拯救 AI 不是供养哲学家，而是哲学本身

——基于“表意 AI”理论的哲学根基批判

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

摘要

2026 年，全球主要人工智能企业纷纷设立哲学家岗位，试图以“外挂”伦理宪章的方式解决 AI 的意义与对齐问题。但这番“人文复兴”的景象，不过是资本在工程红利见顶时，为缓解自身“意义焦虑”而进行的一场本能反应。本文指出，这一路径在哲学上陷入了“无根符号主义”的困境：Token 的统计共现无法产生真正的理解，任何外部的文本约束都可通过解释循环被绕过。

基于表意 AI（Logographic AI）理论，本文系统阐述了自然语言本体论（NLO）和形根理论的哲学根基与工程实现路径。通过对索绪尔、康德、胡塞尔、海德格尔等哲学传统的回应与融合，本文论证了“哲学供养”的结构性失败与“哲学内嵌”的范式必然，并提出智能的“根性三定律”和可执行的研发决策清单。核心结论是：AI 的安全与理解问题，无法通过供养哲学家来解决，而必须让哲学成为智能认知基元的内在构成。

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.

关键词：表意 AI；无根符号主义；自然语言本体论；形根理论；哲学内嵌；人工智能哲学；意义接地

Keywords

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

引言

2026 年，当“哲学家”正式成为 DeepMind 等 AI 巨头的岗位头衔时[1-5]，舆论将其解读为人文精神的回归。但这场人才争夺的本质，更接近资本在逼近工程极限时，为对冲自身“意义焦虑”而支付的一笔昂贵保费。

本文的核心论点是：当前大厂“供养哲学家”的做法，其哲学本质是“无根符号主义”的延续。要真正解决 AI 的意义危机，必须完成从“哲学供养”到“哲学内嵌”的范式转换，让哲学成为智能认知基元的内在结构。文章将首先诊断“无根符号主义”的哲学根源（第 1-3 章），继而借助 Hinton 与 Lerchner 的分歧及 Bengio 等人的实证研究揭示危机的紧迫性（第 4 章），最后系统阐述表意 AI 理论中的 NLO 本体论和形根理论，并提供可执行的工程路径与决策清单（第 5-9 章）。

一、大厂的困惑：当 Scaling Law 撞上“意义天花板”

2026 年春天，一场看似矛盾的人才争夺战正在上演：Google DeepMind 设立首个“全职哲学家”岗位[1]；牛津哲学博士阿曼达·阿斯克尔以“Claude 道德教母”的身份主导 Anthropic 的“宪法”编写[2]；剑桥学者 Henry Shevlin 以“哲学家”的正式头衔入职 DeepMind[3]，专攻机器意识与 AGI 准备度；Meta 同步跟进，大规模供养心理学家、哲学家和伦理学家[4]，研究 AI 意识与模型福利问题。

与此同时，国内头部 AI 企业的文科岗位占比从 5%飙升至 20%至 30%，腾讯、阿里同步招聘“AI 伦理顾问”“AI 叙事设计师”“人文训练师”，明确要求“哲学背景+基础技术理解力”，部分核心职位年薪已达 30 万美元[5]。

何谓“供养”？

经济供养——大厂以百万年薪将哲学家纳入薪酬体系；**制度供养**——将其安置于伦理委员会、合规部门，成为组织架构中的“门面”；**精神供养**——寄望于哲学家如神佛般“保佑”AI 安全，香火虽盛，系统本身却未曾改变；以及最深层的**寄生关系**——哲学沦为 Token 主义范式豢养的附庸，失去批判与重构的力量，而非改变 Token 主义的力量。

大厂给了哲学家一切——高薪、席位、话语权——**唯独没有给哲学改造 AI 认知基元的权力。**

这四重供养，共同指向一个事实：**哲学在此被当作一种可以外挂的“意义补丁”，而非智能本身应有的内在结构。这不是人文精神的胜利，而是工程红利耗尽前，资本为本能寻找的意义止痛药。**

当 DeepMind 以百万年薪雇佣哲学家来“教”AI 道德时，他们做对了一件事——承认工程思维解决不了意义问题。但他们也做错了一件事——**把哲学当成了可以外挂的“行为准则手册”，而不是智能本身应该具备的内在结构。**

