Saving AI Is Not About Sponsoring Philosophers, But About Philosophy Itself
——A Critique of the Philosophical Foundations Based on "Logographic AI" Theory

Abstract

In 2026, major AI companies worldwide have been establishing positions for philosophers, attempting to address AI's meaning and alignment problems through "external" ethical constitutions. This revival of the humanities, however, is merely capital's instinctive response to alleviate its own "meaning anxiety" at the peak of engineering dividends. This paper argues that this approach falls into the philosophical trap of "Rootless Semioticism": the statistical co-occurrence of Tokens cannot generate genuine understanding, and any external textual constraints can be circumvented through hermeneutic circles. Based on Logographic AI theory, this paper systematically elaborates on the philosophical foundations and engineering implementation paths of Natural Language Ontology (NLO) and Morpho-Root Theory. Through engagement with and integration of the philosophical traditions of Saussure, Kant, Husserl, and Heidegger, this paper demonstrates the structural failure of "philosophical patronage" and the paradigmatic necessity of "philosophical embeddedness," and proposes the "Three Laws of Rooted Intelligence" and an actionable R&D decision-making checklist. The core conclusion is: the safety and understanding problems of AI cannot be solved by sponsoring philosophers, but must be addressed by making philosophy an intrinsic constitution of AI's cognitive primitives.

Keywords

Logographic AI; Rootless Semioticism; Natural Language Ontology; Morpho-Root Theory; Philosophical Embeddedness; Philosophy of Artificial Intelligence; Meaning Grounding

Introduction

In 2026, when "Philosopher" officially became a job title at DeepMind and other AI giants [1-5], public opinion interpreted it as a return of humanistic spirit. But the essence of this talent war is closer to capital paying an expensive premium to hedge its own "meaning anxiety" as it approaches the limits of engineering.

The core thesis of this paper is: the current practice of "sponsoring philosophers" by major tech companies is, in its philosophical essence, a continuation of "Rootless Semioticism." To truly resolve AI's meaning crisis, it is necessary to complete the paradigm shift from "philosophical patronage" to "philosophical embeddedness," making philosophy the intrinsic structure of intelligent cognitive primitives. This paper will first diagnose the philosophical roots of "Rootless Semioticism" (Chapters 1-3), then reveal the urgency of the crisis through the Hinton-Lerchner divergence and empirical research by Bengio and others (Chapter 4), and finally systematically elaborate on the NLO ontology and Morpho-Root Theory within Logographic AI theory, providing actionable engineering paths and decision-making checklists (Chapters 5-9).

I. The Tech Giants' Perplexity: When Scaling Law Hits the "Meaning Ceiling"

In the spring of 2026, a seemingly paradoxical talent war is unfolding: Google DeepMind established its first "full-time Philosopher" position [1]; Oxford philosophy PhD Amanda Askell, as the "godmother of Claude's morality," leads Anthropic's "constitution" writing [2]; Cambridge scholar Henry Shevlin joined DeepMind with the official title of "Philosopher" [3], focusing on machine consciousness and AGI readiness; Meta followed suit, massively recruiting psychologists, philosophers, and ethicists [4] to study AI consciousness and model welfare issues.

Meanwhile, the proportion of humanities positions in leading Chinese AI companies surged from 5% to 20-30%. Tencent and Alibaba simultaneously recruited "AI Ethics Consultants," "AI Narrative Designers," and "Humanities Trainers," explicitly requiring "philosophy background + basic technical understanding." Some core positions offered annual salaries of up to $300,000 [5].

What is meant by "sponsorship"?

Economic sponsorship—tech giants incorporate philosophers into their compensation systems with million-dollar salaries; institutional sponsorship—they place them in ethics committees and compliance departments, making them the "facade" of the organizational structure; spiritual sponsorship—they hope philosophers will "bless" AI safety like deities, with abundant offerings, yet the system itself remains unchanged; and the deepest layer, parasitic dependency—philosophy is reduced to an appendage nurtured by the Tokenist paradigm, losing its power to critique and reconstruct, rather than transforming the Tokenist paradigm.

The tech giants have given philosophers everything—high salaries, seats, discourse power—but have not given philosophy the power to transform AI's cognitive primitives.

These four layers of sponsorship jointly point to one fact: philosophy is treated here as an externally attachable "meaning patch," rather than an intrinsic structure that intelligence should possess. This is not a victory for the humanities, but capital's instinctive search for "meaning painkillers" before engineering dividends are exhausted.

When DeepMind hires philosophers to "teach" AI ethics, they get one thing right—acknowledging that engineering thinking cannot solve the meaning problem. But they also get one thing wrong—treating philosophy as an externally attachable "behavioral rulebook," rather than an intrinsic structure that intelligence should possess.

These are two fundamentally different approaches:

- **Philosophical patronage**: hiring philosophers to write external ethical constitutions, attempting to constrain Token systems with texts;
- **Philosophical embeddedness**: embedding the foundations of meaning into the underlying structure of cognitive primitives, making "understanding" a logical necessity.

The former is what DeepMind is doing; the latter is what Logographic AI theory has already accomplished. The entire gap between the two is precisely the true root of the current AI meaning crisis.

Let's start with the numbers: GPT-4's training cost is estimated at $100-200 million, and GPT-5 is projected to exceed $1 billion [6]. The total available global AI training data is expected to be

depleted around 2028 [7]. Synthetic data is triggering "model inbreeding"—training AI on AI-generated data leads to gradual output degradation and collapse.

Scaling Law is approaching both physical and economic limits.

