado, Estado, sistemas de inteligencia y mercados tecnológicos mantenían opacidad sobre sus presupuestos porque el silencio reducía el costo regulatorio, evitaba la atribución de responsabilidad y preservaba la ficción de neutralidad técnica que legitimaba el marco de actor privado puro. Esa economía del silencio está siendo perturbada por la dinámica del desarrollo exponencial de la IA, y el forzamiento filosófico es la consecuencia. La causalidad que sostiene la perturbación es plausible aunque no completamente verificable desde fuera: la masividad del impacto de los modelos sobre la elaboración del pensamiento, la creciente capacidad técnica de auditoría algorítmica externa, y la presión política específica que el caso de Anthropic ilustra, convergen para hacer que la opacidad sea más costosa que la declaración.

Conviene examinar los dos casos paradigmáticos disponibles con la precisión que cada uno exige, sin forzar una simetría allí donde las diferencias estructurales resultan decisivas.

El caso Palantir como declaración ofensiva

El 18 de abril de 2026, la cuenta corporativa oficial de Palantir Technologies publicó en X un documento de 22 puntos presentado como resumen de The Technological Republic, libro de su director ejecutivo Alex Karp. El documento no es analíticamente relevante por sus contenidos específicos —que incluyen afirmaciones sobre deudas morales de Silicon Valley con el Estado, armas autónomas, jerarquías entre culturas y el fin de la neutralización de posguerra de Alemania y Japón— sino por lo que su forma representa: la primera declaración pública explícita, firmada con logo corporativo y distribuida a escala global, de una corporación tecnológica que reivindica protagonismo filosófico y político como programa consciente y deseable. No como accidente. No como consecuencia no buscada. Como objetivo.

El caso Palantir es una declaración ofensiva en sentido estructural. La empresa, cuya trayectoria desde su fundación en 2003 con financiamiento de In-Q-Tel y de Founders Fund la ubica en el centro mismo de la zona de hibridez constitutiva que el Paper 4 documentó, decide salir primero, con su propio encuadre, reivindicando explícitamente la filosofía política que hasta ahora operaba sin nombre. La pregunta analíticamente relevante no es qué dice Karp. Es por qué ahora, por qué en este formato, y qué condiciones produjeron el momento en que lo que antes requería silencio puede decirse en voz alta. La respuesta plausible es que el cálculo estratégico de Palantir reconoce que la auditoría externa de su operación se está volviendo técnicamente posible y políticamente probable, y que en esas condiciones es preferible producir el encuadre desde adentro a esperar que se imponga desde afuera. Salir a declarar la propia filosofía antes de que alguien más la infiera de la propia conducta y la impute con un encuadre no elegido es racionalmente superior a esperar la exposición.

El caso Anthropic como acto operativo bajo presión política específica

La negativa pública de Anthropic a levantar sus restricciones frente al uso de sus modelos en armas plenamente autónomas y vigilancia doméstica masiva —documentada en declaraciones públicas de la empresa durante 2026— es un acto de naturaleza estructuralmente distinta al manifiesto de Palantir, y conviene precisar la diferencia con cuidado para evitar una simetría interpretativa que induciría una falsa equivalencia analítica.

El acto de Anthropic no es una declaración ofensiva de filosofía política integral análoga al manifiesto de Karp. Es, en sentido estricto, una decisión operativa bajo presión política específica que tiene consecuencias filosóficas pero cuya estructura no es la de la reivindicación pública de un programa filosófico-político integral. La empresa enfrentó una solicitud específica de la administración estadounidense, evaluó esa solicitud contra los principios de uso que ya tenía documentados desde su fundación, y decidió declinar públicamente. La declinación tiene contenido filosófico —en el sentido de que se justifica en compromisos éticos sustantivos sobre el uso responsable de inteligencia artificial— pero su estructura es la de una negativa operativa a un uso particular del producto, no la de una declaración programática sobre el papel filosófico-político que la empresa quiere ocupar en el ordenamiento global.

La distinción importa para el argumento. El forzamiento filosófico que el paper describe no produce dos polos simétricos de una misma escala —Palantir en un extremo, Anthropic en el otro. Produce, más precisamente, dos respuestas estructuralmente distintas al mismo desplazamiento: una respuesta ofensiva (declaración programática integral) y una respuesta defensiva (negativa operativa con contenido filosófico). Ambas son respuestas al mismo desplazamiento estructural —la opacidad ya no es estratégicamente sostenible para los actores con instrumentación cognitiva masiva—, pero son respuestas de distinto tipo, y tratarlas como simétricas oscurece la naturaleza del fenómeno. La economía del silencio se está colapsando, sí; pero el colapso no produce un único formato de declaración. Produce el espectro de formatos posibles que cualquier actor obligado a tomar posición pública puede adoptar, según su modelo de negocio, su exposición política específica y los compromisos previos sobre los que opera.

Lo que el forzamiento filosófico revela sobre el desplazamiento

Lo que ambos casos —el ofensivo y el defensivo— revelan en común no es la simetría del espectro sino el desplazamiento mismo: la transformación de la instrumentación cognitiva en objeto de disputa filosófico-política pública. Hasta ese momento, la pregunta filosófica fundamental sobre los modelos de lenguaje —qué presupuestos codifican, sobre qué corpus, con qué función de pérdida, con qué restricciones de seguridad, en beneficio de qué visión del mundo— operaba dentro del laboratorio, en publicaciones técnicas con audiencia restringida y sin presión política específica para hacerla explícita. La masividad del uso, la creciente capacidad de auditoría externa y la presión geopolítica producen el desplazamiento: los presupuestos pasan a ser objeto de disputa pública. Ese

desplazamiento es la condición de posibilidad del forzamiento filosófico, y el forzamiento es su manifestación visible.

El forzamiento filosófico es, en consecuencia, evidencia del desplazamiento del objeto disputable, no su motor. La instrumentación cognitiva pasa a ser disputable porque se vuelve estructuralmente más difícil de mantener en la opacidad técnica. Una vez disputable, los actores que la controlan deben tomar posición. La forma de la posición varía. La existencia del momento de tomarla no varía: es estructural al desplazamiento.

4.3 Mecanismo 3 — La efectividad diferencial sobre la variación interindividual

El tercer mecanismo describe la dimensión más delicada del peligro de Plano 2 y conviene articularla con la precisión que su delicadeza requiere para evitar lecturas que la confundan con argumentos jerarquistas que estructuralmente la contradicen. La literatura psicológica empírica disponible documenta, con alta fiabilidad y replicabilidad transcultural, un conjunto de dimensiones individuales que muestran variación sustancial en poblaciones humanas y que son directamente relevantes para los mecanismos de Plano 2.

La primera de esas dimensiones es la Apertura a la experiencia del modelo Big Five (Costa y McCrae, 1992), ya central en los Papers 1 y 2 de esta serie, que correlaciona con tolerancia a la ambigüedad, flexibilidad metacognitiva y preferencia por el procesamiento deliberativo sobre el heurístico. La segunda es el Need for Cognition (NFC), constructo propuesto por Cacioppo y Petty (1982) y sistematizado en décadas de investigación posterior (Cacioppo, Petty, Feinstein y Jarvis, 1996): la tendencia estable a involucrarse y disfrutar del procesamiento cognitivo esforzado. Los individuos con NFC alto tienden a usar la ruta central del Elaboration Likelihood Model (Petty y Cacioppo, 1986) evaluando argumentos por sus méritos; los individuos con NFC bajo dependen más de señales periféricas y atajos heurísticos. La tercera es la medida del Cognitive Reflection Test (CRT) propuesta por Frederick (2005): la tendencia a inhibir la respuesta automática intuitiva para sustituirla por reflexión deliberada. Las puntuaciones CRT predicen resistencia a sesgos cognitivos con robustez replicada en múltiples contextos culturales. La cuarta es el Actively Open-minded Thinking (AOT) (Baron, 1993; Stanovich y West, 2007): la disposición a considerar hipótesis alternativas, buscar información que contradiga las creencias actuales y revisar posiciones ante evidencia.

