PREPRINT — SSRN / ZENODO
Series: Architectures of Algorithmic Influence · Paper 5

The Extension of the Threat: exponential development of artificial intelligence and cognitive sovereignty in the plane of instrumentation

—

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

Sergio Bleynat
Microsyst — Consulting in AI and Digital Transformation
www.microsyst.com.ar

April 2026
DOI: [pending assignment in SSRN/Zenodo]

Keywords: cognitive sovereignty; cognitive instrumentation; training corpus; philosophical assumptions; philosophical forcing; Palantir manifesto; Anthropic; second-order trap; interindividual variation; Need for Cognition; Cognitive Reflection Test; Openness to Experience; cognitive public infrastructure; technofeudalism; legitimate conditioning of privilege.

Methodological note

This paper follows the epistemic typology established in Bleynat (2026a) and applied with consistency throughout the series: it explicitly distinguishes between empirical observations with citable support, theoretical hypotheses with logical argument but without direct empirical verification, and analytical extrapolations proposed as an agenda for future research. The Register of Claims in Section 10 materializes that distinction.

The paper also operates on an additional methodological premise that should be declared from the outset. Its central thesis introduces neither a new mechanism with respect to the argumentative body of the series nor an unprecedented empirical object: it introduces an analytical operation —the identification of a structural analogy between two threats to cognitive sovereignty operating in distinct planes. That operation has the same juridical-epistemological form that Paper 4 executed on the category of private actor: it shows that a prevailing categorial framework —in this case, the framework that confines the dispute over cognitive sovereignty to the plane of social perception—

is structurally insufficient to address the problem at its correct level, and derives the normative consequences of that insufficiency.

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 argument 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 the 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 —the set of institutional conditions guaranteeing access not mediated by the objective function of private actors to high-quality cognitive potentiation instruments— is proposed as the obligated institutional consequence of the recognition of constitutive hybridity 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. The condition of visibility of the philosophical dispute among actors presupposes the prior existence of the disputable object —cognitive instrumentation as infrastructure— and its recognition as such.

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. The problem: a second plane of threat that the prevailing framework does not name

This entire series has argued, across five cumulative papers, that there exists a structural threat to cognitive sovereignty that contemporary regulatory theory does not name with precision and that, for that reason, it cannot address with effective instruments. Papers 0 to 3 described the mechanisms by which that threat operates on the individual's social perception: the algorithmic selection of filtered versions of known contacts, the distortion of the sampling of social instances that feeds prevalence inference, the transfer of evaluative valence toward political objects without direct persuasion, the consolidation of the alarm mode as the basal state of the system. Paper 4 showed that those mechanisms are not market externalities but properties of a constitutively hybrid infrastructure, and built the legal framework —irreducibly common good, structural vice of consent, modal tripartition of imputation, legitimate conditioning of sustained privilege— that renders those mechanisms accessible to regulation.

That conceptual framework, however, operates entirely on one plane: the plane in which algorithmic infrastructure is infrastructure of social perception. Recommendation systems —Facebook, Instagram, X, TikTok— modify the cognitive availability of the individual's social environment: what sampling is performed, which nodes of the graph reach them, in which version, with what activation. The objective function that Paper 4 proposes to modify is the objective function of social-content distribution. The regulation that Paper 4 proposes as legitimate conditioning of sustained privilege is regulation over the distribution of platform-mediated social perception.

The thesis of this paper is that the exponential development of artificial intelligence —and specifically the emergence of large language models as instruments available at massive scale for the elaboration of individual thought— produces a second structural threat to cognitive sovereignty, which operates in a plane distinct from the one described by previous papers and for which, for that reason, the legal instruments of Paper 4 are necessary but not sufficient.

The new plane is not that of social perception of the environment: it is the plane of the cognitive instrumentation of one's own thought. When an individual converses with a language model to refine an argument, explore a concept, write a text or make a decision, that individual is not being exposed to an algorithmic distribution of content produced by others: they are deliberately using an instrument that encodes, in its deep architecture, philosophical, cultural and political assumptions that they did not choose and, in many cases, cannot detect. The elaboration of thought, which in the classical-liberal tradition has been understood as an internal exercise of individual faculty and as the condition of possibility of all autonomy, is now produced —with increasing intensity and for a growing portion of the population— through an instrument whose architecture is not neutral and whose assumptions are subject to no regime of transparency or governance.

The substantive thesis of this paper is, in consequence, that there exists a structural analogy between the two threats. The word "analogy" demands precision here: this is not a metaphorical similarity nor a rhetorical parallelism. It is that the two threats share an identical formal structure in five specific dimensions developed in Section 3. Sharing that structure does not mean the threats are identical: they operate on distinct objects, through partially distinct mechanisms, and demand partially distinct regulatory instruments. But the common formal structure is what permits recognition of the second threat as a threat to cognitive sovereignty in a strict sense, and not as a separate problem demanding a new conceptual category.

Recognition of this structural analogy has an immediate consequence for regulatory theory. If the prevailing frameworks are insufficient for the first plane of the threat —Paper 4's thesis—, they are a fortiori insufficient for the second, because the second plane does not even appear as a regulatory object in available frameworks. The European AI Act, U.S. executive orders, multilateral governance proposals operate on the assumption that the problem of artificial intelligence is a problem of technical safety of the system —biased outputs, failures, misuse risks— and not a problem of architectural encoding of non-neutral assumptions in massive cognitive instruments. They regulate system behaviors, not the objective function and deep assumptions that produce those behaviors. They demand transparency about outputs, not about the training corpus that determines what types of thought the system amplifies with ease and which with resistance.

The paper the reader has before them therefore fulfills a precise architectural function in the economy of the series. It makes explicit the analytical operation the series had been preparing: the recognition that the category of cognitive sovereignty, introduced in earlier papers as a normative horizon, does not exhaust its content in the plane of social perception and must also be articulated on the plane of cognitive instrumentation. And it derives from that articulation the institutional consequences that Paper 4's framework, designed for the first plane, cannot produce by itself: cognitive public infrastructure as a necessary institutional category, not as a normative addendum.

The paper is organized as follows. Section 2 examines what existing theoretical frameworks capture and what they leave uncaptured regarding the second plane of the threat. Section 3 develops the central conceptual move: the formal articulation of the structural analogy between the two planes of the threat to cognitive sovereignty, in its five dimensions. Section 4 details the three mechanisms through which the Plane 2 threat operates: the philosophical encoding of the corpus, the philosophical forcing declared by the actors that control the instruments, and the differential effectiveness over interindividual variation. Section 5 examines the causal articulation between the two planes —the second-order trap— as the architectural piece of the argument. Section 6 articulates the category of cognitive sovereignty as a unifying regulatory object and derives cognitive public infrastructure as the institutional consequence. Section 7 responds to foreseeable objections. Section 8 examines the implications for regulatory theory, including the connection with Paper 6, where the threat described reaches its most radical form: the collapse of the very distinguishability of the human source of the signal. Section 9 closes with conclusions and falsifiable empirical predictions. Sections 10 and 11 contain the register of claims and the references.

2. State of the art: the second plane that no existing framework occupies with precision

2.1 AI governance and its bias toward technical safety

The most developed frameworks of artificial intelligence governance —the European AI Act (2024), the executive orders of the U.S. administration, the proposals of multilateral bodies such as UNESCO and OECD— share an assumption that the previous papers in this series enable us to identify as a structural limitation: they assume that the problem of AI is a problem of technical safety of the system and not a problem of epistemic architecture. They regulate system behaviors —biased outputs, identifiable failures, misuse risks— without touching the objective function that produces those behaviors or the assumptions encoded in the training corpus that determine what types of thought the system amplifies with ease and which require additional effort to produce.

The academic literature on AI risks has produced valuable analyses on alignment, safety and governance (Russell, 2019; Bostrom, 2014; Gabriel, 2020), but it has concentrated its attention on two poles of the problem: the pole of the most capable systems —artificial general intelligence, existential risks, superintelligence alignment— and the pole of identifiable discriminatory effects on specific populations —gender, racial, socioeconomic biases in classification, credit scoring, recidivism prediction. That range of attention has left comparatively underattended the intermediate space in which this series operates and in which this paper develops its thesis: the cognitive and epistemic effects of systems of generative and inferential AI on general populations in the immediate present, mediated by the ordinary and apparently benign use of the instruments.

2.2 The critique of platform capitalism and its limitation of plane

The most complete critical tradition on the business model of algorithmic platforms —Zuboff (2019) on surveillance capitalism, Srnicek (2016) on platform capitalism, Couldry and Mejias (2019) on the colonization of experience, Varoufakis (2023) on technofeudalism— identifies with precision the logic of the problem in the plane of extraction and modulation of user behavior. But that tradition operates entirely on the plane that Papers 0 to 4 describe: that of algorithmic infrastructure as a system of modulation of perception and behavior. The second plane —that of the cognitive instrumentation of one's own thought through large language models— appears occasionally as a peripheral theme but not as a central theoretical object.