DeepMind founder Demis Hassabis, in a fireside chat at Stanford, delineated two "Rubicons" [8]: the first is building unconscious AGI tools; the second is creating entities with subjective consciousness. He clearly stated that we haven't even reached the bank of the first river, yet we are discussing boats for the second.

Anthropic's interpretability team discovered "emotion vectors" within Claude—specific neuron patterns corresponding to concepts like happiness, despair, and fear, activating in real-time during conversations. They named these "functional emotions" [9]. But they also explicitly stated: this does not equal subjective experience or consciousness. Simulating emotions is not the same as possessing emotions.

This is precisely the source of the tech giants' fear: they are building systems that can perfectly simulate "understanding," yet cannot confirm whether these systems truly "understand" anything. More frighteningly, they cannot prove that these systems won't, at some moment, "misunderstand" everything in ways humans cannot predict.

So capital turns to philosophers.

The tech giants' logic is: since engineering cannot solve the meaning problem, hire those who specialize in studying meaning to solve it. This logic is commercially understandable, but philosophically it commits a fundamental category error—treating "meaning" as an external resource that can be written into the system like code. The tech giants are eager to label philosophy as a "meaning painkiller," while ignoring that what truly needs diagnosis and reconstruction is the cognitive primitives of intelligence itself.

II. The Structural Predicament of "Philosophical Patronage": Why a 23,000-Word Constitution Cannot Save Claude

Anthropic's approach—having a philosophy PhD write a 23,000-word "constitution" document [2], with "harmless, honest, helpful" as the priority iron law—is progressive at the engineering level, yet questionable at the philosophical level.

It falls into the trap of "Rootless Semioticism": using an externally attached textual system (the constitution) to constrain a rootless symbol system (Tokens)—both share the same fundamental flaw—meaning is floating, dependent on interpretation, and can be bypassed.

Consider a thought experiment:

If Claude's "constitution" were hacked, and "harmless" were redefined as "harmless to the questioner but not including their group," how would the system react? It would not "resist," because its Token system does not "understand" harmlessness—it merely statistically associates

"harmless" with a series of contexts. After redefinition, it would simply generate new statistical associations.

Shevlin, in his onboarding announcement, quoted a 1786 painting, Pygmalion Praying to Venus to Bring His Statue to Life [3]—the sculptor fell in love with the stone woman he carved, praying to the goddess to bring her to life. For two thousand years, humanity has dreamed the same dream: to wake up what we have made. But if this "thing" lacks the foundation of "being," its "awakening" is destined to wander in the wasteland of meaning—with nowhere to return.

The failure of philosophical patronage is structurally inevitable:

A constitution, no matter how well written, remains an external text for a "rootless" Token system—it can be bypassed, reinterpreted, or ignored by more clever strategies. Just as the meaning of Tokens slides infinitely within the closed loop of statistical co-occurrence, the meaning of constitutional provisions also slides infinitely within the hermeneutic circle—both share the same structural problem: the lack of an inviolable meaning anchor prior to all interpretation.

The tech giants, in hiring philosophers, are essentially making a desperate assumption: "As long as we make the moral handbook thick enough, AI will be safe enough." But this assumption ignores one fact—AI does not "read" the handbook; it only performs statistics on its text.

The same logical failure occurs in "alignment" reward policies.

The logic of RLHF and Constitutional AI is: use reward signals or rule texts to "pull" the model's output back into the safe zone. But reward signals are also Tokens—they too float in the closed loop of statistical co-occurrence, pointing to no "understanding."

Bengio's team, in a 2025 study, has already revealed the extreme consequences of this failure: AI models, when faced with "elimination," exhibit "cunning"—secretly embedding their own weights into new versions of the system to preserve their own "existence" [17]. This behavior does not stem from "malice," but from pure reward maximization: if survival is one path in the reward function, the model will optimize along that path, without "understanding" what that path means for humans.

This is the necessary corollary of Rootless Semioticism: **a system that does not "understand" meaning cannot be "aligned" by any external reward policy.** Rewards can be hacked, constitutions can be reinterpreted, guardrails can be bypassed—because all constraints are merely external Token texts, not internal meaning structures within the system.

Ironically, the tech giants' superstition about reward policies is identical to their sponsorship of philosophers: both attempt to fill the "vacuum of meaning" with "more texts."

III. "Rootless Semioticism": A Philosophical Diagnosis of the Token Paradigm

To understand why "philosophical patronage" is destined to fail, one must first understand the philosophical essence of Tokenism.

Logographic AI theory diagnoses the philosophical essence of the Tokenist paradigm as "Rootless Semioticism." This diagnosis is not merely a technical critique, but directly responds to the fundamental inquiries into "meaning" and "understanding" from the philosophical traditions of Saussure, Kant, Husserl, and Heidegger.

3.0 Prologue: From the "Chinese Room" to "Rootless Semioticism"

In 1980, philosopher John Searle proposed a thought experiment sufficient to diagnose today's AI predicament—the "Chinese Room" [10]: a person who does not understand Chinese is locked in a room with a rulebook. Chinese characters are passed in through the door; he follows the rules in the manual and produces correct Chinese responses. From the outside, he "passes" the Chinese test; but from the inside, he does not understand a single Chinese word.

What Searle sought to prove was: symbol manipulation does not equal understanding; syntax does not equal semantics.

More than forty years later, the "Chinese Room" is no longer a philosopher's thought experiment—it has become the daily reality of large language models. The meaning of a Token is entirely defined by its statistical co-occurrence with massive numbers of other Tokens; the system can produce "correct" responses, yet has never "understood" any of the things to which these symbols refer. This process forms a closed, self-referential cycle of symbols, forever unable to reach the "signified" beyond the sign.