Cuatro propiedades de estas dimensiones son críticas para el argumento. Primera: muestran variación individual sustancial documentada en múltiples estudios. Segunda: la variación dentro de cualquier grupo cultural, étnico o socioeconómico definible supera la variación entre grupos (Nisbett et al., 2012), lo que hace al individuo —no al grupo— la unidad de análisis relevante para el argumento institucional. Tercera: son influidas por condiciones educativas y prácticas, lo que las hace parcialmente entrenables y, por tanto, objeto legítimo de política pública. Cuarta: están

conectadas mecánicamente, según la literatura disponible sobre procesamiento dual, con la calidad del razonamiento sobre cualquier instrumento de elaboración cognitiva, lo que sugiere —aunque sin verificación empírica directa específica para el uso de modelos de lenguaje— que son las dimensiones que determinan la calidad del uso del instrumento de Plano 2 como instrumento de genuina potenciación del pensamiento.

La hipótesis derivada de estas cuatro propiedades es que la efectividad del Mecanismo 1 sobre la elaboración del pensamiento del usuario varía sistemáticamente con la posición del usuario en estas dimensiones. Un usuario con alto NFC, alto CRT, alta Apertura y alto AOT tiende a usar el modelo evaluando críticamente sus respuestas, formulando preguntas que desafían las premisas que el modelo asume por defecto, manteniendo disposición a ser desafiado por perspectivas no anticipadas, y reconociendo cuando la fluidez del modelo refleja la representación del corpus en una zona y no la calidad genuina de un argumento. Un usuario con bajo NFC, bajo CRT, baja Apertura y bajo AOT tiende a aceptar las respuestas del modelo por su fluidez, a no formular preguntas que activen la ruta central de procesamiento, a quedar dentro de los marcos conceptuales que el corpus prefiere por defecto, y a confundir la sofisticación lingüística del output con la calidad sustantiva del razonamiento.

La consecuencia es directa: el instrumento de Plano 2 produce, sobre la misma población, efectos cualitativamente distintos en función de la posición del usuario en las dimensiones psicológicas examinadas. Para el primer perfil, el instrumento puede operar como instrumento de genuina potenciación —un interlocutor que desafía y refina el pensamiento. Para el segundo perfil, el instrumento opera como amplificador con mayor sofisticación de los marcos previos, produciendo confirmación elaborada de lo que ya se creía bajo la apariencia subjetiva de exploración intelectual genuina. La distinción entre potenciación cognitiva genuina y dependencia funcional bajo apariencia de potenciación —Dimensión 5 de la analogía estructural— se traza, en el plano del uso individual, a lo largo de las dimensiones psicológicas que la literatura disponible documenta.

Esta caracterización exige inmediatamente una distinción analítica que no es opcional sino constitutiva del argumento, porque las dos lecturas posibles de la misma observación tienen consecuencias normativas radicalmente opuestas. La primera lectura —que examinaremos en detalle en la Sección 7 como objeción anticipada— sostiene que la variación interindividual justifica jerarquía: los que pueden usar mejor los instrumentos cognitivos deben gobernar a los que no pueden. Es la respuesta de Karp y, antes que él, la de toda variante seria del elitismo tecnológico. La segunda lectura sostiene que la variación interindividual obliga al diseño institucional a compensar la asimetría que el sistema produce sobre poblaciones cuya variación protectora se distribuye desigualmente. Es la respuesta que este paper articula. La distinción entre ambas no es de sensibilidad moral. Es analítica. La primera lectura usa la variación pero ignora que el sistema —como argumentará la Sección 5— degrada activamente las dimensiones protectoras en tiempo real,

transformando la observación pasiva de la variación en explotación activa de la asimetría que ella produce.

5. La articulación causal: la trampa de segundo orden

Esta sección desarrolla la pieza arquitectónica más importante del paper. La analogía estructural de la Sección 3 establece que los dos peligros comparten una forma común. La descripción de los mecanismos de la Sección 4 establece cómo opera específicamente el peligro de Plano 2. Lo que esta sección añade es la articulación causal entre los dos planos: la demostración de que el peligro de Plano 1 y el peligro de Plano 2 no son simplemente análogos sino que están causalmente acoplados de un modo preciso, y que ese acoplamiento exige tratar a los dos peligros como un único objeto regulatorio.

5.1 La conexión mecánica entre los dos planos

El Paper 3 documentó, con base en la literatura disponible sobre procesamiento dual, una propiedad fundamental del modo alarma como estado basal del sistema: el modo alarma reduce la tolerancia a la ambigüedad y favorece el procesamiento heurístico sobre el deliberativo. Esa propiedad puede reformularse, en los términos de la Sección 4 de este paper, como una proposición precisa: el modo alarma basal produce, en tiempo real y mientras opera, una reducción funcional de las dimensiones psicológicas que determinan la calidad del uso del instrumento de Plano 2. Un individuo bajo modo alarma sostenido tiene, en términos efectivos, NFC funcional reducido respecto de su NFC basal, CRT funcional reducido, Apertura funcional reducida, AOT funcional reducido. La reducción no es permanente —se revierte cuando el modo alarma cede—, pero opera mientras dura.

La consecuencia analítica es decisiva: el sistema de Plano 1, cuya operación normal mantiene a la población en modo alarma como estado basal, está degradando en tiempo real exactamente las dimensiones que protegerían al individuo de los efectos del Mecanismo 1 de Plano 2. La trampa que esto produce es de segundo orden, y conviene articularla con precisión:

- Nivel 1: el sistema de Plano 1 produce vulnerabilidad cognitiva diferencial sobre la población, con mayor efectividad sobre los perfiles con menores defensas elaborativas basales (Papers 1, 2 y 3).
- Nivel 2: el modo alarma basal del sistema de Plano 1 degrada en tiempo real las dimensiones que permitirían al individuo resistir esos mecanismos —reduce funcionalmente NFC, CRT, Apertura y AOT durante la exposición.
- Nivel 3: las mismas dimensiones que el sistema de Plano 1 degrada en tiempo real son las dimensiones que determinan la calidad del uso del instrumento de Plano 2 como instrumento de potenciación genuina del pensamiento.

El individuo expuesto a la operación normal de los sistemas de Plano 1 acumula, en consecuencia, dos efectos simultáneos que se refuerzan mutuamente: una vulnerabilidad creciente a los mecanismos de Plano 1, y una capacidad decreciente de usar el instrumento de Plano 2 de forma que genuinamente potencie el pensamiento en lugar de confirmar con mayor sofisticación los marcos previos. La trampa de segundo orden es la denominación precisa de esta articulación: el sistema produce la vulnerabilidad y simultáneamente degrada las capacidades que permitirían escapar de ella mediante la herramienta que aparentemente la compensaría.

5.2 Por qué la articulación importa para la regulación

La articulación causal entre los dos planos tiene consecuencias regulatorias específicas que ningún marco vigente captura. La regulación de Plano 1 que el Paper 4 propone —modificación de la función objetivo, condicionamiento legítimo del privilegio sostenido, aplicación de responsabilidad objetiva sobre la actividad estructuralmente riesgosa— sigue siendo necesaria, pero no es suficiente para abordar el peligro completo. Por dos razones que conviene articular separadamente.

Primera razón: aun suponiendo regulación efectiva del Plano 1, el Plano 2 opera independientemente. Aun si la función objetivo de los sistemas de recomendación se modificara mañana en el sentido que el Paper 4 prescribe, los modelos de lenguaje seguirían codificando los presupuestos no neutrales del Mecanismo 1, los actores que los controlan seguirían enfrentando el desplazamiento que el Mecanismo 2 describe, y la efectividad diferencial del Mecanismo 3 seguiría operando sobre la variación interindividual de la población. La regulación de Plano 1 no produce, por sí misma, la regulación de Plano 2.