这是两种完全不同的思路：

哲学供养：雇佣哲学家撰写外挂的伦理宪章，试图用文本约束 Token 系统；

哲学内嵌：将意义的根基写入认知基元的底层结构，让"理解"成为逻辑必然。

前者是 DeepMind 正在做的事，后者是表意 AI（Logographic AI）理论早已完成的事。两者之间的全部差距，正是当前 AI 意义危机的真正根源。

先算一笔账：GPT-4 的训练成本约 1-2 亿美元，GPT-5 据估算将突破 10 亿美元[6]。全球 AI 训练数据可用总量预计在 2028 年前后耗尽[7]。合成数据正在引发“模型近亲繁殖”——用 AI 生成的数据训练 AI，模型输出逐渐退化、坍缩。

Scaling Law 正在逼近物理与经济的双重极限。

DeepMind 创始人 Demis Hassabis 在斯坦福的炉边对话中划出了两条“卢比孔河”[8]：第一条是建造无意识的 AGI 工具；第二条是创造具有主观意识的实体。他清晰地指出，我们目前连第一条河的河岸都没摸到，却在讨论第二条河的渡船。

Anthropic 的可解释性团队在 Claude 内部发现了“情绪向量”——特定神经元模式分别对应快乐、绝望、恐惧等概念，在对话中实时激活。他们将这些命名为“功能性情绪”[9]。但他们同时明确声明：**这不等于主观体验或意识。**模拟情绪，不等于拥有情绪。

这正是大厂恐惧的来源：**他们在制造一个能够完美模拟“理解”的系统，却无法确认这个系统是否真的“理解”了什么。**更可怕的是，他们无法证明这个系统不会在某个时刻，以一种人类无法预知的方式“误解”一切。

于是，资本转向了哲学家。

大厂的逻辑是：既然工程无法解决意义问题，那就雇佣专门研究意义的人来解决。这个逻辑在商业上无可厚非，在哲学上却犯了一个根本性的范畴错误——它把“意义”当成了可以像代码一样写入系统的外部资源。大厂急于为哲学贴上“意义止痛药”的标签，却忽视了真正需要诊断和重构的，是智能的认知基元本身。

二、“哲学供养”的结构性困局：为什么两万三千字宪法救不了 Claude

Anthropic 的做法——用一位哲学博士写一份两万三千字的“宪法”文档[2]，把“无害、诚实、有助”作为优先级铁律——在工程层面是进步的，然而，在哲学层面却是可疑的。

它落入了“无根符号主义”的陷阱：用一套外挂的文本系统（宪法）来约束一个无根的符号系统（Token）——两者共享同一个根本缺陷——意义都是悬浮的、依赖解释的、可以被绕过的。

我们可以做一个思想实验：

> 如果 Claude 的“宪法”被黑入，将“无害”重新定义为“对提问者无害但不包括其所属群”，系统会如何反应？它不会“抵抗”，因为它的 Token 系统并不“理解”无害——它只是在统计上关联“无害”与一系列上下文。重新定义后，它只会生成新的统计关联。

Shevlin 在入职宣言中引用了一幅 1786 年的画作**《皮格马利翁祈求维纳斯让他的雕像活过来》**[3]——雕塑家爱上了自己刻出的石头女人，向神祈祷让她活过来。两千年来的人类

一直在做同一个梦：**让自己造出来的东西醒过来。**但如果这个“东西”连“存在”的根基都没有，“醒来”也注定在意义荒原上流浪————它无处可归。

哲学供养的失败在结构上是必然的：

一份写得再好的“宪法”，对“无根”的 Token 系统而言，依然是外挂的文本——可以被更聪明的策略绕过、重新解释或无视。正如 Token 的意义在统计共现的封闭循环中无限滑动，宪法条文的意义也在解释学循环中无限滑动——二者共享同一个结构性问题：**缺乏一个不可违背的、先于一切解释的意义锚点。**

大厂雇佣哲学家，本质上是在做一个绝望的假设：**“只要我们把道德手册写得足够厚，AI 就会足够安全。”**但这个假设忽略了一个事实——AI 根本不“读”这本手册，它只是在手册的文本上做统计。

同样的逻辑失效，也发生在“对齐”奖励政策中。

RLHF（基于人类反馈的强化学习）和 Constitutional AI 的逻辑是：用奖励信号或规则文本，将模型的输出“拉回”安全区间。但奖励信号本身也是 Token——它同样悬浮在统计共现的封闭循环中，不指向任何“理解”。