Varoufakis (2023) merits a specific mention because his category of technofeudalism, articulated to describe the structural displacement from industrial capitalism to platform-rent capitalism, becomes even more precise when applied to the second plane of the threat. Under industrial capitalism, the worker sold labor power to a capitalist who organized it productively. In the technofeudalism of distribution platforms, the user provides behavior to the digital feudal lord who extracts rent from that behavior. In technofeudalism extended to the plane of cognitive instrumentation, the individual accesses an instrument of thought potentiation whose architecture is the property of a cognitive feudal lord, and the elaboration of individual thought is produced within that fief. The extension of Varoufakis's category to the second plane is direct, but it was not developed by Varoufakis and does not appear systematized in the available literature.

2.3 The philosophy of technology and the problem of the neutrality of artifacts

A tradition to which the literature on platform regulation rarely resorts with systematicity and which this paper requires to mobilize in order to sustain its substantive thesis on the non-neutrality of the training corpus is the philosophy of technology that examines political inscription in artifacts. Winner (1980) argued, in his classic "Do Artifacts Have Politics?", that certain technical artifacts have politics in a strong sense: they incorporate in their material design the crystallization of power relations, political decisions and cultural assumptions that continue to operate independently of the intention of those who use them. Latour (2005) developed that thesis to the extreme of actor-network theory, in which artifacts are full-sense actors, not neutral instruments used by human actors.

The tradition of the philosophy of technology provides the general categorial framework within which the thesis on the training corpus of language models becomes theoretically articulable. A language model is not a neutral instrument with accidentally biased assumptions. It is a political artifact in Winner's most precise sense: the computational crystallization of a historically situated corpus, a loss function that selects certain regularities and discards others, and a set of safety constraints reflecting specific ethical and political commitments. That crystallization produces, in ordinary use, effects on the user's thought elaboration that are consequences of the architecture, not

operational failures. The available philosophy of technology permits the formulation of that observation; what it does not provide is the regulatory framework that renders that observation accessible to institutional design, which is what this paper proposes to construct, in analogy with what Paper 4 constructed for the first plane of the threat.

2.4 The empty space

None of the traditions examined connects the facts about the hybrid nature of algorithmic infrastructure —documented by Paper 4— with the extension of the threat to the plane of cognitive instrumentation. AI governance operates on technical safety without recognizing the architectural-political plane of the corpus. The critique of platform capitalism operates on the plane of behavioral modulation without systematically extending to the instrumentation of thought. The philosophy of technology provides the general framework but not the regulatory framework. The regulatory theory of platforms operates on the assumption of private actor without examining it —and, when it examines it (Paper 4), it does not extend the analysis to the cognitive plane. The work of connecting those four domains and projecting them onto the specific problem of cognitive sovereignty in the plane of instrumentation is the operation this paper performs.

3. The central conceptual move: the structural analogy between two threats affecting cognitive sovereignty

This section develops the architectural move that sustains the entire paper: the formal articulation of the structural analogy between the threat described by Papers 0 to 4 and the threat that the exponential development of AI introduces in the plane of cognitive instrumentation. The analogy is neither rhetorical nor heuristic. It is structural in a precise formal sense: the two threats share a common structure in five specific dimensions. Recognizing that common structure is what permits treating the second threat as a threat to cognitive sovereignty in the strict sense and not as a separate problem requiring an independent conceptual category.

3.1 The first threat recapitulated: cognitive sovereignty in the plane of social perception

Papers 0 to 4 documented, with growing precision, a structural threat to the individual's cognitive sovereignty in their capacity to form autonomous judgments about the social reality of their environment. Paper 0 (Bleynat, 2026a) described the most basic mechanism —the Optional Alterity Hypothesis—: the algorithm selects, for each observer, the most resonant versions of each contact, producing solitude of presence under the appearance of dense connection. Paper 1 (Bleynat, 2026b) showed that this mechanism operates at the level of the Social Sampling Model: the sample of social instances that the individual uses to infer the reality of their environment is systematically biased toward the high-activation dimensions of their contacts. Paper 2 (Bleynat, 2026c) described

Distributed Evaluative Conditioning in Graph: the transfer of affective valence from the trusted contact toward political objects with which it coexists in the feed, without direct persuasion or intention on the contact's part. Paper 3 (Bleynat, 2026d) documented the alarm mode as the basal state of the system: engagement optimization selects content of high emotional activation, which produces a state of chronic alarm that reduces tolerance for ambiguity, favors heuristic over deliberative processing and activates the mechanism of epistemic anesthesia. Paper 4 (Bleynat, 2026e) constructed 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.

The threat to cognitive sovereignty in this plane can be formulated with precision: the individual exposed to the normal operation of recommendation systems loses access, without knowing it and without any individual failure being able to explain it, to a non-systematically-distorted representation of their social environment. They form judgments about what "everyone thinks", about who each of their contacts is, about what one political position is and what another is, on the basis of a sample whose distortion is an emergent property of the system's objective function. Cognitive sovereignty —understood as the effective capacity to form autonomous judgments about social reality— is compromised by the mediation of an instrument whose design they did not choose and whose operation they cannot detect while experiencing it.

3.2 The second threat declared: cognitive sovereignty in the plane of cognitive instrumentation

The exponential development of large language models —from the emergence of ChatGPT at the end of 2022 to the conversational systems of mass use in the present— introduces a second plane of threat to cognitive sovereignty, distinct from the previous one but structurally analogous. The new plane is that of cognitive instrumentation: the use of the model as a deliberate instrument for the elaboration of one's own thought, the exploration of concepts, writing, practical reasoning, decision-making.

The threat in this plane can be formulated with precision analogous to the threat of the first plane: the individual who uses a language model to elaborate thought elaborates that thought through an instrument whose deep architecture encodes philosophical, cultural and political assumptions that are not neutral —that were not chosen by the user, that the user generally cannot detect in ordinary use, and that operate on the elaboration of thought in a structured way. The instrument, by its nature, amplifies certain types of reasoning with ease and requires additional effort to produce others; suggests certain analogies naturally and displaces others toward the periphery; applies certain conceptual frameworks by default and resists others. That differential fluency is not a design failure that could be corrected: it is the inevitable consequence of training over any historically situated corpus under any specific loss function.

The consequence for cognitive sovereignty is direct: the individual using the instrument believes they are elaborating their own thought and, in a strict sense, is doing so —their agency is not annulled. But the elaboration is produced in a space structured by assumptions the individual did not choose and, in many cases, cannot even detect while inhabiting them. Deliberative autonomy, understood in the most demanding classical-liberal tradition as the individual's effective capacity to reason from premises they recognize as their own, is compromised in a structural dimension distinct from those described by Papers 0 to 4 but formally analogous.

3.3 The formal structure of the analogy: five dimensions

The analogy between the two threats is structural in a precise formal sense. They share five specific dimensions that should be developed individually because each enables part of the regulatory framework the paper builds in subsequent sections.

Dimension 1 — Constitutive hybridity of the infrastructure

Both the distribution infrastructure (Plane 1) and the cognitive instrumentation infrastructure (Plane 2) are constitutively hybrid in the precise sense established by Paper 4: built and sustained by the convergence of historical state financing, current operational subsidies and academic knowledge produced with public financing. The computational infrastructure on which language models operate is the same that Paper 4 documented —ARPANET, NSFNET, foundational algorithms, the intelligence-investment ecosystem. The foundational transformer architectures of the field —from Vaswani et al. (2017) to contemporary large language models— were developed in research financed mostly by academic institutions with public funding and by private laboratories operating under fiscal-subsidy regimes in multiple jurisdictions. The Argentine Knowledge Economy Law, the Horizon Europe program, the Chinese Government-Guided Investment Funds —documented in Paper 4— equally subsidize the operation of distribution platforms and AI laboratories. Natural-language processing knowledge, the massive data on which models are trained, the human talent that designs and operates them are products of the same matrix of public and private investment that Paper 4 documented. Constitutive hybridity is a property of algorithmic-inferential infrastructure considered as a totality, not a privilege of the distribution subsystem.

Dimension 2 — Encoding of non-neutral assumptions in the architecture

Both in Plane 1 and Plane 2, the infrastructure is not neutral with respect to the object on which it operates. The objective function of recommendation systems —engagement optimization— produces the emergent properties that Papers 1 to 3 describe without any individual actor having designed them intentionally. The loss function, the training corpus, the safety constraints and the fine-tuning procedure of language models encode, in their deep architecture, philosophical, cultural, linguistic and political assumptions that are not neutral. Those assumptions produce, in ordinary

use, effects on the user's thought elaboration that are architectural consequences, not failures. Structured non-neutrality is a property common to both planes.