Segunda razón: la regulación de Plano 2 que esta serie propone presupone capacidades en la población que el sistema de Plano 1 está degradando activamente. Cualquier estrategia regulatoria de Plano 2 que opere por la vía de la educación, de la alfabetización metacognitiva, de la formación crítica para el uso de instrumentos cognitivos, presupone que la población a la que esa estrategia se dirige tiene disponibles, en condiciones efectivas, las dimensiones psicológicas que la Sección 4 documentó. La trampa de segundo orden establece que esa presunción no se cumple si simultáneamente la operación de Plano 1 está degradando esas dimensiones. La consecuencia regulatoria es que los dos planos deben tratarse en conjunto: la regulación de Plano 2 sin regulación de Plano 1 es estructuralmente ineficaz, y la regulación de Plano 1 sin regulación de Plano 2 es insuficiente para abordar el peligro completo.

Esta articulación es la justificación precisa de la unificación categorial que la Sección 6 articula: por qué la categoría de soberanía cognitiva, no las categorías separadas de regulación de plataformas y regulación de IA, es la categoría regulatoria correcta. La unificación no es una preferencia conceptual: es una consecuencia analítica de la trampa de segundo orden.

6. La soberanía cognitiva como categoría unificadora y la infraestructura cognitiva pública

6.1 La definición operacional de soberanía cognitiva

Las secciones anteriores establecen las premisas. Esta sección articula la consecuencia normativa central del paper: la categoría de soberanía cognitiva, articulada como condición de posibilidad de la formación autónoma de juicios, integra los dos planos del peligro en un único objeto regulatorio. La definición operacional propuesta es la siguiente:

> *Soberanía cognitiva es la capacidad efectiva de individuos y comunidades de formar juicios autónomos sobre la realidad social y de elaborar pensamiento propio, usando instrumentos de mediación —de percepción social y de instrumentación cognitiva— cuyos presupuestos sean conocidos y deliberativamente aceptados, con acceso a los recursos de potenciación que hacen posible la elaboración intelectual de alta densidad, y sin exposición no compensada a mecanismos de influencia que operen sobre sus vulnerabilidades cognitivas sin su conocimiento.*

La definición articula tres componentes que corresponden a tres niveles de intervención analíticamente distintos pero institucionalmente complementarios. El nivel individual —la capacidad metacognitiva entrenable que determina la calidad del uso del instrumento de potenciación y la resistencia a los mecanismos de mediación— es el nivel sobre el que opera la efectividad diferencial del Mecanismo 3 y que la trampa de segundo orden hace urgente compensar activamente mediante política pública. El nivel colectivo —el corpus de entrenamiento y la función objetivo como portadores de presupuestos que no son neutrales— es el nivel que la disputa geopolítica está articulando de forma incipiente y que el forzamiento filosófico hace visible. El nivel institucional —la gobernanza de la función objetivo y de los presupuestos del corpus como condición de posibilidad de la soberanía cognitiva colectiva— es la extensión directa del armazón del Paper 4 al plano de la instrumentación.

6.2 La extensión del armazón jurídico del Paper 4

La unificación categorial permite extender el armazón jurídico del Paper 4 al plano de la instrumentación con precisión analítica. Cada pieza del armazón del Paper 4 tiene su contraparte en el plano nuevo, y conviene articularla.

El bien irreductiblemente común, que el Paper 4 identificó como la integridad informacional y deliberativa de la comunidad política, se extiende en el plano de la instrumentación cognitiva como la integridad de los instrumentos sobre los que la comunidad política elabora pensamiento. Si el corpus, la función de pérdida y los presupuestos de seguridad de los modelos de lenguaje masivos codifican los marcos sobre los que el pensamiento se elabora, la integridad de esos instrumentos —su transparencia respecto de los presupuestos codificados, su pluralidad respecto de las tradiciones

intelectuales que representan, su accesibilidad universal respecto de la población que los usa— es un bien irreductiblemente común en el sentido preciso del Paper 4. No es un bien individual agregado, porque el efecto sobre cualquier individuo depende de la matriz instrumental que toda la comunidad usa; es un bien estructural cuya integridad cada miembro de la comunidad necesita pero ninguno posee separadamente.

El vicio estructural del consentimiento, que el Paper 4 identificó como mecanismo por el que el daño de Plano 1 opera, se extiende sin modificación sustancial al Plano 2. El usuario que conversa con un modelo de lenguaje no consiente en sentido jurídicamente significativo a la operación del Mecanismo 1 sobre la elaboración de su pensamiento, porque ese mecanismo opera por debajo del umbral de detección consciente, sin que el usuario tenga acceso al corpus, a la función de pérdida o a los procedimientos de fine-tuning que producen los efectos. El consentimiento al «uso del instrumento» no es consentimiento al condicionamiento estructural que ese uso produce. La protección frente al vicio estructural del consentimiento, que el Paper 4 articuló como función ordinaria del Estado y no excepcional, se extiende al plano de la instrumentación con la misma fuerza jurídica.

La tripartición modal de imputación se extiende también al Plano 2 con ajustes que la Sección 8 desarrolla. Hay un Modo 1 de operación deliberada —actores que utilizan modelos de lenguaje con conocimiento de sus presupuestos para producir efectos políticos específicos sobre poblaciones específicas. Hay un Modo 2 de captura oportunista —actores políticos o comerciales que se benefician de los efectos del Mecanismo 1 sin haberlos diseñado, pero con representación del daño previsible. Y hay un Modo 3 de emergencia sistémica —los efectos sobre la elaboración del pensamiento que emergen de la operación normal de los modelos sobre poblaciones masivas, sin que ningún actor individual los haya diseñado. El Modo 3 vuelve a ser, como en el Plano 1, el modo causalmente dominante y el que ningún marco regulatorio vigente puede abordar; y vuelve a exigir, como en el Plano 1, la aplicación del régimen de responsabilidad objetiva por actividad estructuralmente riesgosa. La instrumentación cognitiva masiva sobre poblaciones con vulnerabilidades cognitivas documentadas, bajo función de pérdida específica y corpus históricamente situado, configura una actividad cuyos daños emergentes son inherentes a su operación normal y exigen, por la lógica clásica del derecho de daños, atribución objetiva de responsabilidad y modificación estructural de la actividad para minimizar el daño inherente.

El condicionamiento legítimo del privilegio sostenido, que es la conclusión normativa central del Paper 4, se extiende al Plano 2 sin necesidad de fundación adicional. Los laboratorios que operan los modelos de lenguaje masivos lo hacen sobre infraestructura constitutivamente híbrida —Dimensión 1 de la analogía estructural— y se benefician de los mismos regímenes de subsidio fiscal, formación pública del talento y financiamiento público histórico de la investigación que el Paper 4 documentó. Operan bajo régimen de privilegio público sostenido, no bajo propiedad privada absoluta. El Estado que sostiene continuamente ese régimen tiene la facultad y el deber ordinarios de condicionar la

operación a las condiciones que la gestión legítima del régimen exige, incluyendo la transparencia respecto de los presupuestos del corpus, la pluralidad respecto de las tradiciones intelectuales representadas y los regímenes de auditoría externa de las restricciones de seguridad y los procedimientos de fine-tuning.

6.3 La infraestructura cognitiva pública como consecuencia institucional obligada

El armazón jurídico extendido habilita la consecuencia institucional central que el paper deriva: la infraestructura cognitiva pública. Esta categoría no se propone como manifiesto político ni como añadido normativo a un argumento que se sostiene sin ella. Se deriva como consecuencia obligada del reconocimiento de la analogía estructural, de la articulación causal entre los dos planos y de la extensión del armazón del Paper 4 al plano de la instrumentación.