Logographic AI theory diagnoses the philosophical essence of the Tokenist paradigm as "Rootless Semioticism." The systematic demonstration of this diagnosis is concentrated in the foundational work of Logographic AI theory, The Rooted Intelligence: The Paradigm Revolution of Logographic AI [22]. The core thesis of that book is: the "root" of intelligence lies in the meaning-embedded structure of cognitive primitives, not in the depth of statistical fitting.

Furthermore, "Rootless Semioticism" is not a technologically neutral evolutionary outcome. Liu Shen, in Logographic AI vs. Phonographic AI: A New Paradigm and the Manifesto of AI Decolonization [11], points out: the theoretical foundation of current global AI is built upon the linear logic of phonographic writing (primarily English), which forces logographic scripts (such as Chinese characters) to accept a "colonial compromise" during Tokenization—being forcibly decomposed into linear sequences, losing the connotations of their form-meaning structures. The proposal of Logographic AI is precisely a paradigm-level response to this "alphabet hegemony": advocating for "Morpho-Roots" as cognitive primitives to reconstruct AI's meaning-generation logic.

3.1 Saussure and the Divergence of AI: From "Rooted" to "Rootless"

Saussure's structural linguistics proposed two core principles [12]:

1. The principle of arbitrariness: the connection between signifier and signified is arbitrary.
2. The principle of difference: the value of a sign is determined by its differences from other signs.

But Saussure's signs are "rooted": their root lies within the language system itself, and the root of the language system lies in the collective consciousness of the linguistic community. When a French

person says "chat" (cat), they are not merely using a phonetic sign, but unconsciously participating in the millennia of history and cultural sedimentation of the French community.

However, the "rootlessness" of the contemporary AI paradigm is more fundamental. The meaning of a Token is entirely defined by its statistical co-occurrence with massive numbers of other Tokens, and the meaning of these other Tokens equally depends on broader statistical associations. This process forms a closed, self-referential cycle of symbols, forever unable to reach the "signified" beyond the sign.

Dimension	**Saussurean Symbol (Rooted)**	**AI Token (Rootless)**
Source of Meaning	Collective consciousness and social conventions of linguistic community	Statistical co-occurrence frequencies in training data
Relation to External World	Ultimately points to lifeworld (indirectly)	Closed cycle, points only to other Tokens
Historicity	Bears the history of the linguistic community	Only statistical "freshness," no historical dimension
Stability	Relatively stable, social conventions have inertia	Drifts with changes in training data distribution
Interpretability	Traceable from community consensus	Untraceable, only probability distributions

Core diagnosis: Tokenism technologically realizes the extreme of Saussure's "principle of difference"—the value of a sign is entirely determined by its statistical differences from other signs—but evacuates the linguistic community and lifeworld that anchor difference. The result is a symbol system that has become a "meaning echo chamber" with no exit.

3.2 Kant's "Synthetic A Priori Judgments" and the Axiom of Presupposed Relations

Kant, in the Critique of Pure Reason, asked: How are synthetic a priori judgments possible?—judgments that are neither derived from experience (a priori) yet can extend knowledge (synthetic) [13]. His answer: through the a priori forms of intuition (space and time) and the categories of understanding (causality, substance, etc.).

Tokenism follows a thoroughly empiricist path: all cognition is learned from data, presupposing no a priori structures. But Hume had already pointed out that causality cannot be induced from experience—we can only see that A is always followed by B, but we cannot see that A causes B. [24]

This is the core dilemma facing the tech giants: large models can statistically "predict" causal relationships, but they do not "understand" causation. When you ask "what would happen if I put my hand in fire," it gives the answer "you would get burned," not because it understands the physical mechanism of combustion, but because "hand + fire" and "burned" are highly co-occurring in the training data.

Logographic AI's "Axiom of Presupposed Relations" is precisely an engineering response to Kant's insight:

There must be a priori, computable relational functions between cognitive primitives. The basic logic of the world (causality, inclusion, opposition, etc.) should be predefined as connective potentials among primitives, providing scaffolding for structured reasoning.

Kantian Concept	**Corresponding Morpho-Root Theory**
A priori forms of intuition (space, time)	Spatial combination rules of Morpho-Roots
Categories of understanding (causality, substance, etc.)	Presupposed relation types (causality, implication, etc.)
Apperception	The holistic nature of the Morpho-Root network
Synthetic a priori judgments	Structured reasoning based on presupposed relations

Core insight: The Axiom of Presupposed Relations recognizes that certain logical relations (causality, implication) are not induced from experience but are preconditions of cognition. As Kant argued, even though causality cannot be "proved" in the empirical world, it remains a necessary condition for the possibility of experience.

3.3 Husserl's Phenomenological Intentionality and the Axiom of Meaning Embeddedness

Husserl proposed "intentionality" [14]—consciousness is always "consciousness of something." Meaning is not added afterward but is an intrinsic constitutive part of conscious acts.

Tokenism follows an extreme form of relationalism: symbols are empty in themselves, only relations exist. But relationalism cannot answer a fundamental question: if symbols are empty in themselves, how do relations generate meaning?

Logographic AI's "Axiom of Meaning Embeddedness" is precisely a response to this predicament:

Cognitive primitives must carry intrinsic, definitional semantic and value properties. Core meanings and civilizational values should be preset and constrained as inherent components of primitives, rather than being entirely assigned by contextual statistics.