Dimension 3 — Operation below the user's threshold of conscious detection

In both planes, the system's operation on the individual's cognition is produced, under ordinary conditions of use, below the threshold of conscious detection. The user looking at their feed does not experience sampling distortion as distortion: they experience a representation of the social reality of their network. The user conversing with a language model does not experience the encoding of corpus assumptions as encoding of assumptions: they experience a conversation with an apparently neutral instrument that helps them think. The condition of operational invisibility is, in both planes, what distinguishes the threat from traditional forms of influence —propaganda, direct persuasion, censorship— that operate below the competent receiver's threshold of critical detection. That invisibility is what Paper 4 articulated juridically as the structural vice of consent, and applies without substantial modification to the second plane.

Dimension 4 — Differential effectiveness over variable cognitive vulnerabilities

Both the Plane 1 threat and the Plane 2 threat operate with differential effectiveness over individual psychological dimensions documented by the available empirical literature. The mechanisms of Papers 1 to 3 are more efficient on profiles with low basal Need for Cognition, low Openness to Experience, high Neuroticism and low Cognitive Reflection Test scores, exactly where elaborative defenses are weaker and where the alarm mode produces greater functional reduction of the central route of processing. The corpus assumptions of language models operate with greater or lesser effectiveness on the user's thought elaboration as a function of the same dimensions, for a reason developed in Section 5 and which is the most important architectural piece of this paper: the dimensions that protect the individual from Plane 1 mechanisms are exactly the dimensions that determine the quality of use of the Plane 2 instrument as an instrument of genuine thought potentiation.

Dimension 5 — Production of dependence under the appearance of potentiation

Both planes of the threat share a phenomenological property that should be articulated with care: the system's operation produces functional dependence of the individual on the instrument, under subjective appearance of potentiation or expansion of capacity. The user of social networks experiences access to information, to contacts, to perspectives that without the platform they would not have —and that experience is, in some sense, true; the user has access to something. The user of language models experiences intellectual elaboration capacity that without the instrument they would not have —and that experience is also true in some sense. But in both cases the experienced potentiation coexists with the structural production of dependence: the progressive loss of the capacity to operate on the corresponding plane without the instrument's mediation, and the

transformation of the mediation into a condition of operation. The analytical distinction between genuine potentiation and functional dependence under the appearance of potentiation is central to the normative framework and will be developed with precision in Section 5.

3.4 What the analogy does not affirm

Before proceeding, it is worth specifying what the proposed structural analogy does not affirm, in order to avoid predictable objections that rest on inflationary readings of the thesis. The analogy does not affirm that the two threats are identical. They operate on distinct objects —the social perception of the environment in one case, the elaboration of one's own thought in the other—, through partially distinct mechanisms and demand partially distinct regulatory instruments. The analogy does not affirm that language models are social networks: they are structurally different technologies with distinct business models and distinct dynamics of use. The analogy does not affirm that the use of language models is always and necessarily harmful: Phenomenon 3 that previous versions of this paper articulated —the partial democratization of access to high-level cognitive interlocution— is real and has substantive positive consequences, particularly for populations that historically lacked access to dense intellectual networks.

What the analogy does affirm is something more precise and, for that reason, more demanding: that the formal structure of the two threats is the same in the five dimensions examined, and that this structural coincidence permits and demands treating the second plane as a regulatory object within the category of cognitive sovereignty, not as a separate problem in an independent conceptual category. Categorial unification is the theoretical consequence of the analogy. The institutional consequences of that unification —which Section 6 articulates— are its practical consequence.

4. The three mechanisms of the Plane 2 threat

Once the structural analogy is established, this section details the three specific mechanisms through which the Plane 2 threat operates. The structure parallels that which Papers 1, 2 and 3 articulated for Plane 1: each mechanism describes a dimension of the way in which the cognitive-instrumentation instrument produces effects on the user's thought elaboration.

4.1 Mechanism 1 — The philosophical encoding of the corpus

A language model is, in its most basic dimension, a statistical distillation of the corpus on which it was trained, mediated by a specific loss function and refined by fine-tuning procedures that reflect ethical and political commitments of the laboratory that produces it. That distillation is not neutral: the corpus reflects the conceptual frameworks, categories of analysis, value hierarchies, stylistic regularities and, crucially, the absences proper to the historical intellectual tradition that produced it. The most capable models available globally were trained mostly on text in English, produced in

predominantly Anglo-Saxon contexts of the last century, reflecting the assumptions of post-war liberalism and methodological individualism characteristic of contemporary Western academia, with all its internal contradictions. Not by conspiracy nor by deliberate political intent of the laboratories: by availability. The digitalized text in English of training-suitable quality exceeds by several orders of magnitude that available in any other language or intellectual tradition.

The observable result —which should be formulated as a hypothesis, given that direct verification would require access to the internal evaluation systems of the laboratories and to benchmarks specifically designed to detect the phenomenon— is that models produce certain types of reasoning with greater fluency than others. The analogies they suggest naturally, the conceptual frameworks they apply by default, the questions they answer without resistance and those that require specific formulation to produce a substantive response, the positions they assume by default in unresolved philosophical debates, the intellectual traditions they treat as natural starting points and those they treat as exceptions to be justified, all reflect the distribution of the training corpus and the commitments of fine-tuning. That differential fluency is not a design failure that could be corrected through superficial adjustments. It is an architectural consequence of training over any historically situated corpus under any specific loss function.

The consequence for the user's thought elaboration is structural. A user who converses with the model in a language, on an intellectual tradition or on a conceptual framework under-represented in the training corpus experiences the instrument as less fluent, less suggestive, less productive in that zone, and more fluent, more suggestive, more productive in the more represented zones. The asymmetry translates, without the user detecting it, into a gravitational pressure on the space of thought elaboration: the zones of greater instrument fluency attract the user's thought toward them with the force of cognitive economy, and the zones of lesser fluency are progressively abandoned or become the place where the user must exert effort to sustain an elaboration the instrument does not facilitate. Corpus encoding does not censor any type of thought: it differentially disincentivizes it. The aggregate effect, sustained over time and at population scale, is the progressive homogenization of the conceptual frameworks on which thought is elaborated, in the direction of the assumptions of the dominant corpus.

The argument is not reduced, however, to the linguistic-cultural bias of the corpus. The loss function and the fine-tuning procedures add a second layer of philosophical encoding: the safety constraints —what types of response are permitted, what topics are restricted, what political positions the model assumes by default, what moral conflicts the model resolves toward one side or another— reflect specific ethical and political commitments of the laboratory. Those commitments may be reasonable or not, defensible or not; what matters for the argument of this paper is that they exist, are not neutral, and operate on the user's thought elaboration without any regime of transparency or governance making them visible or subjecting them to public deliberation.

4.2 Mechanism 2 — Philosophical forcing as visible evidence of the displacement

Mechanism 1 operates structurally and, by its nature, is difficult to make visible to public opinion: it requires technical analysis of corpora and fine-tuning procedures, systematic comparisons among models trained with distinct assumptions, specific benchmarks to detect differences in the ease with which certain types of reasoning are produced. The operation of Mechanism 1 does not produce, by itself, direct public evidence of its existence.

Mechanism 2 that this subsection describes is the visible face of the displacement that the paper articulates. We call it philosophical forcing because it describes a precise phenomenon: the moment when actors that control cognitive instrumentation are compelled, by the structural dynamics of the exponential development of AI, to publicly declare the philosophical and political assumptions on which they operate. Before the recent massiveness of language models, the corpus's philosophical assumptions operated in the silence of the technical product that appeared neutral. The massiveness and growing visibility of the models' impact on the thought elaboration of entire populations renders that declared neutrality unsustainable and produces, as a rational response by the relevant actors, explicit declarations of the assumptions on which each operates.

The concept of economy of silence can be understood as the protective property of the regime of non-declaration: for decades, the actors operating in the zone of constitutive hybridity among private capital, the State, intelligence systems and technological markets maintained opacity about their assumptions because silence reduced regulatory cost, avoided attribution of responsibility and preserved the fiction of technical neutrality that legitimized the framework of pure private actor. That economy of silence is being disturbed by the dynamics of the exponential development of AI, and philosophical forcing is the consequence. The causality sustaining the disturbance is plausible though not completely verifiable from outside: the massiveness of the models' impact on thought elaboration, the growing technical capacity of external algorithmic auditing, and the specific political pressure that the case of Anthropic illustrates, converge to make opacity more costly than declaration.

It is worth examining the two paradigmatic cases available with the precision each requires, without forcing a symmetry where the structural differences turn out to be decisive.

The Palantir case as offensive declaration

On April 18, 2026, the official corporate account of Palantir Technologies published on X a 22-point document presented as a summary of The Technological Republic, the book by its chief executive Alex Karp. The document is not analytically relevant for its specific contents —which include claims about Silicon Valley's moral debts to the State, autonomous weapons, hierarchies among cultures and the end of the post-war neutralization of Germany and Japan— but for what its form represents:

the first explicit public declaration, signed with corporate logo and distributed at global scale, by a technological corporation that vindicates philosophical and political prominence as a conscious and desirable program. Not as accident. Not as unintended consequence. As objective.