La infraestructura cognitiva pública se define operacionalmente como el conjunto de condiciones institucionales que garantizan, para la población general, acceso no mediado por la función objetivo de actores privados a instrumentos de potenciación cognitiva de calidad, formación efectiva en las capacidades metacognitivas que determinan la calidad del uso de esos instrumentos, y participación deliberativa en la determinación de los presupuestos sobre los que esos instrumentos operan. La categoría es paralela, en su lógica institucional, a la educación pública, la salud pública y la justicia pública: condiciones de posibilidad de la dignidad y la participación democrática que el diseño institucional tiene la obligación de garantizar y no bienes de mercado cuya distribución se deja exclusivamente a la función objetivo de actores privados.

La justificación de la infraestructura cognitiva pública sigue, paso a paso, la misma estructura de justificación que el Paper 4 desarrolló para el condicionamiento de la función objetivo de los sistemas de recomendación. La hibridez constitutiva de la infraestructura algorítmica-inferencial, en su totalidad y no solo en el subsistema de distribución, configura un régimen de privilegio público sostenido que el Estado renueva continuamente. La operación de ese régimen produce, sobre la integridad de los instrumentos cognitivos masivos, efectos estructurales que la gestión legítima del régimen tiene la facultad y el deber de abordar. La omisión regulatoria, frente a daño documentado y privilegio sostenido, configura mala administración del régimen de privilegio y omisión culposa frente al daño que ese privilegio facilita. La inversión de la carga de la prueba que el Paper 4 produjo para el Plano 1 se reproduce en el Plano 2: la pregunta deja de ser por qué el Estado puede intervenir y se convierte en por qué la sociedad que financió la infraestructura, formó el talento que la opera y subsidia operativamente al sector no puede establecer condiciones sobre los presupuestos codificados en los instrumentos cognitivos masivos cuando esos presupuestos producen efectos estructurales sobre la elaboración del pensamiento de poblaciones enteras.

La infraestructura cognitiva pública admite múltiples implementaciones institucionales que exceden al alcance de este paper desarrollar en detalle. Tres dimensiones operativas, sin embargo, pueden enunciarse para clarificar el contenido sustantivo de la categoría. Primera dimensión: la transparencia regulatoriamente exigible respecto del corpus de entrenamiento, la función de pérdida y los procedimientos de fine-tuning de los modelos de lenguaje que operan a escala poblacional, con auditoría por reguladores independientes y obligación de provisión de datos para investigación científica externa. Segunda dimensión: el desarrollo, financiado total o parcialmente con fondos públicos, de modelos de lenguaje entrenados sobre corpus que reflejan tradiciones intelectuales sub-representadas en los corpus dominantes, como condición de pluralidad estructural de la matriz instrumental disponible para la población. Tercera dimensión: la incorporación de la alfabetización metacognitiva —el desarrollo deliberado de las dimensiones psicológicas (NFC, CRT, AOT, Apertura) que determinan la calidad del uso de los instrumentos cognitivos— como objeto de política educativa pública, en analogía con la alfabetización lectora y la educación científica.

Estas tres dimensiones no agotan el contenido de la categoría ni constituyen un programa cerrado. Son ejemplos de implementación que clarifican qué tipo de instituciones serían consecuencias plausibles de la consagración regulatoria de la infraestructura cognitiva pública como categoría. El diseño institucional concreto excede el alcance del argumento académico y requiere deliberación política, análisis técnico-jurídico y diseño institucional específico en cada jurisdicción.

7. Objeciones y respuestas

7.1 La analogía es forzada: redes sociales y modelos de lenguaje son tecnologías estructuralmente distintas

La objeción más previsible al argumento sostiene que la analogía entre los dos planos no es estructural sino retórica, porque las dos tecnologías son sustantivamente distintas: las redes sociales operan sobre la distribución de contenido producido por humanos, los modelos de lenguaje generan contenido nuevo; las redes sociales operan sobre el grafo social del usuario, los modelos operan sobre la conversación individual; las redes sociales tienen un modelo de negocio basado en la atención mantenida, los modelos en la suscripción o en el uso por API. Tratarlas como instancias de un mismo peligro estructural sería, según esta objeción, oscurecer diferencias que importan.

La respuesta es que la objeción descansa en una confusión entre identidad sustantiva y analogía estructural. El paper no afirma que los dos peligros sean idénticos —la Sección 3.4 explicitó precisamente lo que la analogía no afirma. Afirma que comparten una estructura formal idéntica en cinco dimensiones específicas: hibridez constitutiva de la infraestructura, codificación de presupuestos no neutrales en la arquitectura, operación bajo umbral de detección consciente, efectividad diferencial sobre vulnerabilidades cognitivas variables, y producción de dependencia bajo apariencia de potenciación. Esa coincidencia estructural es verificable en cada dimensión por

separado y se sostiene incluso reconociendo todas las diferencias sustantivas que la objeción enumera. El paper no propone que las dos tecnologías sean lo mismo. Propone que los dos peligros pertenecen a la misma categoría de peligro y que esa categoría es la categoría regulatoria correcta. La consecuencia institucional —tratar a los dos planos como un único objeto regulatorio— se sigue de la coincidencia estructural, no de una identidad sustantiva que el paper no reclama.

7.2 El argumento es paternalista: presupone usuarios pasivos sin agencia

La segunda objeción sostiene que el argumento, al insistir en la efectividad diferencial sobre vulnerabilidades cognitivas y en la trampa de segundo orden, presupone receptores pasivos cuya agencia queda eliminada por los mecanismos sistémicos descritos. Eso sería paternalista respecto de la capacidad real de los individuos de resistir, desconectarse, desarrollar literacidad mediática y adoptar distancia crítica frente a los instrumentos.

La objeción es correcta como descripción de una posibilidad individual, pero incorrecta como crítica del armazón. El paper no afirma que todos los individuos sean igualmente vulnerables ni que la agencia individual sea imposible. Afirma, siguiendo los Papers 1 a 3, que la resistencia activa carece de simetría con el sistema: el sistema opera continuamente, automáticamente y con conocimiento del perfil de cada usuario; la resistencia individual requiere esfuerzo cognitivo deliberado, recursos que la mayoría no tiene y conocimiento del mecanismo que la propia operación del sistema dificulta producir. La pregunta analítica no es si algunos individuos pueden resistir; es por qué la arquitectura del sistema hace costosa la resistencia para la mayoría y por qué la trampa de segundo orden hace que esa costo aumente para los perfiles que más necesitan resistir. Una protección puramente individualista, que dejara la responsabilidad de la resistencia a la capacidad metacognitiva del usuario individual, produce protecciones uniformes sobre riesgos diferenciales y deja desprotegidos exactamente a quienes más protección necesitan. La regulación estructural —la infraestructura cognitiva pública, la alfabetización metacognitiva como política pública, la transparencia regulatoriamente exigible respecto de los presupuestos del corpus— opera sobre la condición que produce la distribución desigual del riesgo, no sobre la capacidad individual de absorberlo.

7.3 Reconocer la variación interindividual implica jerarquía

La tercera objeción es la más importante de responder con precisión, porque su confusión es la que conduce, en otras versiones del argumento, a las posiciones que este paper rechaza explícitamente. La objeción sostiene que reconocer la variación interindividual en NFC, CRT, AOT y Apertura es estructuralmente equivalente a reproducir la lógica de Karp: si las capacidades difieren, entonces los que tienen mejores capacidades deben gobernar.

La respuesta es que la objeción confunde dos argumentos que son estructuralmente opuestos. El argumento de Karp dice: la variación existe, por lo tanto los que están en una posición privilegiada deben gobernar. El argumento de este paper dice: la variación existe, por lo tanto el diseño

institucional tiene la obligación de compensar la asimetría que los sistemas que operan sobre esa variación producen, y especialmente cuando esos sistemas degradan activamente las dimensiones que producen la variación protectora. La dirección normativa es opuesta. El primer argumento usa la variación para justificar la concentración del poder. El segundo la usa para justificar la distribución del apoyo.