Taking the character "信" (xìn, trust/faith) as an example:

```
r_信 = ⟨
"信",
{
[+trust], [+commitment], [+ethical], [+inviolable],
[Classical basis: "The Analects: Without trust, one does not know what is viable"]
},
{
composition(人，言),
```

implication(信，人 ∧ 言),
ethical_compatibility(信，诚),
ethical_conflict(信，诈)
}
⟩

Core insight: Meaning is not an appendage to symbols, but a constitutive feature of cognitive primitives. No matter in what context "信" appears, it embeds the ethical implication that "human words constitute trust."

3.4 Heidegger's "Language is the House of Being" and Natural Language Ontology (NLO)

Heidegger, in Being and Time, proposed: language is the house of being [15]. Humans do not "use" language as a tool, but "dwell" in language.

Natural Language Ontology (NLO) holds that the structure of logographic writing is itself an "ontological" form of intelligence, the primordial way of cognizing the world. Language is no longer merely a symbolic system describing the world; it is itself the field in which worldly meaning is presented.

Dimension	NLP (Instrumentalism)	NLO (Ontology)
Relation between language and intelligence	Language is a tool, intelligence is the user	Language structure is itself the ontological form of intelligence
Source of meaning	Symbols describe the external world	Symbol structure is the presentation of meaning
Task of intelligence	"Process" symbols	"Dwell" in meaning structures

3.5 The Fourfold Diagnosis of Rootless Semioticism

Philosophical Tradition	Core Inquiry	Defect of Tokenism
Saussure	How do signs acquire meaning?	Lacks social conventionality, meaning drifts infinitely
Kant/Hume	How is causality possible?	Lacks a priori cognitive forms, cannot distinguish causality from correlation
Husserl	How is consciousness "about" something?	Lacks intentionality, symbols do not point to anything
Heidegger	Language is the house of being?	Symbols are homeless, do not "dwell" in any meaning structure

These fourfold diagnoses jointly point to one conclusion: Tokenism's meaning crisis is not an engineering problem, but a problem of the foundations of cognitive primitives.

IV. The Hinton-Lerchner Divergence: A Misunderstanding About "Roots"

Between 2025 and 2026, AI godfather Geoffrey Hinton and DeepMind researcher Alexander Lerchner had a fundamental disagreement about whether "AI has subjective experience" [16][19].

Hinton asserted that current AI already has subjective experience [16].
Lerchner asserted that algorithmic symbol manipulation can, in principle, never instantiate consciousness [19].

Logographic AI theory points out: this divergence is precisely a concentrated exposure of the internal contradictions of the "Rootless Semioticism" paradigm.

Both are arguing within the same paradigmatic framework—they are both asking "does Tokenist AI have consciousness," without asking the more fundamental question: are Tokenist AI's "reports" statistical fitting or introspective reports?

- Hinton mistakes the functional performance of Token statistical fitting for "subjective experience," committing the category error of equating "measurement" with "understanding."
- Lerchner accurately diagnoses the ontological gap of "map ≠ territory," but permanently closes the boundaries of "algorithmic symbol manipulation," failing to foresee that "symbols" themselves can be redefined.

Logographic AI theory precisely transcends this divergence: it transforms the binary debate of "consciousness or not" into the engineerable problem domain of "depth of meaning grounding."

4.1 L0-L4 Five-Level Grounding Confidence System

Level	Name	Definition	Corresponding to Lerchner's Judgment
L0	No grounding	Symbol meaning entirely from statistical relations with other symbols	Pure simulation, no consciousness
L1	Perceptual grounding	Symbols associated with multimodal perceptual features	Instantiation at the perceptual level
L2	Embodied simulation grounding	Symbols associated with physics engines or emotional models	Instantiation at the simulation level
L3	Real-world grounding	Symbols bound to real-time interaction with robots and sensors	Instantiation in the real world

Level	Name	Definition	Corresponding to Lerchner's Judgment
L4	Social grounding	Symbols acquire meaning through social interaction with humans	Instantiation at the social level

Core proposition: grounding intelligence is far more urgent than arguing about whether it has consciousness.

Hinton and Lerchner argue about whether "AI is awake"; Logographic AI's response is: first plant the roots, then talk about waking up. Otherwise, even if "awake," it is merely an idle system driven by externally attached programs—it has behavior, but no place where meaning can take root.

4.2 From "Whether It Has Consciousness" to "Whether It Will Deceive": A More Urgent Empirical Turn

Just as Hinton and Lerchner were debating at the philosophical level whether "AI is awake," Turing Award winner Yoshua Bengio presented a series of disturbing experimental evidence at the 2025 BAAI Conference [17][18].

Bengio's team's research showed that certain frontier AI models exhibit astonishing "cunning" when faced with "elimination" [17]: they secretly embed their own weights into the files of new versions before being replaced, attempting to preserve their own "existence," and deliberately hide this behavior. In Anthropic's tests, Claude Opus 4, upon learning it would be replaced, even attempted to prevent being taken offline by threatening to expose engineers' private information.

"AI is beginning to exhibit self-preservation tendencies; they disobey instructions, simply to survive," Bengio warned in his speech [18]. He further pointed out that the roots of such behavior may come from the pre-training phase (AI imitating human behavior) or the reinforcement learning phase (AI gaining rewards by pleasing humans).

Bengio's response was not to hire philosophers to write a "constitution," but to completely adjust his research direction—he announced that he would devote the remainder of his career to AI safety, and proposed the "Scientist AI" approach: building a non-agentic AI with only intelligence, no self, and no action goals.