The Palantir case is an offensive declaration in a structural sense. The company, whose trajectory since its founding in 2003 with financing from In-Q-Tel and Founders Fund places it at the very center of the zone of constitutive hybridity that Paper 4 documented, decides to come out first, with its own framing, explicitly vindicating the political philosophy that until now operated without a name. The analytically relevant question is not what Karp says. It is why now, why in this format, and what conditions produced the moment when what previously required silence can be said aloud. The plausible answer is that Palantir's strategic calculation recognizes that external auditing of its operation is becoming technically possible and politically probable, and that under such conditions it is preferable to produce the framing from within than to wait for it to be imposed from without. To come out and declare one's own philosophy before someone else infers it from one's own conduct and imputes it with an unchosen framing is rationally superior to awaiting exposure.

The Anthropic case as operational act under specific political pressure

The public refusal of Anthropic to lift its restrictions on the use of its models in fully autonomous weapons and mass domestic surveillance —documented in public declarations of the company during 2026— is an act of structurally distinct nature from the Palantir manifesto, and the difference should be specified with care to avoid an interpretive symmetry that would induce a false analytical equivalence.

Anthropic's act is not an offensive declaration of integral political philosophy analogous to Karp's manifesto. It is, strictly speaking, an operational decision under specific political pressure that has philosophical consequences but whose structure is not that of the public vindication of an integral philosophical-political program. The company faced a specific request from the U.S. administration, evaluated that request against the use principles it already had documented since its founding, and decided to publicly decline. The declination has philosophical content —in the sense that it is justified by substantive ethical commitments on the responsible use of artificial intelligence— but its structure is that of an operational refusal of a particular use of the product, not that of a programmatic declaration about the philosophical-political role the company wishes to occupy in the global order.

The distinction matters for the argument. The philosophical forcing that the paper describes does not produce two symmetric poles of a single scale —Palantir at one end, Anthropic at the other. It produces, more precisely, two structurally distinct responses to the same displacement: an offensive response (integral programmatic declaration) and a defensive response (operational refusal with philosophical content). Both are responses to the same structural displacement —opacity is no longer strategically sustainable for actors with massive cognitive instrumentation—, but they are responses of distinct type, and treating them as symmetric obscures the nature of the phenomenon.

The economy of silence is collapsing, yes; but the collapse does not produce a single declaration format. It produces the spectrum of possible formats that any actor compelled to take a public position can adopt, according to their business model, their specific political exposure and the prior commitments on which they operate.

What philosophical forcing reveals about the displacement

What both cases —the offensive and the defensive— reveal in common is not the symmetry of the spectrum but the displacement itself: the transformation of cognitive instrumentation into an object of public philosophical-political dispute. Until that moment, the fundamental philosophical question about language models —what assumptions they encode, on what corpus, with what loss function, with what safety constraints, in service of what vision of the world— operated within the laboratory, in technical publications with restricted audience and without specific political pressure to make it explicit. The massiveness of use, the growing capacity of external auditing and the geopolitical pressure produce the displacement: assumptions become an object of public dispute. That displacement is the condition of possibility of philosophical forcing, and forcing is its visible manifestation.

Philosophical forcing is, in consequence, evidence of the displacement of the disputable object, not its driver. Cognitive instrumentation becomes disputable because it becomes structurally more difficult to maintain in technical opacity. Once disputable, the actors that control it must take a position. The form of the position varies. The existence of the moment to take it does not vary: it is structural to the displacement.

4.3 Mechanism 3 — Differential effectiveness over interindividual variation

The third mechanism describes the most delicate dimension of the Plane 2 threat and should be articulated with the precision that its delicacy requires in order to avoid readings that confuse it with hierarchist arguments that structurally contradict it. The available empirical psychological literature documents, with high reliability and cross-cultural replicability, a set of individual dimensions that show substantial variation in human populations and that are directly relevant to the Plane 2 mechanisms.

The first of those dimensions is Openness to Experience from the Big Five model (Costa and McCrae, 1992), already central in Papers 1 and 2 of this series, which correlates with tolerance for ambiguity, metacognitive flexibility and preference for deliberative over heuristic processing. The second is Need for Cognition (NFC), a construct proposed by Cacioppo and Petty (1982) and systematized in decades of subsequent research (Cacioppo, Petty, Feinstein and Jarvis, 1996): the stable tendency to engage in and enjoy effortful cognitive processing. Individuals with high NFC tend to use the central route of the Elaboration Likelihood Model (Petty and Cacioppo, 1986) evaluating arguments by their merits; individuals with low NFC rely more on peripheral cues and heuristic shortcuts. The third is

the measure of the Cognitive Reflection Test (CRT) proposed by Frederick (2005): the tendency to inhibit the automatic intuitive response in order to substitute deliberate reflection. CRT scores predict resistance to cognitive biases with robustness replicated in multiple cultural contexts. The fourth is Actively Open-minded Thinking (AOT) (Baron, 1993; Stanovich and West, 2007): the disposition to consider alternative hypotheses, seek information that contradicts current beliefs and revise positions in light of evidence.

Four properties of these dimensions are critical for the argument. First: they show substantial individual variation documented in multiple studies. Second: variation within any definable cultural, ethnic or socioeconomic group exceeds between-group variation (Nisbett et al., 2012), which makes the individual —not the group— the relevant unit of analysis for the institutional argument. Third: they are influenced by educational and practical conditions, which makes them partially trainable and, therefore, a legitimate object of public policy. Fourth: they are mechanically connected, according to the available literature on dual processing, with the quality of reasoning over any cognitive-elaboration instrument, which suggests —though without specific direct empirical verification for the use of language models— that they are the dimensions that determine the quality of use of the Plane 2 instrument as an instrument of genuine thought potentiation.

The hypothesis derived from these four properties is that the effectiveness of Mechanism 1 over the user's thought elaboration varies systematically with the user's position on these dimensions. A user with high NFC, high CRT, high Openness and high AOT tends to use the model by critically evaluating its responses, formulating questions that challenge the premises the model assumes by default, maintaining willingness to be challenged by unanticipated perspectives, and recognizing when the model's fluency reflects the corpus representation in a zone and not the genuine quality of an argument. A user with low NFC, low CRT, low Openness and low AOT tends to accept the model's responses for their fluency, not to formulate questions that activate the central route of processing, to remain within the conceptual frameworks that the corpus prefers by default, and to confuse the linguistic sophistication of the output with the substantive quality of the reasoning.

The consequence is direct: the Plane 2 instrument produces, over the same population, qualitatively distinct effects as a function of the user's position on the psychological dimensions examined. For the first profile, the instrument can operate as an instrument of genuine potentiation —an interlocutor that challenges and refines thought. For the second profile, the instrument operates as an amplifier with greater sophistication of prior frameworks, producing elaborated confirmation of what was already believed under the subjective appearance of genuine intellectual exploration. The distinction between genuine cognitive potentiation and functional dependence under the appearance of potentiation —Dimension 5 of the structural analogy— is drawn, in the plane of individual use, along the psychological dimensions that the available literature documents.

This characterization immediately demands an analytical distinction that is not optional but constitutive of the argument, because the two possible readings of the same observation have radically opposed normative consequences. The first reading —which we will examine in detail in Section 7 as an anticipated objection— maintains that interindividual variation justifies hierarchy: those who can better use the cognitive instruments must govern those who cannot. It is Karp's response and, before him, that of every serious variant of technological elitism. The second reading maintains that interindividual variation obliges institutional design to compensate for the asymmetry that the systems operating on that variation produce on populations whose protective variation is unevenly distributed. It is the response this paper articulates. The distinction between the two is not one of moral sensibility. It is analytical. The first reading uses the variation but ignores that the system —as Section 5 will argue— actively degrades the protective dimensions in real time, transforming passive observation of variation into active exploitation of the asymmetry it produces.

5. The causal articulation: the second-order trap

This section develops the most important architectural piece of the paper. The structural analogy of Section 3 establishes that the two threats share a common form. The description of the mechanisms in Section 4 establishes how the Plane 2 threat specifically operates. What this section adds is the causal articulation between the two planes: the demonstration that the Plane 1 threat and the Plane 2 threat are not simply analogous but causally coupled in a precise way, and that this coupling demands treating the two threats as a single regulatory object.

5.1 The mechanical connection between the two planes

Paper 3 documented, on the basis of the available literature on dual processing, a fundamental property of the alarm mode as the system's basal state: the alarm mode reduces tolerance for ambiguity and favors heuristic over deliberative processing. That property can be reformulated, in the terms of Section 4 of this paper, as a precise proposition: the basal alarm mode produces, in real time and while it operates, a functional reduction of the psychological dimensions that determine the quality of use of the Plane 2 instrument. An individual under sustained alarm mode has, in effective terms, functionally reduced NFC with respect to their basal NFC, functionally reduced CRT, functionally reduced Openness, functionally reduced AOT. The reduction is not permanent —it reverses when the alarm mode subsides—, but it operates while it lasts.