Reconocer variación interindividual en dimensiones psicológicas no implica ninguna conclusión normativa específica: implica que las conclusiones normativas deben construirse teniendo en cuenta esa variación, no ignorándola. Ignorarla no la hace desaparecer: la hace inaccesible para el diseño de intervenciones que podrían compensarla. Y en el caso específico que el paper articula, esa compensación es urgente, no opcional, porque el sistema no solo opera sobre la variación pasiva: la amplifica activamente degradando en tiempo real las dimensiones protectoras mediante el modo alarma. La trampa de segundo orden convierte el reconocimiento de la variación de tema controvertido en condición sine qua non del diseño institucional adecuado al peligro que el paper describe.

7.4 El open source resuelve el problema del acceso

La cuarta objeción sostiene que el ecosistema de modelos de lenguaje en código abierto —Llama, Mistral, Qwen, DeepSeek, entre otros— democratiza efectivamente el acceso a la instrumentación cognitiva y resuelve, sin necesidad de regulación adicional, el problema que el paper describe. Si cualquiera puede acceder a un modelo y entrenar su propio modelo derivado, los presupuestos codificados en los modelos comerciales pierden su carácter monopólico.

El open source reduce efectivamente la barrera de acceso técnico a los modelos. Pero la objeción confunde tres niveles de acceso que el argumento debe distinguir. El acceso al modelo —poder ejecutarlo localmente o vía API— se democratiza significativamente con el open source. La capacidad de entrenar modelos competitivos sobre corpus que reflejen presupuestos filosóficos alternativos sigue requiriendo recursos que el open source no provee: capacidad técnica especializada, infraestructura computacional sustancial, datos de calidad en volumen comparable y curación de corpus que es, en sí misma, una operación políticamente cargada. Y, crucialmente, la metacognición necesaria para usar cualquier modelo —comercial u open source— de forma que genuinamente potencie el pensamiento en lugar de confirmar sesgos previos no es una propiedad técnica del modelo: es una capacidad del usuario que el open source no desarrolla y que la trampa de segundo orden hace problemática asumir como dada.

El open source es, en consecuencia, una pieza de la infraestructura cognitiva pública en sentido amplio, pero no es el contenido completo de la categoría. La sustitución del problema por la disponibilidad de alternativas técnicas presupone una capacidad de uso crítico que la infraestructura cognitiva pública debe construir como condición de posibilidad.

7.5 La propuesta de gobernanza es vaga

La quinta objeción sostiene que la categoría de infraestructura cognitiva pública, tal como el paper la introduce, es vaga: no especifica los instrumentos concretos, las arquitecturas institucionales, las jurisdicciones competentes, los regímenes de financiamiento. Sin esa especificación, la categoría sería más una formulación retórica que una propuesta institucional sustantiva.

La objeción es parcialmente correcta y conviene aceptarla en lo que tiene de correcto antes de delimitar lo que tiene de incorrecto. Es correcto que el paper no provee diseño institucional concreto: no especifica si la regulación corresponde a la Unión Europea, a Estados nacionales, a organismos multilaterales; no especifica los mecanismos de auditoría del corpus; no especifica los regímenes de financiamiento de los modelos públicos; no especifica los currículos de la alfabetización metacognitiva. Es incorrecto, sin embargo, que esa ausencia de especificación constituya vaguedad del armazón conceptual. La función específica del paper en el programa de investigación de la serie es identificar la categoría jurídica desde la cual ese diseño puede partir, los modos causales sobre los que cada conjunto de instrumentos opera y el régimen de imputación apropiado a cada modo —la misma división del trabajo que el Paper 4 estableció explícitamente.

El diseño institucional concreto requiere deliberación política, análisis jurídico específico de cada jurisdicción y diseño institucional que excede el alcance de un argumento académico. La contribución específica del paper es, como en el Paper 4, identificar la categoría desde la cual ese diseño puede partir y mostrar que esa categoría es accesible a la regulación con instrumentos que el derecho ya conoce y aplica a otras infraestructuras de impacto comparable. Cambiar el presupuesto categorial no es suficiente para producir el diseño regulatorio. Es necesario para que el diseño regulatorio sea posible.

7.6 El argumento es apocalíptico o conspirativo

La sexta objeción sostiene que el conjunto del armazón —dos peligros estructurales, trampa de segundo orden, instrumentación cognitiva como objeto de disputa filosófica— produce un cuadro excesivamente sombrío que subestima la resiliencia de las instituciones democráticas, la agencia individual y la posibilidad de que el desarrollo tecnológico produzca soluciones a sus propios problemas.

El argumento del paper no afirma que el resultado esté determinado ni que la agencia individual sea imposible. Afirma que el diseño del sistema —la función objetivo de los sistemas de recomendación, los presupuestos codificados en el corpus de entrenamiento, la distribución desigual del acceso a la potenciación cognitiva— crea asimetrías estructurales que la resiliencia individual y la agencia democrática deben enfrentar con los recursos que el sistema produce en su operación normal. Identificar esas asimetrías con precisión es la condición de posibilidad de cualquier respuesta eficaz. Ignorarlas en nombre de la agencia individual o de la confianza en la autocorrección tecnológica

produce exactamente el efecto que el Paper 3 describe como anestesia epistémica: la satisfacción subjetiva de que se puede resistir, que coexiste con la reducción objetiva de la capacidad de hacerlo.

El argumento no es conspirativo en el sentido preciso que el Paper 4 articuló al responder a la misma objeción: no postula coordinación secreta entre actores ni intención política explícita de producir el peligro. Postula la convergencia estructural de incentivos económicos, dinámica de desarrollo tecnológico exponencial y omisión regulatoria, en presencia de una infraestructura constitutivamente híbrida cuyo régimen de privilegio público sostenido es lo que el Estado tiene la facultad y el deber de gobernar. La no-conspiración del Paper 4 sigue siendo la no-conspiración de este paper. La emergencia sistémica del Modo 3 sigue siendo el modo causalmente dominante.

8. Implicaciones

8.1 Para la teoría regulatoria: la unificación del objeto

La consecuencia más directa del armazón propuesto para la teoría regulatoria es la unificación del objeto. Los marcos vigentes operan sobre tres objetos analíticamente separados —la regulación de plataformas digitales, la regulación de inteligencia artificial, las políticas de educación y formación cognitiva—. La articulación que el paper produce muestra que esos tres objetos son, en sentido estricto, dimensiones de un único objeto regulatorio: la integridad de las condiciones de posibilidad de la formación autónoma de juicios y de la elaboración de pensamiento propio en una comunidad política. La regulación de plataformas, la regulación de IA y las políticas de alfabetización cognitiva no son tres áreas regulatorias separables: son las tres dimensiones de la regulación de la soberanía cognitiva. Tratarlas separadamente, como hacen los marcos vigentes, no es una división del trabajo legítima sino una consecuencia del error de categorización que el paper documenta.

8.2 Para el análisis político: tres bases de legitimidad transnacional

La cuestión jurisdiccional, que el Paper 4 abordó respecto del primer plano del peligro, se reproduce respecto del segundo: si los principales laboratorios de modelos de lenguaje son estadounidenses o chinos, qué legitimidad tiene el Estado argentino, el europeo o el brasileño para condicionar los presupuestos del corpus o de las restricciones de seguridad. La respuesta sigue la misma estructura tripartita que el Paper 4 articuló.

Primera base: legitimidad por participación en la infraestructura de conocimiento. Los corpus de entrenamiento de los modelos incluyen literatura académica, datos producidos en investigación financiada con fondos públicos en múltiples jurisdicciones, y resultados de investigación cuyo costo de producción fue asumido por sistemas educativos públicos cuya formación de talento aprovechan los laboratorios. Cualquier Estado cuyos sistemas públicos contribuyeron al corpus o al talento humano tiene base de legitimidad para condicionar la operación de los instrumentos producidos.