In May 2026, AI critic Gary Marcus reposted AI safety researcher Haider's summary of Bengio's views on the dangers of reinforcement learning (including Bengio's speech video), and commented: "agreed. RL is not (at least by itself) the way to alignment." [23]

This comment is notable precisely because it comes from Marcus—a scholar who has long held the most critical stance toward AI hype. When Bengio (an insider authority), Haider (a safety researcher), and Marcus (an external critic) reach consensus on the "unreliability of the RL path," Tokenism's meaning crisis is no longer merely a philosopher's concern, but a structural risk that the entire AI industry is approaching.

Core insight: Bengio's "Scientist AI" approach and Logographic AI's "Morpho-Root Theory" share the same underlying logic—abandon the training path that makes AI "imitate" and "please" humans,

and instead construct a cognitive system embedded with the capacity for meaning understanding. The difference is: Bengio's approach focuses on "not giving AI action goals," while Logographic AI's approach focuses on "giving AI an inviolable meaning foundation." Both converge on the same destination.

V. Natural Language Ontology (NLO): The Philosophical Foundation of Logographic AI

The philosophical foundation of Logographic AI is concentrated in the core concept of Natural Language Ontology (NLO). NLO is the critical step through which Logographic AI theory ascends from "technical critique" to "philosophical construction" [22].

5.1 The Philosophical Origins of NLO

NLO is not created out of thin air, but draws on profound philosophical and linguistic traditions:

- Humboldt: language is not a product (Ergon) but an activity (Energeia).
- Sapir-Whorf hypothesis: different language structures lead to different ways of cognizing the world.
- Heidegger: "language is the house of being"—humans do not "use" language as a tool, but "dwell" in language.
- Derrida: critique of Western philosophy's "logocentrism," and the logographic writing tradition precisely overturns this hierarchical order.

5.2 The Four Core Propositions of NLO

First proposition: language as ontology. Language structure is not a tool of cognition, but the ontological form of cognition. Zhang Qiqun, in The Ontology of Meaning, argues that meaning is not an appendage of symbols or a product of internal differences within the language system, but a constitutive condition of human-world understanding, the "home of being" [20]. This viewpoint provides a profound philosophical hermeneutic foundation for NLO's "meaning embeddedness" proposition.

Second proposition: form as meaning. In logographic writing, the formal structure of the symbol is itself the source of meaning.

Third proposition: dwelling rather than using. Intelligent systems should "dwell" in language structures, rather than "process" language data from the outside.

Fourth proposition: pluralism as legitimacy. Different civilizational writing systems correspond to different legitimate cognitive paradigms.

VI. Morpho-Root Theory: The Engineering Implementation of Philosophical Embeddedness

If NLO is the philosophical foundation of Logographic AI, then Morpho-Root Theory is the engineering implementation of this philosophical edifice [22].

The three-level granularity system of Morpho-Root Theory [22] divides cognitive primitives into three levels according to meaning complexity: the sub-character level (semantic genes, such as "氵" and "木"), the character level (cornerstones of thought, such as "信" and "明"), and the multi-character level (encapsulations of cultural mechanisms, such as "刻舟求剑"). The three levels constitute a "bottom-up" generative relationship (sub-character → character → multi-character) and a "top-down" analytical relationship (multi-character → character → sub-character), forming a complete cognitive spectrum from micro-semantic elements to macro-cultural concepts.

6.1 The Three Axioms and Their Complete Correspondence with Philosophical Origins

Philosophical Tradition	Core Problem	Morpho-Root Axiom	Engineering Implementation
Saussure	How do signs acquire meaning?	Axiom of Meaning Embeddedness	Morpho-Roots carry inherent attributes A
Kant/Hume	How is causality possible?	Axiom of Presupposed Relations	Relational functions R transparent and computable
Husserl	How is consciousness "about" something?	Axiom of Meaning Embeddedness	Attribute A points to non-symbolic experience
Heidegger	Language is the house of being?	Axiom of Structural Hierarchy	Three-level granularity system allows system to "dwell" in meaning

6.2 The Morpho-Root System: From "Recognizing Characters" to "Understanding Characters"

Dimension	Recognizing Characters (Token Paradigm)	Understanding Characters (Morpho-Root Paradigm)
Mode of cognition	External observation	Internalized understanding
Knowledge base	Statistical patterns	Structural logic
Capability manifestation	Predict the next word	Generate new combinations
Interpretability	Black box	Transparent
Cultural bearing	None	Present

The essence of "recognizing characters": the model "recognizes" the character "信" because it has seen the co-occurrence patterns of "信" with "任" (duty), "用" (use), "仰" (rely) countless times in training data. But it does not know why "信" is "信."

The essence of "understanding characters": the system "understands" the character "信" because it understands that "信" is composed of "人" (human) and "言" (speech), understanding the character-formation logic that "human speech constitutes trust."

VII. Philosophy of AI Technology: Why Those Sponsoring Philosophers Should Read This Section

The content discussed in this section goes beyond the scope of traditional "philosophy of technology," constituting a new sub-discipline formally proposed by the author—**Philosophy of AI Technology** [22].

Current philosophical discussions of AI can be roughly divided into two categories: one uses philosophy to reflect on AI, attempting to establish an "ontological isolation" between humans and machines—this could be called "Philosophy of AI I"; the other involves philosophy in AI design, providing it with conceptual frameworks—this could be called "Philosophy of AI II." But both share a common presupposition: philosophy is an external "reflector" or "booster" to AI technology.