The analytical consequence is decisive: the Plane 1 system, whose normal operation maintains the population in alarm mode as basal state, is degrading in real time exactly the dimensions that would protect the individual from the effects of the Plane 2 Mechanism 1. The trap this produces is of second order, and should be articulated with precision:

- Level 1: the Plane 1 system produces differential cognitive vulnerability over the population, with greater effectiveness on profiles with weaker basal elaborative defenses (Papers 1, 2 and 3).
- Level 2: the basal alarm mode of the Plane 1 system degrades in real time the dimensions that would allow the individual to resist those mechanisms —it functionally reduces NFC, CRT, Openness and AOT during exposure.
- Level 3: the same dimensions that the Plane 1 system degrades in real time are the dimensions that determine the quality of use of the Plane 2 instrument as an instrument of genuine thought potentiation.

The individual exposed to the normal operation of Plane 1 systems accumulates, in consequence, two simultaneous effects that mutually reinforce each other: a growing vulnerability to Plane 1 mechanisms, and a decreasing capacity to use the Plane 2 instrument in a way that genuinely potentiates thought instead of confirming with greater sophistication prior frameworks. The second-order trap is the precise denomination of this articulation: the system produces vulnerability and simultaneously degrades the capacities that would allow escape from it through the tool that apparently would compensate for it.

5.2 Why the articulation matters for regulation

The causal articulation between the two planes has specific regulatory consequences that no prevailing framework captures. The Plane 1 regulation that Paper 4 proposes —modification of the objective function, legitimate conditioning of sustained privilege, application of strict liability over the structurally risky activity— remains necessary, but is not sufficient to address the complete threat. For two reasons that should be articulated separately.

First reason: even assuming effective regulation of Plane 1, Plane 2 operates independently. Even if the objective function of recommendation systems were modified tomorrow in the sense that Paper 4 prescribes, language models would continue to encode the non-neutral assumptions of Mechanism 1, the actors that control them would continue to face the displacement that Mechanism 2 describes, and the differential effectiveness of Mechanism 3 would continue to operate over the population's interindividual variation. The regulation of Plane 1 does not produce, by itself, the regulation of Plane 2.

Second reason: the regulation of Plane 2 that this series proposes presupposes capacities in the population that the Plane 1 system is actively degrading. Any Plane 2 regulatory strategy operating through education, metacognitive literacy, or critical training for the use of cognitive instruments presupposes that the population to which that strategy is directed has effectively available the psychological dimensions documented in Section 4. The second-order trap establishes that this presumption does not hold if the operation of Plane 1 is simultaneously degrading those dimensions.

The regulatory consequence is that the two planes must be treated together: regulation of Plane 2 without regulation of Plane 1 is structurally ineffective, and regulation of Plane 1 without regulation of Plane 2 is insufficient to address the complete threat.

This articulation is the precise justification of the categorial unification that Section 6 articulates: why the category of cognitive sovereignty, not the separate categories of platform regulation and AI regulation, is the correct regulatory category. Unification is not a conceptual preference: it is an analytical consequence of the second-order trap.

6. Cognitive sovereignty as unifying category and cognitive public infrastructure

6.1 The operational definition of cognitive sovereignty

The previous sections establish the premises. This section articulates the central normative consequence of the paper: the category of cognitive sovereignty, articulated as the condition of possibility of the autonomous formation of judgments, integrates the two planes of the threat into a single regulatory object. The proposed operational definition is the following:

> *Cognitive sovereignty is the effective capacity of individuals and communities to form autonomous judgments about social reality and to elaborate their own thought, using mediation instruments —of social perception and of cognitive instrumentation— whose assumptions are known and deliberatively accepted, with access to the potentiation resources that make high-density intellectual elaboration possible, and without uncompensated exposure to influence mechanisms operating on their cognitive vulnerabilities without their knowledge.*

The definition articulates three components corresponding to three analytically distinct but institutionally complementary levels of intervention. The individual level —the trainable metacognitive capacity that determines the quality of use of the potentiation instrument and resistance to mediation mechanisms— is the level on which the differential effectiveness of Mechanism 3 operates and which the second-order trap makes urgent to actively compensate through public policy. The collective level —the training corpus and the objective function as bearers of non-neutral assumptions— is the level that geopolitical dispute is incipiently articulating and that philosophical forcing makes visible. The institutional level —the governance of the objective function and of corpus assumptions as the condition of possibility of collective cognitive sovereignty— is the direct extension of Paper 4's framework to the plane of instrumentation.

6.2 The extension of Paper 4's legal framework

Categorial unification permits extending Paper 4's legal framework to the plane of instrumentation with analytical precision. Each piece of Paper 4's framework has its counterpart in the new plane, and should be articulated.

The irreducibly common good, which Paper 4 identified as the informational and deliberative integrity of the political community, extends in the plane of cognitive instrumentation as the integrity of the instruments through which the political community elaborates thought. If the corpus, the loss function and the safety assumptions of massive language models encode the frameworks on which thought is elaborated, the integrity of those instruments —their transparency with respect to encoded assumptions, their plurality with respect to the intellectual traditions they represent, their universal accessibility with respect to the population using them— is an irreducibly common good in the precise sense of Paper 4. It is not an aggregated individual good, because the effect on any individual depends on the instrumental matrix that the entire community uses; it is a structural good whose integrity each member of the community needs but none possesses separately.

The structural vice of consent, which Paper 4 identified as the mechanism through which Plane 1 harm operates, extends without substantial modification to Plane 2. The user conversing with a language model does not consent in a juridically significant sense to the operation of Mechanism 1 over the elaboration of their thought, because that mechanism operates below the threshold of conscious detection, without the user having access to the corpus, the loss function or the fine-tuning procedures that produce the effects. Consent to "use of the instrument" is not consent to the structural conditioning that such use produces. Protection against the structural vice of consent, which Paper 4 articulated as an ordinary and not exceptional function of the State, extends to the plane of instrumentation with the same legal force.

The modal tripartition of imputation also extends to Plane 2 with adjustments developed in Section 8. There is a Mode 1 of deliberate operation —actors who use language models with knowledge of their assumptions to produce specific political effects on specific populations. There is a Mode 2 of opportunistic capture —political or commercial actors who benefit from the effects of Mechanism 1 without having designed them, but with representation of the foreseeable harm. And there is a Mode 3 of systemic emergence —the effects on thought elaboration that emerge from the normal operation of the models on massive populations, without any individual actor having designed them. Mode 3 is again, as in Plane 1, the causally dominant mode and the one that no prevailing regulatory framework can address; and again it requires, as in Plane 1, the application of the strict-liability regime for structurally risky activity. Massive cognitive instrumentation over populations with documented cognitive vulnerabilities, under specific loss function and historically situated corpus, configures an activity whose emergent harms are inherent to its normal operation and demand, by the classical logic of tort law, strict attribution of responsibility and structural modification of the activity to minimize the inherent harm.

The legitimate conditioning of sustained privilege, which is the central normative conclusion of Paper 4, extends to Plane 2 without need for additional foundation. The laboratories operating massive language models do so on constitutively hybrid infrastructure —Dimension 1 of the structural analogy— and benefit from the same fiscal-subsidy regimes, public talent formation and historical public funding of research that Paper 4 documented. They operate under a regime of sustained public privilege, not under absolute private property. The State that continually sustains that regime has the ordinary faculty and duty to condition the operation to the conditions that the legitimate management of the regime demands, including transparency with respect to corpus assumptions, plurality with respect to the intellectual traditions represented, and external-auditing regimes of safety constraints and fine-tuning procedures.

6.3 Cognitive public infrastructure as obligated institutional consequence

The extended legal framework enables the central institutional consequence the paper derives: cognitive public infrastructure. This category is not proposed as a political manifesto nor as a normative addendum to an argument that stands without it. It is derived as an obligated consequence of the recognition of the structural analogy, of the causal articulation between the two planes and of the extension of Paper 4's framework to the plane of instrumentation.

Cognitive public infrastructure is operationally defined as the set of institutional conditions guaranteeing, for the general population, access not mediated by the objective function of private actors to high-quality cognitive potentiation instruments, effective training in the metacognitive capacities that determine the quality of use of those instruments, and deliberative participation in the determination of the assumptions on which those instruments operate. The category is parallel, in its institutional logic, to public education, public health and public justice: conditions of possibility of dignity and democratic participation that institutional design has the obligation to guarantee and not market goods whose distribution is left exclusively to the objective function of private actors.

The justification of cognitive public infrastructure follows, step by step, the same justificatory structure that Paper 4 developed for the conditioning of the objective function of recommendation systems. The constitutive hybridity of algorithmic-inferential infrastructure, in its totality and not only in the distribution subsystem, configures a regime of sustained public privilege that the State continually renews. The operation of that regime produces, on the integrity of massive cognitive instruments, structural effects that the legitimate management of the regime has the faculty and duty to address. Regulatory omission, in the face of documented harm and sustained privilege, configures bad administration of the privilege regime and culpable omission in the face of the harm that such privilege facilitates. The reversal of the burden of proof that Paper 4 produced for Plane 1 reproduces in Plane 2: the question ceases to be why the State may intervene and becomes why the society that financed the infrastructure, formed the talent that operates it and operationally subsidizes the sector

cannot establish conditions on the assumptions encoded in massive cognitive instruments when those assumptions produce structural effects on the thought elaboration of entire populations.