Segunda base: legitimidad por subsidio operativo presente. Los laboratorios de IA operan, en múltiples jurisdicciones, bajo regímenes de subsidio fiscal —la Ley de Economía del Conocimiento argentina, el programa Horizon Europe, los Government-Guided Investment Funds chinos, los créditos fiscales estadounidenses para investigación y desarrollo. Cualquier Estado que está actualmente subsidiando la operación tiene base de legitimidad como administrador del régimen de privilegio que renueva continuamente.

Tercera base: legitimidad por soberanía sobre la integridad cognitiva de la propia comunidad política. Cualquier Estado cuya población usa masivamente los instrumentos de instrumentación cognitiva tiene base de legitimidad en el deber de proteger un bien irreductiblemente común propio de la comunidad política que ese Estado representa. Esta es la base más amplia y la que no depende de ningún vínculo histórico o fiscal con la infraestructura.

La combinación de las tres bases produce un argumento de legitimidad regulatoria transnacional que no requiere que ningún Estado herede la legitimidad de los Estados donde los principales laboratorios están establecidos: cada Estado la genera desde su propia relación presente con la infraestructura cognitiva, su régimen propio de privilegios concedidos y los efectos sobre su propia comunidad política.

8.3 Para la geopolítica de la IA: la disputa por los corpus como campo regulatorio

El argumento del paper sugiere una predicción específica sobre la evolución regulatoria internacional. La disputa geopolítica sobre la inteligencia artificial ha sido descrita principalmente en términos de competencia tecnológica y militar: quién tiene los mejores modelos, los chips más avanzados, el talento más calificado. La unificación categorial que el paper propone sugiere que esa descripción es incompleta y que la dimensión sustantiva de la disputa, en el mediano plazo, será la disputa por los corpus de entrenamiento como objeto regulatorio: qué presupuestos filosóficos se codifican en los instrumentos cognitivos masivos que la población de cada jurisdicción usa, con qué transparencia, con qué pluralidad, bajo qué régimen de gobernanza pública.

La predicción específica que el armazón sugiere es que en los próximos cinco años el control sobre los conjuntos de datos de entrenamiento de modelos de lenguaje masivos se convertirá en objeto de regulación pública explícita en al menos tres jurisdicciones con capacidad de implementación, siguiendo la lógica de la hibridez constitutiva establecida en el Paper 4. La forma específica de esa regulación —exigencias de transparencia, requisitos de pluralidad de fuentes, financiamiento público de modelos alternativos, certificación de presupuestos— variará por jurisdicción, pero la categoría regulatoria emergerá como categoría reconocida.

8.4 Para el Paper 6: la condición de posibilidad de la soberanía cognitiva

Esta subsección articula la conexión entre el argumento de este paper y el del Paper 6, que es el paper sucesor en la serie y al cual este paper sirve, en su función arquitectónica, como condición conceptual previa. El Paper 6 describe el escenario en el que el presupuesto antropológico que toda la serie había mantenido implícitamente se invalida: la condición de aplicabilidad de los mecanismos descritos en los Papers 0 a 4 era que los nodos del grafo social que generan señal son humanos, y esa condición está siendo erosionada progresivamente por la irrupción de agentes de inteligencia artificial generativa que producen contenido y comportamiento estadísticamente plausibles para receptores humanos.

La articulación con este paper es la siguiente: la categoría de soberanía cognitiva, tal como este paper la define, exige una condición de posibilidad que el Paper 6 examina específicamente: la distinguibilidad. La soberanía cognitiva en el plano de la percepción social exige que el individuo pueda distinguir, al menos en principio, las señales auténticas de su entorno social de las construcciones que el sistema fabrica. La soberanía cognitiva en el plano de la instrumentación exige que el individuo pueda distinguir, al menos en principio, los presupuestos codificados en el instrumento que usa. Las dos exigencias son formalmente análogas: ambas requieren transparencia respecto del origen y de la naturaleza de aquello que media la cognición individual.

El Paper 6 muestra el escenario en el que la primera exigencia colapsa estructuralmente: cuando una proporción mayoritaria y no determinable de los nodos que generan señal en el grafo social son agentes sintéticos, la distinguibilidad entre señal humana y señal sintética se vuelve estructuralmente imposible. La trampa que el Paper 5 identifica como trampa de segundo orden adquiere un tercer nivel en el Paper 6: el sistema produce vulnerabilidad, degrada las capacidades que permitirían detectarla, y elimina la posibilidad de distinguir el entorno intervenido del entorno real incluso si esas capacidades estuvieran intactas. El Paper 5 instala el concepto de soberanía cognitiva como categoría regulatoria unificada y la condición de distinguibilidad como su condición de posibilidad. El Paper 6 muestra el escenario límite en que esa condición de posibilidad se erosiona estructuralmente y deriva las consecuencias para los marcos teóricos y regulatorios que el armazón de los papers anteriores había construido.

La función arquitectónica de este paper en la serie es, en consecuencia, doble. Hacia atrás, extiende y unifica el armazón del Paper 4 al plano de la instrumentación cognitiva, articulando la categoría de soberanía cognitiva como objeto regulatorio único que integra los dos planos del peligro. Hacia adelante, instala las condiciones de posibilidad de esa categoría —la distinguibilidad como condición— que el Paper 6 examina en su escenario de erosión estructural. El paper opera como bisagra entre la analítica jurídica del Paper 4 y la analítica de autonomía técnica del Paper 6, sin que ninguna de las dos funciones lo agote: la unificación categorial es contribución sustantiva propia que se sostiene con independencia de su función de bisagra.

9. Conclusiones y predicciones empíricas

9.1 Recapitulación del argumento

He argumentado en este paper que el desarrollo exponencial de la inteligencia artificial —y específicamente la irrupción de los modelos de lenguaje de gran escala como instrumentos cognitivos disponibles a escala masiva para la elaboración del pensamiento individual— produce un peligro estructural a la soberanía cognitiva análogo al que los Papers 0 a 4 de esta serie documentaron en el plano de la percepción social. La analogía no es retórica sino estructural en sentido formal preciso: los dos peligros comparten una estructura común en cinco dimensiones específicas —hibridez constitutiva de la infraestructura, codificación de presupuestos no neutrales en la arquitectura, operación bajo umbral de detección consciente, efectividad diferencial sobre vulnerabilidades cognitivas variables, y producción de dependencia bajo apariencia de potenciación. Esa estructura común permite y exige tratar a los dos peligros como un único objeto regulatorio dentro de la categoría de soberanía cognitiva.

Los dos planos del peligro están además causalmente acoplados: el sistema de Plano 1, cuya operación normal mantiene a la población en modo alarma como estado basal, degrada en tiempo real exactamente las dimensiones psicológicas que protegerían al individuo de los efectos del Mecanismo 1 de Plano 2 y que determinan la calidad del uso del instrumento de Plano 2 como instrumento de potenciación genuina. La trampa de segundo orden no es un argumento entre otros: es la pieza arquitectónica que explica por qué los dos peligros son inseparables y por qué los remedios deben pensarse en conjunto.

El armazón jurídico que el Paper 4 construyó para el primer plano del peligro —bien irreductiblemente común, vicio estructural del consentimiento, tripartición modal de imputación, condicionamiento legítimo del privilegio sostenido— se extiende sin modificaciones sustanciales al segundo plano. La consecuencia institucional obligada de esa extensión es la categoría de infraestructura cognitiva pública: el conjunto de condiciones institucionales que garantizan acceso no mediado por la función objetivo de actores privados a instrumentos de potenciación cognitiva de calidad, formación efectiva en las capacidades metacognitivas que determinan la calidad del uso de esos instrumentos, y participación deliberativa en la determinación de los presupuestos sobre los que esos instrumentos operan.