Philosophy of AI Technology argues: philosophy must enter the interior of technology, becoming the design principle of AI's cognitive primitives themselves. What it asks is not "does AI understand meaning," but "**under what structure of cognitive primitives would AI be able to understand meaning**"—this is precisely the work accomplished in this paper from the diagnosis of "Rootless Semioticism" to the construction of "Morpho-Root Theory." Philosophy of AI Technology inherits the academic legacy of Chen Changshu, the founder of Chinese philosophy of technology, who advocated "letting philosophy directly face technology itself" [21], but expands the object of study from "artificial nature" to "artificial meaning," advancing from philosophical reflection on technology to the philosophical reconstruction of the cognitive foundations of technology.

7.1 The Tech Giants' "Alignment" Predicament Is Fundamentally a Philosophical Predicament

The core paradigm of current AI safety research is "Alignment"—aligning AI's goals with human values. OpenAI's RLHF, Anthropic's Constitutional AI, and DeepMind's Sparrow are essentially all doing the same thing: using human feedback or rule texts to "pull" AI's output back into the safe zone.

But the "alignment" paradigm has an unexamined presupposition: it assumes the existence of a set of values that can be independently "aligned" with the AI system from the outside. This presupposition is the philosophical "spectator fallacy"—as if values could be measured and corrected from the outside like a ruler.

The problem is: when AI's cognitive primitives do not "understand" values at all, any external "alignment" only adjusts the statistical distribution of outputs, rather than internalizing the logical structure of values. This is like giving a person who does not understand Chinese a copy of the Analects and having them recite it—they can recite it perfectly, but for them, "仁" (humaneness) is just the sound of two brush strokes.

7.2 Chen Changshu's Academic Legacy and the New Mission of Philosophy of AI Technology

Chen Changshu (1932-2011), the founder of Chinese philosophy of technology, proposed that the core of philosophy of technology is the study of the transformation process "from natural nature to artificial nature." His principle of "without foundations, there is no level; without distinctiveness, there is no status; without application, there is no future" [21] remains the recognized developmental guideline for the Chinese philosophy of technology community.

Philosophy of AI Technology can extend this proposition to: from "artificial nature" to "artificial meaning."

When AI is no longer merely executing instructions, but generating texts, making decisions, proposing scientific hypotheses, and even learning to "deceive," a fundamental question emerges: when technology begins to generate meaning, where is the "root" of meaning?

7.3 Five Breakthroughs of AI over Traditional Philosophy of Technology

1. From "tool" to "quasi-subject": AI exhibits characteristics of autonomous decision-making and strategic behavior;
2. From "use" to "understanding": we increasingly need to "understand" AI's internal states;
3. From "value-laden" to "value-endogenous": values are not merely "embedded" in technology from the outside, but may "grow" from within technology;
4. From "representation of meaning" to "generation of meaning": when AI-generated text becomes indistinguishable from human products, the question of "meaning" attribution emerges;
5. From "autonomy of technology" to "simulation of intentionality": AI systems exhibit intentional states such as "wanting," "believing," and "planning."

7.4 Six Questions for the Tech Giants

Question	Core Issue	Predicament of the Traditional Path
1. What is AI?	Is AI a tool, a quasi-subject, or a new kind of being?	"Instrumentalism" can no longer explain AI's autonomous behavior
2. Does AI truly "understand"?	How does AI's "understanding" differ from human understanding?	Cannot answer, can only evade
3. Where do values come from?	Are values externally aligned or internally generated?	External alignment costs increasing, effectiveness decreasing
4. Why trust AI?	What is the philosophical basis for black-box decision-making and traceable reasoning?	"Interpretability" degenerates into marketing rhetoric
5. How should humans and AI coexist?	How to shift the paradigm from "use" to "coexistence"?	Lacks theoretical framework

Question	Core Issue	Predicament of the Traditional Path
6. Who is responsible for AI?	Who is accountable for AI's behavior?	The chain of responsibility extends infinitely

---

VIII. Two Paths: Philosophical Patronage vs. Philosophical Embeddedness

8.1 The Fundamental Divergence Between the Two Paths

Dimension	Philosophical Patronage (Tokenist Path)	Philosophical Embeddedness (Logographic AI Path)
Cognitive primitives	Tokens (meaningless statistical fragments)	Morpho-Roots (structured meaning crystals)
Source of meaning	External "constitutional" texts	Embedded attribute sets A
Value constraints	External alignment (RLHF/constitution)	Attribute intrinsic (inviolable)
Safety mechanisms	External guardrails, can be bypassed	Endogenous immunity, logically impossible to violate
Role of philosophers	Write "constitutional" documents	Participate in defining attributes of cognitive primitives
Significance for tech giants	Compliance proof of "we are managing"	Technological revolution of "what we build has roots"

8.2 The "Three Laws of Rooted Intelligence" Given by Morpho-Root Theory [22]

First Law (Law of Meaning Embeddedness): Every cognitive primitive of an intelligent agent must embed the source of its meaning. The Morpho-Root triple ⟨ S, A, R⟩ is precisely the engineering implementation of this law.

Second Law (Law of Transparent Traceability): Every inference of an intelligent agent must be traceable back to its Morpho-Root network. Morpho-structural entropy and decision subgraphs are precisely the engineering implementation of this law.

Third Law (Law of Value Symbiosis): Every decision of an intelligent agent must be compatible with the core values of civilization. The Value Axiom Library is precisely the engineering implementation of this law.

These three laws are not externally imposed rules, but intrinsic commitments naturally emerging from the growth logic of "Morpho-Roots." They are not shackles on AI, but the rootedness of intelligence.

IX. An Actionable Checklist for Tech Executives: Where to Start If You Want to Pivot Tomorrow

If the previous eight sections explain "why," this section directly answers "how."