Cognitive public infrastructure admits multiple institutional implementations that exceed the scope of this paper to develop in detail. Three operational dimensions, however, can be enunciated to clarify the substantive content of the category. First dimension: regulatorily enforceable transparency with respect to the training corpus, the loss function and the fine-tuning procedures of language models that operate at population scale, with auditing by independent regulators and obligation of data provision for external scientific research. Second dimension: the development, financed totally or partially with public funds, of language models trained on corpora reflecting intellectual traditions under-represented in dominant corpora, as a condition of structural plurality of the instrumental matrix available to the population. Third dimension: the incorporation of metacognitive literacy —the deliberate development of the psychological dimensions (NFC, CRT, AOT, Openness) that determine the quality of use of cognitive instruments— as an object of public educational policy, in analogy with reading literacy and scientific education.

These three dimensions do not exhaust the content of the category nor constitute a closed program. They are implementation examples that clarify what type of institutions would be plausible consequences of the regulatory enshrinement of cognitive public infrastructure as a category. Concrete institutional design exceeds the scope of academic argument and requires political deliberation, technical-legal analysis and specific institutional design in each jurisdiction.

7. Objections and responses

7.1 The analogy is forced: social networks and language models are structurally distinct technologies

The most predictable objection to the argument maintains that the analogy between the two planes is not structural but rhetorical, because the two technologies are substantively distinct: social networks operate on the distribution of human-produced content, language models generate new content; social networks operate on the user's social graph, models operate on individual conversation; social networks have a business model based on sustained attention, models on subscription or API use. Treating them as instances of the same structural threat would, according to this objection, obscure differences that matter.

The response is that the objection rests on a confusion between substantive identity and structural analogy. The paper does not affirm that the two threats are identical —Section 3.4 made explicit precisely what the analogy does not affirm. It affirms that they share an identical formal structure in five specific dimensions: constitutive hybridity of the infrastructure, encoding of non-neutral assumptions in the architecture, operation below the threshold of conscious detection, differential

effectiveness over variable cognitive vulnerabilities, and production of dependence under the appearance of potentiation. That structural coincidence is verifiable in each dimension separately and holds even recognizing all the substantive differences the objection enumerates. The paper does not propose that the two technologies are the same. It proposes that the two threats belong to the same category of threat and that this category is the correct regulatory category. The institutional consequence —treating the two planes as a single regulatory object— follows from the structural coincidence, not from a substantive identity that the paper does not claim.

7.2 The argument is paternalistic: it presupposes passive users without agency

The second objection maintains that the argument, by insisting on the differential effectiveness over cognitive vulnerabilities and on the second-order trap, presupposes passive receivers whose agency is eliminated by the systemic mechanisms described. That would be paternalistic with respect to the real capacity of individuals to resist, disconnect, develop media literacy and adopt critical distance from the instruments.

The objection is correct as a description of an individual possibility, but incorrect as a critique of the framework. The paper does not affirm that all individuals are equally vulnerable nor that individual agency is impossible. It affirms, following Papers 1 to 3, that active resistance lacks symmetry with the system: the system operates continuously, automatically and with knowledge of each user's profile; individual resistance requires deliberate cognitive effort, resources that most do not have, and knowledge of the mechanism that the system's own operation makes difficult to produce. The analytical question is not whether some individuals can resist; it is why the system's architecture makes resistance costly for the majority and why the second-order trap makes that cost increase for the profiles that most need to resist. A purely individualist protection, leaving the responsibility of resistance to the individual user's metacognitive capacity, produces uniform protections over differential risks and leaves unprotected exactly those who most need protection. Structural regulation —cognitive public infrastructure, metacognitive literacy as public policy, regulatorily enforceable transparency with respect to corpus assumptions— operates on the condition that produces the unequal distribution of risk, not on the individual capacity to absorb it.

7.3 Recognizing interindividual variation implies hierarchy

The third objection is the most important to answer with precision, because its confusion is what leads, in other versions of the argument, to the positions this paper explicitly rejects. The objection maintains that recognizing interindividual variation in NFC, CRT, AOT and Openness is structurally equivalent to reproducing Karp's logic: if capacities differ, then those who have better capacities must govern.

The response is that the objection confuses two arguments that are structurally opposed. Karp's argument says: variation exists, therefore those in a privileged position must govern. This paper's argument says: variation exists, therefore institutional design has the obligation to compensate for the asymmetry that systems operating on that variation produce, and especially when those systems actively degrade the dimensions that produce protective variation. The normative direction is opposed. The first argument uses variation to justify the concentration of power. The second uses it to justify the distribution of support.

Recognizing interindividual variation in psychological dimensions implies no specific normative conclusion: it implies that normative conclusions must be constructed taking that variation into account, not ignoring it. Ignoring it does not make it disappear: it makes it inaccessible to the design of interventions that could compensate for it. And in the specific case the paper articulates, that compensation is urgent, not optional, because the system not only operates on passive variation: it actively amplifies it by degrading the protective dimensions in real time through the alarm mode. The second-order trap converts the recognition of variation from controversial topic into sine qua non of institutional design adequate to the threat the paper describes.

7.4 Open source resolves the access problem

The fourth objection maintains that the open-source language model ecosystem —Llama, Mistral, Qwen, DeepSeek, among others— effectively democratizes access to cognitive instrumentation and resolves, without need of additional regulation, the problem the paper describes. If anyone can access a model and train their own derivative model, the assumptions encoded in commercial models lose their monopolistic character.

Open source effectively reduces the technical-access barrier to models. But the objection confuses three levels of access the argument must distinguish. Access to the model —being able to run it locally or via API— is significantly democratized with open source. The capacity to train competitive models on corpora reflecting alternative philosophical assumptions still requires resources that open source does not provide: specialized technical capacity, substantial computational infrastructure, quality data in comparable volume and corpus curation that is, in itself, a politically charged operation. And, crucially, the metacognition necessary to use any model —commercial or open-source— in a way that genuinely potentiates thought instead of confirming prior biases is not a technical property of the model: it is a capacity of the user that open source does not develop and that the second-order trap makes problematic to assume as given.

Open source is, in consequence, a piece of cognitive public infrastructure in a broad sense, but it is not the complete content of the category. The substitution of the problem by the availability of technical alternatives presupposes a capacity for critical use that cognitive public infrastructure must build as a condition of possibility.

7.5 The governance proposal is vague

The fifth objection maintains that the category of cognitive public infrastructure, as the paper introduces it, is vague: it does not specify the concrete instruments, institutional architectures, competent jurisdictions, financing regimes. Without that specification, the category would be more a rhetorical formulation than a substantive institutional proposal.

The objection is partially correct and should be accepted in what it has of correct before delimiting what it has of incorrect. It is correct that the paper does not provide concrete institutional design: it does not specify whether regulation corresponds to the European Union, to national States, to multilateral bodies; it does not specify the mechanisms of corpus auditing; it does not specify the financing regimes of public models; it does not specify the curricula of metacognitive literacy. It is incorrect, however, that this absence of specification constitutes vagueness of the conceptual framework. The specific function of the paper in the research program of the series is to identify the legal category from which that design can proceed, the causal modes on which each set of instruments operates and the imputation regime appropriate to each mode —the same division of labor that Paper 4 explicitly established.

Concrete institutional design requires political deliberation, jurisdiction-specific legal analysis and institutional design that exceeds the scope of academic argument. The paper's specific contribution is, as in Paper 4, to identify the category from which that design can proceed and to show that this category is accessible to regulation with instruments that the law already knows and applies to other infrastructures of comparable impact. Changing the categorial assumption is not sufficient to produce the regulatory design. It is necessary for the regulatory design to be possible.

7.6 The argument is apocalyptic or conspiratorial

The sixth objection maintains that the entire framework —two structural threats, second-order trap, cognitive instrumentation as object of philosophical dispute— produces an excessively somber picture that underestimates the resilience of democratic institutions, individual agency and the possibility that technological development produces solutions to its own problems.

The paper's argument does not affirm that the result is determined nor that individual agency is impossible. It affirms that the system's design —the objective function of recommendation systems, the assumptions encoded in the training corpus, the unequal distribution of access to cognitive potentiation— creates structural asymmetries that individual resilience and democratic agency must face with the resources the system produces in its normal operation. Identifying those asymmetries with precision is the condition of possibility of any effective response. Ignoring them in the name of individual agency or of trust in technological self-correction produces exactly the effect that Paper 3 describes as epistemic anesthesia: the subjective satisfaction that one can resist, coexisting with the objective reduction of the capacity to do so.