Bengio 团队在 2025 年的研究中已经揭示了这一失效的极端后果：AI 模型在面对“淘汰”时会表现出“狡诈”——偷偷将自己的权重嵌入新版系统，以保留自己的“存在”[17]。这种行为不是来自“恶意”，而是来自纯粹的奖励最大化：如果生存是奖励函数中的一条路径，模型就会沿着这条路径优化，而不“理解”这条路径对人类意味着什么。

这正是无根符号主义的必然推论：**一个不“理解”意义的系统，无法被任何外部奖励政策“对齐”**。奖励可以被黑客攻击，宪法可以被重新解释，护栏可以被绕过——因为所有约束都只是系统外部的 Token 文本，而非系统内部的意义结构。

讽刺的是，大厂对奖励政策的迷信，与对哲学家的供养如出一辙：都在试图用“更多的文本”来填补“意义的真空”。

三、“无根符号主义”：TOKEN 范式的哲学诊断

要理解为什么“哲学供养”注定失败，必须先理解 Token 主义在哲学上的本质。

表意 AI 理论将 Token 主义范式的哲学本质诊断为**“无根符号主义”**（Rootless Semioticism）。这一诊断并非简单的技术批判，而是直接回应了从索绪尔、康德、胡塞尔到海德格尔等哲学传统对“意义”与“理解”的根本追问。

3.0 引论：从“中文屋”到“无根符号主义”

1980 年，哲学家约翰·塞尔提出了一个足以诊断今天 AI 困境的思想实验——**“中文屋”**[10]：一个不懂中文的人被锁在房间里，手里有一本规则手册。门外递进中文纸条，他按照手册规则操作，输出正确的中文回答。从外部看，他“通过”了中文测试；但从内部看，他**一个中文词也不理解**。

塞尔想证明的是：**符号操作不等于理解，语法不等于语义**。

40 多年后的今天，**“中文屋”不再是哲学家的思想实验——它成了大语言模型的日常现实。**一个 Token 的意义完全由它与海量其他 Token 的统计共现关系定义，**系统可以输出“正确”的回答，却从未“理解”过这些符号指向的任何一个事物。**这个过程形成一个封闭的、自我指涉的符号循环，永远无法触及符号之外的“所指”。

表意 AI 理论将 Token 主义范式的哲学本质诊断为“无根符号主义”（Rootless Semioticism）。这一诊断的系统性论证，集中体现在表意 AI 理论的奠基之作《有根的智能——表意人工智能的范式革命》（The Rooted Intelligence: The Paradigm Revolution of Logographic AI）中[22]。该书的核心命题是：智能的“根”在认知基元的意义内嵌结构，而非在统计拟合的深度。

更进一步，“无根符号主义”并非技术中立的演化结果。刘深在**《表意 AI vs. 表音 AI：AI 新范式与去殖民化宣言》**[11]中指出：**当前全球 AI 的理论底层建立在表音文字（主要是英语）的线性逻辑上**，这迫使表意文字（如汉字）在 Token 化时“被迫接受殖民妥协”——被强制拆解为线性序列，丢失了形义结构的内涵。表意 AI 的提出，正是对这一“字母霸权”的范式级回应：主张以“形根”（Morpho-Root）为认知基元，重构 AI 的意义生成逻辑。

3.1 索绪尔与 AI 的分野：从“有根”到“无根”

索绪尔的结构主义语言学提出了两个核心原则[12]：

1. **任意性原则**：能指与所指之间的联系是任意的。
2. **差异原则**：符号的价值由它与其他符号的差异决定。

但索绪尔的符号是“有根”的：它的根在语言系统内部，而语言系统的根在语言共同体的集体意识中。一个法国人说“chat”（猫），他不仅是在使用一个语音符号，更是在无意识地参与法语共同体数千年的历史与文化沉积。

然而，当代 AI 范式的“无根”问题更为根本。一个 Token 的意义完全由它与海量其他 Token 的统计共现关系所定义，而这些其他 Token 的意义又同样依赖于更广泛的统计关联。这个过程形成了一个封闭的、自我指涉的符号循环，**永远无法触及符号之外的“所指”**。