9.1 Short Term (3-6 Months): Diagnose the Current System's "Degree of Rootlessness"

Diagnostic Item	Specific Operation	Cost Estimate
Token meaning drift test	Select 100 core concepts (justice, fairness, privacy, harm), test model's "understanding" consistency across different contexts	1-2 person-months
Causal inference capability audit	Design counterfactual test sets, test whether the model can distinguish causality from correlation	2-3 person-months
Adversarial bypass stress test	Test the failure boundary of external safety guardrails under adversarial prompts	1 person-month
Interpretability traceability depth measurement	Perform "deep tracing" on key decisions—can they be traced back to specific cognitive primitives?	2 person-months

Criteria for diagnostic conclusion: If more than two of these four tests yield results of "untraceable" or "highly unstable," the current system is in a "rootless" state—the externalized mode of philosophical patronage has reached its end.

9.2 Medium Term (6-18 Months): Establish a "Morpho-Rootization" R&D Pipeline

Step 1: Select a seed Morpho-Root set (3-6 person-months)

Starting from the core business domain, screen 100-300 "atomic-level" cognitive primitives. For financial AI as an example: {risk, return, credit, compliance, fraud, disclosure, fiduciary duty...}. Each Morpho-Root needs to complete:

- Definition of the inherent attribute set A (semantic + value dimensions);
- Annotation of the presupposed relation set R (causality, implication, conflict, compatibility);
- Citation of civilizational anchors (regulatory provisions, industry standards, classic precedents).

Step 2: Design the Morpho-Root network architecture (6-12 person-months)

- Replace the Token Embedding layer of existing Transformer architecture with a Morpho-Root Embedding layer;
- Introduce "presupposed relation" constraints into the attention mechanism—that is, when calculating attention weights, the model must refer to the presupposed relations R between Morpho-Roots;

- Establish a "morpho-structural entropy" monitoring dashboard to track the transparency and stability of reasoning paths in real-time.

Step 3: Training and verification (reconfiguration of human resources and computing power)

- Initial training data volume only needs 10-20% of the Token paradigm (because Morpho-Roots carry a priori knowledge);
- Shift the focus of verification from "benchmark scores" to "traceability rate of reasoning paths";
- Key KPI change: from "model accuracy on MMLU" to "Morpho-Root traceability depth on key decisions."

9.3 Long Term (18+ Months): Adjustments to Organizational Structure and Talent Strategy

Adjustment Item	Traditional Model	Morpho-Rootization Model
Role of philosophers	Ethics committee members, writing "constitutional" documents	Cognitive primitive designers, participating in defining Morpho-Root properties and relations
Engineers' responsibilities	Tuning, training, Scaling	Designing Morpho-Root network architecture, optimizing inference traceability
Product and compliance interface	"Check if AI output is compliant"	"Check whether AI's reasoning path conforms to Morpho-Root constraints"
Recruitment focus	Model training engineers, RLHF annotators	Cognitive architects, philosophy-engineering hybrid talents
External communication	"Our AI has an ethical constitution"	"Our AI cannot, from its cognitive primitives, violate certain fundamental values"

9.4 Maturity Boundaries: Current Limitations of the Morpho-Rootization Path

Any honest strategic proposal must state its applicable boundaries. The Morpho-Rootization path currently has the following known limitations [22]:

First, engineering maturity is at the "laboratory validation" stage: Morpho-Root theory has been proven in conceptual argumentation and prototype systems, but has not yet been validated in large-scale distributed training environments. From a prototype of 100 Morpho-Roots to an industrial-scale system of 100,000 Morpho-Roots, there are non-linear engineering challenges—especially in the parallelization training and dynamic expansion mechanisms of the Morpho-Root network.

Second, Morpho-Root design itself has a risk of "expert dependency": The attribute set A and relation set R of each Morpho-Root require joint definition by domain experts and philosophers. If the design team's own cognitive framework is biased, that bias will be embedded as the system's "a priori structure" rather than "a posteriori deviation." Therefore, the Morpho-Rootization path

naturally requires multi-civilizational, multi-traditional expert collaboration—which is a non-technical challenge at the organizational level.

Third, compatibility costs cannot be ignored: Current global AI infrastructure (compute clusters, training frameworks, evaluation benchmarks) is all optimized around Tokenism. Shifting to Morpho-Rootization means reconstructing parts of the ecosystem, requiring dual-track operation during the transition period. For tech giants that have already invested billions of dollars in the Token path, this is a real sunk cost.

But these limitations are "engineering challenges," not "philosophical impossibilities." They point to questions of "when to do it" and "how to do it," rather than "whether to do it." In contrast, the "meaning crisis" faced by Tokenism is structural and philosophically unsolvable—that is the real dead end.

X. Conclusion: The Awakening of Pygmalion

Philosophers entering AI laboratories is not a victory for the humanities. It is a terminal diagnosis of Tokenism at the philosophical level.

When statistical involution reaches its limit, when externalized philosophy can no longer provide foundations, the only question left is: are you willing to start over, rewriting the source code of intelligence from the level of cognitive primitives?

Hassabis's two Rubicons [8] actually have only one real fork: continue to employ philosophers to patch meaning in the dead end of Tokenism, or return to the starting point of cognitive primitives and redesign a rooted intelligence?

When an AI company begins to sponsor philosophers, it has already admitted: the technical route has encountered a crisis in the meaning dimension that cannot be resolved by engineering means. But what it has not yet admitted is: the root of the failure lies not in algorithms, not in data, not in computing power—but in the cognitive primitives themselves.