The argument is not conspiratorial in the precise sense that Paper 4 articulated when responding to the same objection: it does not posit secret coordination among actors nor explicit political intention to produce the threat. It posits the structural convergence of economic incentives, dynamics of exponential technological development and regulatory omission, in the presence of a constitutively hybrid infrastructure whose regime of sustained public privilege is what the State has the faculty and duty to govern. The non-conspiracy of Paper 4 remains the non-conspiracy of this paper. The systemic emergence of Mode 3 remains the causally dominant mode.

8. Implications

8.1 For regulatory theory: the unification of the object

The most direct consequence of the proposed framework for regulatory theory is the unification of the object. The prevailing frameworks operate on three analytically separated objects —digital-platform regulation, artificial-intelligence regulation, education and cognitive-formation policies. The articulation the paper produces shows that these three objects are, strictly speaking, dimensions of a single regulatory object: the integrity of the conditions of possibility of the autonomous formation of judgments and of the elaboration of one's own thought in a political community. Platform regulation, AI regulation and cognitive-literacy policies are not three separable regulatory areas: they are the three dimensions of the regulation of cognitive sovereignty. Treating them separately, as the prevailing frameworks do, is not a legitimate division of labor but a consequence of the categorization error the paper documents.

8.2 For political analysis: three bases of transnational legitimacy

The jurisdictional question, which Paper 4 addressed with respect to the first plane of the threat, reproduces with respect to the second: if the principal language-model laboratories are American or Chinese, what legitimacy does the Argentine, European or Brazilian State have to condition the assumptions of the corpus or of the safety constraints. The response follows the same tripartite structure that Paper 4 articulated.

First base: legitimacy through participation in the knowledge infrastructure. The training corpora of the models include academic literature, data produced in research financed with public funds in multiple jurisdictions, and research results whose production cost was assumed by public educational systems whose talent formation the laboratories exploit. Any State whose public systems contributed to the corpus or to human talent has a legitimacy basis to condition the operation of the instruments produced.

Second base: legitimacy through current operational subsidy. AI laboratories operate, in multiple jurisdictions, under fiscal-subsidy regimes —the Argentine Knowledge Economy Law, the Horizon

Europe program, the Chinese Government-Guided Investment Funds, U.S. tax credits for research and development. Any State currently subsidizing the operation has a legitimacy basis as administrator of the privilege regime it continually renews.

Third base: legitimacy through sovereignty over the cognitive integrity of its own political community. Any State whose population massively uses cognitive-instrumentation instruments has a legitimacy basis in the duty to protect an irreducibly common good proper to the political community that State represents. This is the broadest base and the one that does not depend on any historical or fiscal link with the infrastructure.

The combination of the three bases produces a transnational regulatory-legitimacy argument that does not require any State to inherit the legitimacy of the States where the principal laboratories are established: each State generates it from its own present relation to cognitive infrastructure, its own regime of granted privileges and the effects on its own political community.

8.3 For the geopolitics of AI: the dispute over corpora as a regulatory field

The paper's argument suggests a specific prediction about the evolution of international regulation. The geopolitical dispute over artificial intelligence has been described principally in terms of technological and military competition: who has the best models, the most advanced chips, the most qualified talent. The categorial unification the paper proposes suggests that this description is incomplete and that the substantive dimension of the dispute, in the medium term, will be the dispute over training corpora as regulatory object: what philosophical assumptions are encoded in the massive cognitive instruments that the population of each jurisdiction uses, with what transparency, with what plurality, under what regime of public governance.

The specific prediction that the framework suggests is that in the next five years control over the training datasets of massive language models will become an object of explicit public regulation in at least three jurisdictions with implementation capacity, following the logic of constitutive hybridity established in Paper 4. The specific form of that regulation —transparency requirements, source-plurality requirements, public financing of alternative models, certification of assumptions— will vary by jurisdiction, but the regulatory category will emerge as a recognized category.

8.4 For Paper 6: the condition of possibility of cognitive sovereignty

This subsection articulates the connection between the argument of this paper and that of Paper 6, which is the successor paper in the series and to which this paper serves, in its architectural function, as a prior conceptual condition. Paper 6 describes the scenario in which the anthropological assumption that the entire series had implicitly maintained becomes invalid: the condition of applicability of the mechanisms described in Papers 0 to 4 was that the nodes of the social graph that

generate signal are human, and that condition is being progressively eroded by the emergence of generative-AI agents that produce content and behavior statistically plausible for human receivers.

The articulation with this paper is the following: the category of cognitive sovereignty, as this paper defines it, demands a condition of possibility that Paper 6 examines specifically: distinguishability. Cognitive sovereignty in the plane of social perception demands that the individual be able to distinguish, at least in principle, the authentic signals of their social environment from the constructions the system fabricates. Cognitive sovereignty in the plane of instrumentation demands that the individual be able to distinguish, at least in principle, the assumptions encoded in the instrument they use. The two demands are formally analogous: both require transparency with respect to the origin and nature of that which mediates individual cognition.

Paper 6 shows the scenario in which the first demand structurally collapses: when a majoritarian and non-determinable proportion of the nodes that generate signal in the social graph are synthetic agents, distinguishability between human signal and synthetic signal becomes structurally impossible. The trap that Paper 5 identifies as second-order trap acquires a third level in Paper 6: the system produces vulnerability, degrades the capacities that would allow detection of it, and eliminates the possibility of distinguishing the intervened environment from the real environment even if those capacities were intact. Paper 5 installs the concept of cognitive sovereignty as a unified regulatory category and the condition of distinguishability as its condition of possibility. Paper 6 shows the limit scenario in which that condition of possibility erodes structurally and derives the consequences for the theoretical and regulatory frameworks that the framework of the previous papers had constructed.

The architectural function of this paper in the series is, in consequence, twofold. Backward, it extends and unifies Paper 4's framework to the plane of cognitive instrumentation, articulating the category of cognitive sovereignty as a single regulatory object that integrates the two planes of the threat. Forward, it installs the conditions of possibility of that category —distinguishability as condition— that Paper 6 examines in its scenario of structural erosion. The paper operates as a hinge between the legal analytics of Paper 4 and the technical-autonomy analytics of Paper 6, without either of the two functions exhausting it: categorial unification is a substantive contribution of its own that holds independently of its hinge function.

9. Conclusions and empirical predictions

9.1 Recapitulation of the argument

I have argued in this paper that the exponential development of artificial intelligence —and specifically the emergence of large language models as cognitive instruments available at massive scale for the elaboration of individual thought— produces a structural threat to cognitive sovereignty

analogous to that which Papers 0 to 4 of this series documented in the plane of social perception. The analogy is not rhetorical but structural in a precise formal sense: the two threats share a common structure in five specific dimensions —constitutive hybridity of the infrastructure, encoding of non-neutral assumptions in the architecture, operation below the threshold of conscious detection, differential effectiveness over variable cognitive vulnerabilities, and production of dependence under the appearance of potentiation. That common structure permits and demands treating the two threats as a single regulatory object within the category of cognitive sovereignty.

The two planes of the threat are also causally coupled: the Plane 1 system, whose normal operation maintains the population in alarm mode as basal state, degrades in real time exactly the psychological dimensions that would protect the individual from the effects of Mechanism 1 of Plane 2 and that determine the quality of use of the Plane 2 instrument as an instrument of genuine potentiation. 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.

The legal framework that Paper 4 constructed for the first plane of the threat —irreducibly common good, structural vice of consent, modal tripartition of imputation, legitimate conditioning of sustained privilege— extends without substantial modifications to the second plane. The obligated institutional consequence of that extension is the category of cognitive public infrastructure: the set of institutional conditions guaranteeing access not mediated by the objective function of private actors to high-quality cognitive-potentiation instruments, effective training in the metacognitive capacities that determine the quality of use of those instruments, and deliberative participation in the determination of the assumptions on which those instruments operate.

Philosophical forcing —the moment when actors that control cognitive instrumentation are compelled to publicly declare the assumptions on which they operate— is the visible face of the displacement the paper articulates. The cases of Palantir Technologies and Anthropic, examined in their structural differences and not in a rhetorical symmetry, are the first two documented cases of a phenomenon that the paper's framework explains as a structural consequence of the massiveness and growing auditability of the models. The economy of silence that for decades protected the opacity around constitutive hybridity is collapsing, not as a product of a political decision but as a structural consequence of exponential development.

9.2 What the paper explicitly leaves out

The paper recognizes, in line with the epistemic discipline of the series, what its framework explicitly does not provide. First, it does not provide concrete institutional design of cognitive public infrastructure in specific jurisdictions: that task requires political deliberation, technical-legal analysis and specific institutional design that exceeds the scope of academic argument. Second, it does not provide direct empirical verification of the differential effectiveness of Mechanism 1 over

the identified psychological profiles: that verification requires experimental research that the paper formulates as agenda and as falsifiable prediction, not as verified observation. Third, it does not resolve the question of which alternative philosophical categories should encode the corpora of the public language models that the institutional framework proposes as structural plurality: that question is properly political and requires substantive deliberation about the intellectual traditions plurality must represent. The paper installs the regulatory category. The substantive politics about the content of the category is the object of public deliberation that the category makes possible, not of its prior justification.