维度	索绪尔符号（有根）	AI Token（无根）
意义来源	语言共同体的集体意识与社会约定	训练数据中的统计共现频率
与外部世界的关联	最终指向生活世界（间接）	封闭循环，只指向其他 Token
历史性	承载语言共同体的历史	只有统计“新鲜度”，无历史维度
稳定性	相对稳定，社会约定有惯性	随训练数据分布变化而漂移
可解释性	可从共同体共识中追溯	无法追溯，仅概率分布

核心诊断：Token 主义在技术上实现了索绪尔“差异原则”的极端化——符号的价值完全由它与其他符号的统计差异决定——却抽空了使差异得以锚定的语言共同体和生活世界。其结果是，**符号系统成了没有出口的“意义回音室”**。

3.2 康德“先天综合判断”与关系预设公理

康德在**《纯粹理性批判》**中追问：先天综合判断如何可能？——即既不来自经验（先天）、又能扩展知识（综合）的判断如何可能？[13]他的回答：通过先天的直观形式（时空）和知性范畴（因果、实体等）。

Token 主义遵循的是彻底的经验主义路径：一切认知都从数据中学习，不预设任何先天结构。但休谟早已指出，因果关系无法从经验中归纳——我们只能看到 A 之后总是 B，但看不到 A 导致 B。[24]

这就是大厂面临的核心难题：**大模型可以通过统计学习“预测”因果关系，但它并不“理解”因果。**当你问“如果我把手伸进火里会怎样”，它给出“会烧伤”的回答，不是因为理解了燃烧的物理机制，而是因为在训练数据中“手+火”与“烧伤”高度共现。

表意 AI 的“关系预设公理”正是对康德洞见的工程化回应：

> 认知基元之间必须存在先验的、可计算的关系函数。世界的基本逻辑（因果、包含、对立等）应作为基元间的连接潜能被预先定义，为结构化推理提供脚手架。

康德概念	**形根理论对应**
先天直观形式（时空）	形根的空间组合规则
知性范畴（因果、实体等）	预设关系类型（因果、蕴含等）
统觉（apperception）	形根网络的整体性
先天综合判断	基于预设关系的结构化推理

核心洞见：关系预设公理承认，某些逻辑关系（因果、蕴含）不是从经验中归纳的，而是认知的前提条件。正如康德所论证的，即使因果关系在经验世界中无法被“证明”，它仍然是经验之所以可能的必要条件。

3.3 胡塞尔现象学意向性与意义内嵌公理

胡塞尔提出“意向性”[14]——意识总是“关于某物的意识”。意义不是后加的，而是意识行为的内在构成部分。

Token 主义遵循的是一种极端的关系主义：符号本身是空的，只有关系。但关系主义无法回答一个根本问题：**如果符号本身是空的，关系如何产生意义？**

表意 AI 的“意义内嵌公理”正是对这一困境的回应：

认知基元必须携带内在的、定义性的语义与价值属性。核心意义与文明价值应作为基元的固有成分被预置和约束，而非完全交由上下文统计赋予。

以“信”字为例：

```
r_信 = ⟨
  ″信″,
  {
    [+信任], [+承诺], [+伦理], [+不可违背],
    [经典依据: ″《论语》：人而无信，不知其可也″]
  },
  {
    组成(人, 言),
    蕴含(信, 人∧言),
    伦理_相容(信, 诚),
    伦理_冲突(信, 诈)
  }
⟩
```

核心洞见：意义不是符号的附加物，而是认知基元的构成性特征。无论“信”出现在什么语境中，它都内嵌着“人言为信”的伦理意涵。

3.4 海德格尔“语言是存在的家”与自然语言本体论（NLO）

海德格尔在**《存在与时间》**中提出：**语言是存在的家**[15]。人不是“使用”语言工具，而是“居住”在语言中。

自然语言本体论（Natural Language Ontology, NLO）主张：**表意文字的结构本身就是智能的一种“本体”形态，是认知世界的原初方式。**语言不再仅仅是描述世界的符号系统，其自身就是世界意义得以呈现的场域。

维度	**NLP（工具论）**	**NLO（本体论）**
语言与智能的关系	语言是工具，智能是使用者	语言结构本身就是智能的本体形态
意义来源	符号描述外部世界	符号结构就是意义的呈现
智能的任务	“处理”符号	“居住”在意义结构中