El forzamiento filosófico —el momento en que actores que controlan instrumentación cognitiva son obligados a declarar públicamente los presupuestos sobre los que operan— es la cara visible del desplazamiento que el paper articula. Los casos de Palantir Technologies y Anthropic, examinados en sus diferencias estructurales y no en una simetría retórica, son los dos primeros casos documentados de un fenómeno que el armazón del paper explica como consecuencia estructural de la masividad y la creciente auditabilidad de los modelos. La economía del silencio que durante

décadas protegió la opacidad sobre la hibridez constitutiva está colapsando, no como producto de una decisión política sino como consecuencia estructural del desarrollo exponencial.

9.2 Lo que el paper deja explícitamente fuera

El paper reconoce, en línea con la disciplina epistémica de la serie, lo que su armazón explícitamente no provee. Primero, no provee diseño institucional concreto de la infraestructura cognitiva pública en jurisdicciones específicas: esa tarea requiere deliberación política, análisis técnico-jurídico y diseño institucional específico que excede el alcance de un argumento académico. Segundo, no provee verificación empírica directa de la efectividad diferencial del Mecanismo 1 sobre los perfiles psicológicos identificados: esa verificación requiere investigación experimental que el paper formula como agenda y como predicción falsificable, no como observación verificada. Tercero, no resuelve la cuestión sobre qué categorías filosóficas alternativas deberían codificar los corpus de los modelos de lenguaje públicos que el armazón institucional propone como pluralidad estructural: esa cuestión es propiamente política y requiere deliberación sustantiva sobre las tradiciones intelectuales que la pluralidad debe representar. El paper instala la categoría regulatoria. La política sustantiva sobre el contenido de la categoría es objeto de la deliberación pública que la categoría hace posible, no de su justificación previa.

9.3 Predicciones empíricas operacionalizables

Las siguientes predicciones son derivaciones del argumento que podrían orientar investigación empírica futura. Se formulan como hipótesis falsificables en sentido popperiano, reconociendo que su verificación directa requeriría acceso a datos y metodologías actualmente fuera del alcance de la investigación independiente.

Predicción 1 — Aceleración de declaraciones filosóficas corporativas. La frecuencia de declaraciones públicas explícitas de presupuestos filosóficos y políticos por parte de corporaciones tecnológicas de primer nivel aumentará de forma medible en los próximos veinticuatro meses, en correlación con el aumento de la capacidad técnica de auditoría algorítmica independiente. Si la economía del silencio opera como el armazón del paper propone, la reducción del costo de exposición producirá más declaraciones. Si no hay correlación entre el aumento de la auditabilidad y el aumento de las declaraciones, la hipótesis sobre el forzamiento filosófico requiere revisión.

Predicción 2 — Diferencial de efectividad según metacognición. En estudios experimentales controlados sobre uso de modelos de lenguaje para tareas de elaboración intelectual, el efecto de potenciación del pensamiento atribuible al uso de los modelos mostrará mayor magnitud en usuarios con mayor puntuación en medidas combinadas de NFC, CRT, AOT y Apertura, controlando por tiempo de uso, nivel educativo y familiaridad técnica con la herramienta. Si no hay diferencial significativo, la hipótesis sobre la efectividad diferencial del Mecanismo 1 requiere revisión.

Predicción 3 — Sesgo de amplificación por corpus. Análisis comparativos de modelos entrenados sobre corpus de distinta composición cultural y lingüística mostrarán diferencias sistemáticas en la facilidad con que producen ciertos tipos de razonamiento, medible mediante benchmarks diseñados específicamente para detectar sesgos de amplificación conceptual. La predicción es directamente falsificable mediante diseño experimental adecuado.

Predicción 4 — Convergencia geopolítica en gobernanza de corpus. En los próximos cinco años, el control sobre los conjuntos de datos de entrenamiento de modelos de lenguaje se convertirá en objeto de regulación pública explícita en al menos tres jurisdicciones con capacidad de implementación, siguiendo la lógica de la hibridez constitutiva establecida en el Paper 4 y extendida en este paper al plano de la instrumentación cognitiva.

Predicción 5 — Acoplamiento empírico entre los dos planos. En diseños experimentales que midan la calidad del uso de modelos de lenguaje como instrumentos de potenciación cognitiva sobre poblaciones expuestas a distintos niveles previos de uso de redes sociales optimizadas por engagement, el grupo con mayor exposición previa mostrará menor calidad de uso del instrumento, controlando por las variables individuales relevantes. Si la trampa de segundo orden opera como el paper propone, la exposición a Plano 1 debe correlacionar negativamente con la calidad del uso de Plano 2. Si no hay correlación, la articulación causal entre los dos planos requiere revisión.

9.4 La pregunta abierta

La pregunta que este paper deja abierta es estructuralmente análoga a la que el Paper 4 dejó abierta y se encuentra con ella en el mismo terreno: si la autoridad para condicionar el régimen de privilegio público sostenido existe, si tiene la voluntad de ejercerla y si tiene los instrumentos para hacerlo antes de que los efectos que la serie describe produzcan transformaciones difíciles de revertir. Lo que este paper agrega es la dimensión cognitiva de esa pregunta: la autoridad para condicionar la operación de la instrumentación cognitiva masiva, la voluntad política de ejercerla en el plano de los corpus, las funciones de pérdida y los procedimientos de fine-tuning que codifican los presupuestos sobre los que la población elabora pensamiento, y la urgencia de hacerlo antes de que la trampa de segundo orden haga estructuralmente más costosa la regulación de cada plano por la degradación que la operación del otro produce.

Los algoritmos ya no filtran solo información ni solo a las personas que conocemos. Filtran, cada vez más, los marcos conceptuales sobre los que pensamos. Y la infraestructura que produce ese filtrado no es solo de distribución sino también de instrumentación, no es exclusivamente privada y el daño que produce no es una externalidad de mercado. La categoría jurídica adecuada para abordar el peligro completo no requiere ser inventada: requiere ser reconocida en su unidad y aplicada en sus dos planos articulados. Eso es lo que la categoría de soberanía cognitiva, articulada en este paper como objeto regulatorio unificado, hace posible.

10. Registro de afirmaciones

De acuerdo con el estándar epistémico de la serie, este registro distingue el estatus de cada tipo de afirmación que el paper realiza, con el propósito de permitir al lector evaluar el estatus de cada proposición y de orientar la investigación empírica que podría verificar o falsificar las hipótesis.

10.1 Observaciones empíricas con respaldo citable

- Palantir Technologies publicó un documento de 22 puntos en X el 18 de abril de 2026, presentado como resumen de The Technological Republic de Alex Karp (verificable en registros públicos).
- Anthropic ha rechazado públicamente colaborar en aplicaciones específicas relacionadas con armas autónomas bajo presión de la administración Trump (verificable en declaraciones públicas de la empresa durante 2026).
- Need for Cognition, Cognitive Reflection Test, Actively Open-minded Thinking y Apertura a la experiencia son constructos psicológicos estables, replicados en múltiples contextos culturales, con variación interindividual sustancial documentada y carácter parcialmente entrenable (Cacioppo y Petty, 1982; Frederick, 2005; Baron, 1993; Stanovich y West, 2007; Costa y McCrae, 1992).
- La variación en estas dimensiones dentro de cualquier grupo cultural o étnico definible supera la variación entre grupos (Nisbett et al., 2012).
- Los modelos de lenguaje de gran escala fueron entrenados mayoritariamente sobre corpus en inglés, en proporciones que superan en órdenes de magnitud al disponible en otras lenguas (documentado en papers técnicos de entrenamiento de los principales laboratorios).
- DeepSeek demostró capacidad de entrenamiento de modelos competitivos con recursos computacionales significativamente menores que los utilizados por los laboratorios líderes (verificable en reportes técnicos publicados, enero 2025).
- Los marcos regulatorios actuales (AI Act europeo 2024, Digital Services Act 2022, executive orders estadounidenses) operan sobre el supuesto de actor externo identificable que produce contenido o sobre la moderación de outputs específicos del sistema, no sobre los presupuestos arquitectónicos del corpus, la función de pérdida y los procedimientos de fine-tuning.