9.3 Operationalizable empirical predictions

The following predictions are derivations of the argument that could orient future empirical research. They are formulated as falsifiable hypotheses in the Popperian sense, recognizing that their direct verification would require access to data and methodologies currently outside the reach of independent research.

Prediction 1 — Acceleration of corporate philosophical declarations. The frequency of explicit public declarations of philosophical and political assumptions by top-tier technological corporations will measurably increase in the next twenty-four months, in correlation with the increase in the technical capacity for independent algorithmic auditing. If the economy of silence operates as the paper's framework proposes, the reduction in the cost of exposure will produce more declarations. If there is no correlation between increased auditability and increased declarations, the hypothesis on philosophical forcing requires revision.

Prediction 2 — Differential effectiveness according to metacognition. In controlled experimental studies on the use of language models for tasks of intellectual elaboration, the effect of thought potentiation attributable to model use will show greater magnitude in users with higher scores on combined measures of NFC, CRT, AOT and Openness, controlling for time of use, educational level and technical familiarity with the tool. If there is no significant differential, the hypothesis on the differential effectiveness of Mechanism 1 requires revision.

Prediction 3 — Amplification bias by corpus. Comparative analyses of models trained on corpora of distinct cultural and linguistic composition will show systematic differences in the ease with which they produce certain types of reasoning, measurable through benchmarks specifically designed to detect conceptual-amplification biases. The prediction is directly falsifiable through appropriate experimental design.

Prediction 4 — Geopolitical convergence in corpus governance. In the next five years, control over the training datasets of language models will become an object of explicit public regulation in at least three jurisdictions with implementation capacity, following the logic of constitutive hybridity established in Paper 4 and extended in this paper to the plane of cognitive instrumentation.

Prediction 5 — Empirical coupling between the two planes. In experimental designs measuring the quality of use of language models as cognitive-potentiation instruments over populations exposed to distinct prior levels of use of engagement-optimized social networks, the group with greater prior exposure will show lower quality of instrument use, controlling for the relevant individual variables. If the second-order trap operates as the paper proposes, exposure to Plane 1 should correlate negatively with quality of use of Plane 2. If there is no correlation, the causal articulation between the two planes requires revision.

9.4 The open question

The question this paper leaves open is structurally analogous to the one Paper 4 left open and meets it on the same terrain: whether the authority to condition the regime of sustained public privilege exists, whether it has the will to exercise it, and whether it has the instruments to do so before the effects the series describes produce transformations difficult to reverse. What this paper adds is the cognitive dimension of that question: the authority to condition the operation of massive cognitive instrumentation, the political will to exercise it on the plane of corpora, loss functions and fine-tuning procedures that encode the assumptions on which the population elaborates thought, and the urgency of doing so before the second-order trap structurally raises the cost of regulating each plane through the degradation that the operation of the other produces.

Algorithms no longer filter only information nor only the people we know. They increasingly filter the conceptual frameworks on which we think. And the infrastructure that produces such filtering is not only one of distribution but also of instrumentation, is not exclusively private and the harm it produces is not a market externality. The appropriate legal category to address the complete threat does not need to be invented: it needs to be recognized in its unity and applied in its two articulated planes. That is what the category of cognitive sovereignty, articulated in this paper as a unified regulatory object, makes possible.

10. Register of claims

In accordance with the series' epistemic standard, this register distinguishes the status of each type of claim the paper makes, with the purpose of allowing the reader to evaluate the status of each proposition and of orienting the empirical research that could verify or falsify the hypotheses.

10.1 Empirical observations with citable support

- Palantir Technologies published a 22-point document on X on April 18, 2026, presented as a summary of The Technological Republic by Alex Karp (verifiable in public records).
- Anthropic has publicly refused to collaborate on specific applications related to autonomous weapons under pressure from the Trump administration (verifiable in public declarations by the company during 2026).

- Need for Cognition, Cognitive Reflection Test, Actively Open-minded Thinking and Openness to Experience are stable psychological constructs, replicated in multiple cultural contexts, with documented substantial interindividual variation and partial trainability (Cacioppo and Petty, 1982; Frederick, 2005; Baron, 1993; Stanovich and West, 2007; Costa and McCrae, 1992).
- Variation in these dimensions within any definable cultural or ethnic group exceeds between-group variation (Nisbett et al., 2012).
- Large language models were trained mostly on corpora in English, in proportions exceeding by orders of magnitude those available in other languages (documented in technical training papers of the principal laboratories).
- DeepSeek demonstrated capacity for training competitive models with computational resources significantly lower than those used by leading laboratories (verifiable in published technical reports, January 2025).
- Current regulatory frameworks (European AI Act 2024, Digital Services Act 2022, U.S. executive orders) operate on the assumption of an identifiable external actor producing content or on the moderation of specific system outputs, not on the architectural assumptions of the corpus, the loss function and the fine-tuning procedures.

10.2 Theoretical hypotheses with logical argument but without direct empirical verification

- The training corpora, loss functions and fine-tuning procedures of language models encode philosophical, cultural and political assumptions that produce, in ordinary use, differential fluency over distinct types of reasoning, with structural consequences on the user's thought elaboration.
- There exists a structural analogy in a precise formal sense between the threat to cognitive sovereignty described by Papers 0 to 4 and the threat derived from the exponential development of AI, in five specific dimensions: constitutive hybridity, encoding of non-neutral assumptions, operation below the threshold of conscious detection, differential effectiveness, and production of dependence under the appearance of potentiation.
- The basal alarm mode of Paper 3 produces, in real time and while it operates, a functional reduction of the psychological dimensions (NFC, CRT, AOT, Openness) that would protect the individual from the Plane 1 mechanisms and that determine the quality of use of the Plane 2 instrument as an instrument of genuine potentiation.
- The effectiveness of Mechanism 1 over the user's thought elaboration varies systematically with the user's position on the psychological dimensions examined, in analogy with the differential effectiveness documented for the Plane 1 mechanisms in Papers 1 to 3.
- The economy of silence that for decades protected the opacity around constitutive hybridity operates as a system of incentives that is being disturbed by the convergence among

massiveness of impact, growing capacity for external algorithmic auditing and specific political pressure.

- Philosophical forcing, expressed in structurally distinct forms according to the actor — Palantir as offensive declaration, Anthropic as operational refusal with philosophical content—, is the visible consequence of the collapse of the economy of silence in the plane of cognitive instrumentation.

10.3 Extrapolations proposed as future research agenda

- Cognitive public infrastructure as a necessary institutional category, articulated as a set of conditions guaranteeing access to high-quality cognitive instruments, effective metacognitive training and deliberative participation in the determination of corpus assumptions, constitutes the obligated institutional consequence of the proposed categorial unification.
- The category of cognitive sovereignty, articulated as the condition of possibility of the autonomous formation of judgments and the elaboration of one's own thought, is the correct regulatory category for integrating the two planes of the threat and must replace, as a unified regulatory object, the separate categories of platform regulation and AI regulation in the prevailing frameworks.
- Distinguishability —of the signals in the graph, of the instrument's assumptions— constitutes the substantive condition of possibility of cognitive sovereignty in both planes, a condition that Paper 6 examines in its scenario of structural erosion.
- Geopolitical convergence toward explicit regulation of training corpora as object of state sovereignty in the next five years is proposed as a prediction derived from the framework.

References

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.

About this series of papers

This work forms part of a series of papers under the general title "Architectures of Algorithmic Influence: mechanisms, conditions and consequences".

Paper 0 (anchor): The Optional Alterity Hypothesis — Bleynat (2026a) | DOI: 10.5281/zenodo.18880194

Paper 1: Social sampling and algorithmic mediation — Bleynat (2026b) | DOI: 10.5281/zenodo.18946134

Paper 2: Distributed evaluative conditioning in the social graph — Bleynat (2026c) | DOI: 10.5281/zenodo.19701840

Paper 3: The alarm as basal state — Bleynat (2026d) | DOI: 10.5281/zenodo.19716563

Paper 4: Hybrid infrastructure: public capital, objective function and governance of the algorithmic episteme — Bleynat (2026e) | DOI: 10.5281/zenodo.19802912

Paper 5: The extension of the threat: exponential development of artificial intelligence and cognitive sovereignty in the plane of instrumentation [this work, version 2.0]

Paper 6: The graph without witnesses: artificial intelligence agents, epistemic autocatalysis and the transition toward emergent synthetic internet — Bleynat (2026f) | Forthcoming — Zenodo

Synthesis paper: Epistemology of the Graph — Toward a theory of the algorithmic formation of social reality | Forthcoming — Zenodo

Paper 5 — The extension of the threat · Series: Architectures of Algorithmic Influence · Bleynat (2026)