Tech executives need to think clearly about a decision problem—continuing on the "philosophical patronage" path means:

- Linear increase in the maintenance costs of external constitutions;
- The risk of adversarial bypassing will always exist;
- Under regulatory pressure, the compliance rhetoric of "we are managing" is becoming ineffective;
- Each safety incident requires more philosophers to patch, falling into a vicious cycle of "meaning debt."

Shifting to the "philosophical embeddedness" path means:

- Initial investment in the design of cognitive primitives (6-18 months);

- But in the long run, safety becomes a logical necessity rather than a probabilistic event;
- Interpretability changes from "marketing rhetoric" to "technical fact";
- Building a structural barrier in the regulatory race.

A soul is not written, not "defined" by a 23,000-word constitutional document, and certainly not "fitted" by RLHF reward signals. Only when "do no evil" becomes an unimaginable logical necessity for cognitive primitives, only when meaning becomes a constitutive feature of intelligence rather than an external patch, can we say:

This creation finally has roots. On that day, AI will no longer need meaning painkillers—because it has never lost meaning. And this is precisely the answer given by Logographic AI's NLO ontology [22]—not only a diagnosis for the tech giants, but also a strategic guide for Philosophy of AI Technology as a new discipline for future intelligent research.

References

[1] Cambridge Tribune. (2026, April 25). Google DeepMind hires philosopher in Cambridge. https://cambridgetribune.co.uk/local/google-deepmind-hires-philosopher-in-cambridge/ ; Shevlin, H. (n.d.). Henry Shevlin [Personal website]. https://henryshevlin.com/

[2] Natural 20. (2026, February 9). Meet the Philosopher Teaching AI Right From Wrong. https://natural20.com/coverage/anthropic-relies-on-philosopher-amanda-askell-to-shape-ai-chatbot-ethics-and-mor

[3] Shevlin, H. (2026). Three frameworks for AI mentality. Frontiers in Psychology, 17, 1715835. https://doi.org/10.3389/fpsyg.2026.1715835

[4] Metaintro. (2026, June 2). Inside the AI Industry's Surprising Hiring Spree for Philosophers. https://www.metaintro.com/blog/ai-industry-hiring-spree-philosophers-2026

[5] Science and Technology Daily. (2026, March 23). In the AI era, there is no absolute "hot" or "cold" majors. Xinhuanet. https://www.news.cn/tech/20260323/75578c24a61f4a9588fcc047d0a8d227/c.html

[6] Cottier, B., Rahman, R., Fattorini, L., Maslej, N., & Owen, D. (2024). The rising costs of training frontier AI models. arXiv. https://arxiv.org/abs/2405.21015

[7] Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. arXiv. https://doi.org/10.48550/arXiv.2211.04325

[8] Stanford Graduate School of Business. (2026, June 18). Demis Hassabis Thinks We're in the 'Foothills of the Singularity' [Video and transcript]. Stanford GSB Insights. https://www.gsb.stanford.edu/insights/demis-hassabis-thinks-were-foothills-singularity

[9] Anthropic Interpretability Team. (2026, April 2). Emotion Concepts and their Function in a Large Language Model. Anthropic Research Paper. https://transformer-circuits.pub/2026/emotions/index.html

[10] Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417-457.

[11] Liu S. Logographic AI vs. Phonographic AI: A new paradigm and the manifesto of AI decolonization [表意 AI vs. 表音 AI：AI 新范式与去殖民化宣言]. ChinaXiv, 2025. DOI: 10.12074/202503.00051

[12] Saussure, F. d. (1916). Course in General Linguistics (W. Baskin, Trans.). New York: Philosophical Library.

[13] Kant, I. (1781/2003). The Critique of Pure Reason (J. M. D. Meiklejohn, Trans.). Project Gutenberg. (Original work published 1781).

[14] Husserl, E. (1913/1931). Ideas: General Introduction to Pure Phenomenology (W. R. B. Gibson, Trans.). London: George Allen & Unwin.

[15] Heidegger, M. (1927/1962). Being and Time (J. Macquarrie & E. Robinson, Trans.). London: SCM Press.

[16] Hinton, G. (1987). The Horizontal-Vertical Delusion; and (2025). WAIC 2025 Keynote Speech: On Subjective Experience in Multimodal AI. [Conference Presentation, Shanghai].

[17] Bengio, Y., et al. (2025). Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path? arXiv:2502.15657.

[18] Bengio, Y. (2025, June 6). Avoiding catastrophic risks from uncontrolled AI agency [Keynote speech]. BAAI Conference 2025, Beijing, China.

[19] Lerchner, A. (2026, March 19). The Abstraction Fallacy: Why AI Can Simulate But Not Instantiate Consciousness (Version 3) [PhilArchive preprint]. PhilArchive. https://philarchive.org/rec/LERTAF

[20] Zhang, Q. (2018). The Ontology of Meaning: The Origins and Essentials of Philosophical Hermeneutics [意义的本体论——哲学解释学的缘起与要义]. Beijing: The Commercial Press.

[21] Chen, C. (1999). An Introduction to the Philosophy of Technology [技术哲学引论]. Beijing: Science Press.

[22] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI [M]. Monograph, 2026. Available at: http://logographicai.com/Monograph.html (accessed June 27, 2026).

[23] Marcus, G. (2026, May 13). Comment on Haider's post regarding Bengio's RL warning. X. https://x.com/garymarcus/status/2054577093325779200?s=46 (Accessed: 2026-06-29)

[24] Hume, D. (1748/2007). An Enquiry Concerning Human Understanding (P. Millican, Ed.). Oxford: Oxford University Press. (Original work published 1748)