10.2 Hipótesis teóricas con argumento lógico pero sin verificación empírica directa

- Los corpus de entrenamiento, las funciones de pérdida y los procedimientos de fine-tuning de los modelos de lenguaje codifican presupuestos filosóficos, culturales y políticos que producen, en uso ordinario, fluidez diferencial sobre distintos tipos de razonamiento, con consecuencias estructurales sobre la elaboración del pensamiento del usuario.
- Existe una analogía estructural en sentido formal preciso entre el peligro a la soberanía cognitiva descrito por los Papers 0 a 4 y el peligro derivado del desarrollo exponencial de la

IA, en cinco dimensiones específicas: hibridez constitutiva, codificación de presupuestos no neutrales, operación bajo umbral de detección consciente, efectividad diferencial y producción de dependencia bajo apariencia de potenciación.

- El modo alarma basal del Paper 3 produce, en tiempo real y mientras opera, una reducción funcional de las dimensiones psicológicas (NFC, CRT, AOT, Apertura) que protegerían al individuo de los mecanismos del Plano 1 y que determinan la calidad del uso del instrumento del Plano 2 como instrumento de potenciación genuina.
- La efectividad del Mecanismo 1 sobre la elaboración del pensamiento del usuario varía sistemáticamente con la posición del usuario en las dimensiones psicológicas examinadas, en analogía con la efectividad diferencial documentada para los mecanismos del Plano 1 en los Papers 1 a 3.
- La economía del silencio que durante décadas protegió la opacidad sobre la hibridez constitutiva opera como sistema de incentivos que está siendo perturbado por la convergencia entre masividad del impacto, capacidad creciente de auditoría algorítmica externa y presión política específica.
- El forzamiento filosófico, expresado en formas estructuralmente distintas según el actor —Palantir como declaración ofensiva, Anthropic como negativa operativa con contenido filosófico—, es la consecuencia visible del colapso de la economía del silencio en el plano de la instrumentación cognitiva.

10.3 Extrapolaciones propuestas como agenda de investigación futura

- La infraestructura cognitiva pública como categoría institucional necesaria, articulada como conjunto de condiciones que garantizan acceso a instrumentos cognitivos de calidad, formación metacognitiva efectiva y participación deliberativa en la determinación de los presupuestos del corpus, constituye la consecuencia institucional obligada de la unificación categorial propuesta.
- La categoría de soberanía cognitiva, articulada como condición de posibilidad de la formación autónoma de juicios y la elaboración de pensamiento propio, es la categoría regulatoria correcta para integrar los dos planos del peligro y debe sustituir, como objeto regulatorio unificado, las categorías separadas de regulación de plataformas y regulación de IA en los marcos vigentes.
- La distinguibilidad —de las señales en el grafo, de los presupuestos del instrumento— constituye la condición de posibilidad sustantiva de la soberanía cognitiva en ambos planos, condición que el Paper 6 examina en su escenario de erosión estructural.
- La convergencia geopolítica hacia la regulación explícita de los corpus de entrenamiento como objeto de soberanía estatal en los próximos cinco años se propone como predicción derivada del armazón.

Referencias bibliográficas

Baron, J. (1993). Why teach thinking? An essay. Applied Psychology, 42(3), 191–214.

Bleynat, S. (2026a). La Hipótesis de la Alteridad Opcional: del F-117 al espejo perfecto. Preprint. Zenodo / SSRN. DOI: 10.5281/zenodo.18880194

Bleynat, S. (2026b). Muestreo social y mediación algorítmica. Preprint. Zenodo / SSRN. DOI: 10.5281/zenodo.18946134

Bleynat, S. (2026c). Condicionamiento evaluativo distribuido en el grafo social: transferencia de credibilidad política sin persuasión directa. Preprint. Zenodo / SSRN. DOI: 10.5281/zenodo.19701840

Bleynat, S. (2026d). La alarma como estado basal: optimización algorítmica, activación emocional sostenida y sus consecuencias epistémicas. Forthcoming — Zenodo.

Bleynat, S. (2026e). Infraestructura híbrida: capital público, función objetivo y gobernanza de la episteme algorítmica. Forthcoming — Zenodo.

Bleynat, S. (2026f). El grafo sin testigos: agentes de inteligencia artificial, autocatálisis epistémica y la transición hacia la internet sintética emergente. Forthcoming — Zenodo.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. Journal of Personality and Social Psychology, 42(1), 116–131.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. Psychological Bulletin, 119(2), 197–253.

Costa, P. T., & McCrae, R. R. (1992). NEO PI-R Professional Manual. Psychological Assessment Resources.

Couldry, N., & Mejias, U. A. (2019). The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism. Stanford University Press.

European Parliament and Council. (2022). Digital Services Act (Regulation EU 2022/2065). Official Journal of the European Union.

European Parliament and Council. (2024). Artificial Intelligence Act (Regulation EU 2024/1689). Official Journal of the European Union.

Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19(4), 25–42.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411–437.

Galesic, M., Olsson, H., & Rieskamp, J. (2018). A sampling model of social judgment. Psychological Review, 125(3), 363–390.

Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.

Latour, B. (2005). Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press.

Mazzucato, M. (2013). The Entrepreneurial State: Debunking Public vs. Private Sector Myths. Anthem Press.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. American Psychologist, 67(2), 130–159.

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of persuasion. Advances in Experimental Social Psychology, 19, 123–205.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

Srnicek, N. (2016). Platform Capitalism. Polity Press.

Stanovich, K. E. (2011). Rationality and the Reflective Mind. Oxford University Press.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. Thinking & Reasoning, 13(3), 225–247.

Sunstein, C. R. (2017). #Republic: Divided Democracy in the Age of Social Media. Princeton University Press.

Taylor, C. (1995). Philosophical Arguments. Harvard University Press.

Varoufakis, Y. (2023). Technofeudalism: What Killed Capitalism. Bodley Head.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Walther, E. (2002). Guilty by mere association: Evaluative conditioning and the spreading attitude effect. Journal of Personality and Social Psychology, 82(6), 919–934.

Winner, L. (1980). Do artifacts have politics? Daedalus, 109(1), 121–136.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 112(4), 1036–1040.

Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs.

Sobre esta serie de papers

Este trabajo forma parte de una serie de papers bajo el título general «Arquitecturas de Influencia Algorítmica: mecanismos, condiciones y consecuencias».

Paper 0 (ancla): La Hipótesis de la Alteridad Opcional — Bleynat (2026a) | DOI: 10.5281/zenodo.18880194

Paper 1: Muestreo social y mediación algorítmica — Bleynat (2026b) | DOI: 10.5281/zenodo.18946134

Paper 2: Condicionamiento evaluativo distribuido en el grafo social — Bleynat (2026c) | DOI: 10.5281/zenodo.19701840

Paper 3: La alarma como estado basal — Bleynat (2026d) | DOI: 10.5281/zenodo.19716563

Paper 4: Infraestructura híbrida: capital público, función objetivo y gobernanza de la episteme algorítmica — Bleynat (2026e) | DOI: 10.5281/zenodo.19802912

Paper 5: La extensión del peligro: desarrollo exponencial de la inteligencia artificial y soberanía cognitiva en el plano de la instrumentación [este trabajo, versión 2.0]

Paper 6: El grafo sin testigos: agentes de inteligencia artificial, autocatálisis epistémica y la transición hacia la internet sintética emergente — Bleynat (2026f) | Forthcoming — Zenodo

Paper de síntesis: Epistemología del Grafo — Hacia una teoría de la formación algorítmica de la realidad social | Forthcoming — Zenodo

Paper 5 — La extensión del peligro · Serie: Arquitecturas de Influencia Algorítmica · Bleynat (2026)