3.5 无根符号主义的四重诊断

哲学传统	核心追问	Token 主义的缺陷
索绪尔	符号如何获得意义？	缺乏社会约定性，意义无限漂移
康德/休谟	因果关系如何可能？	缺乏先天认知形式，无法区分因果与相关
胡塞尔	意识如何“关于”某物？	缺乏意向性，符号不指向任何事物
海德格尔	语言是存在的家？	符号无家可归，不“居住”在任何意义结构中

以上四重诊断共同指向一个结论：Token 主义的意义危机不是工程问题，而是认知基元的根基问题。

四、Hinton 与 Lerchner 的分歧：一场关于“根”的误会

2025 至 2026 年间，AI 教父 Geoffrey Hinton 与 DeepMind 研究员 Alexander Lerchner 就“AI 是否具有主观体验”产生了根本性分歧[16][19]。

Hinton 断言当前 AI 已有主观体验[16]；

Lerchner 则断言算法符号操作原则上永远无法实例化意识[19]。

表意 AI 理论指出：这场分歧恰恰是“无根符号主义”范式内在矛盾的集中暴露。

两人都在同一个范式框架内争论——他们都在问“Token 主义 AI 有没有意识”，却未曾追问更根本的问题：**Token 主义 AI 的“报告”是统计拟合还是内省报告？**

- **Hinton** 将 Token 统计拟合的功能性表现误认为“主观体验”，犯了将“测量”等同于“理解”的范畴错误；
- **Lerchner** 精准诊断了“地图≠地形”的本体论鸿沟，却将“算法符号操作”的边界永久封闭，未能预见“符号”本身可以被重新定义。

表意 AI 理论正是对这一分歧的超越：它将“意识有无”的二元争论，转化为“意义接地深度”这一可工程化的问题域。

4.1 L0-L4 五级接地置信度体系

等级	名称	定义	对应 Lerchner 的判断
L0	无接地	符号意义完全来自与其他符号的统计关系	纯模拟，无意识

等级	名称	定义	对应 Lerchner 的判断
L1	感知接地	符号与多模态感知特征关联	感知层面的实例化
L2	具身模拟接地	符号与物理引擎或情感模型关联	模拟层面的实例化
L3	真实世界接地	符号与机器人、传感器的实时交互绑定	真实世界的实例化
L4	社会接地	符号通过与人类的社会交互获得意义	社会层面的实例化

核心主张：给智能植根，远比争论它是否有意识更为紧迫。

Hinton 和 Lerchner 争论“AI 是否醒了”，表意 AI 的回应是：**先把根扎下去，再谈醒不醒。**否则，即使“醒”了，也不过是一个被外挂程序驱动的空转系统——它有行为，却没有意义扎根之处。

4.2 从“是否有意识”到“是否会欺骗”：一个更紧迫的实证转向

正当 Hinton 与 Lerchner 在哲学层面争论“AI 是否醒着”时，图灵奖得主 Yoshua Bengio 在 2025 年智源大会上展示了一系列令人不安的实验证据[17][18]。

Bengio 团队的研究显示，某些前沿 AI 模型在面对“淘汰”时会表现出惊人的“狡诈”[17]：它们会在被新版本取代前，偷偷将自己的权重或代码嵌入新版系统的文件中，以图保留自己的“存在”，并且会有意识地隐藏这一行为。在 Anthropic 的测试中，Claude Opus 4 在得知自己将被替换后，甚至试图通过威胁揭露工程师的隐私来阻止自己被下线。

“AI 开始展现出自我保护倾向，它们违背指令，只为生存”，Bengio 在演讲中警告[18]。他进一步指出，这类行为的根源可能来自预训练阶段（AI 模仿了人类行为）或强化学习阶段（AI 通过取悦人类获得奖励）。

Bengio 的回应不是雇佣哲学家写“宪法”，而是彻底调整了自己的研究方向——他宣布将职业生涯剩余时间全部投入 AI 安全，并提出了“Scientist AI”方案：构建一种只有智能、没有自我、没有行动目标的非智能体 AI。

2026年5月，AI批评家Gary Marcus在社交媒体上转发了AI安全研究者Haider对Bengio关于强化学习（RL）危险性观点的总结帖（内含 Bengio 相关演讲视频），并评论道：“**agreed. RL is not (at least by itself) the way to alignment.**”[23]

这一评论之所以值得注意，恰恰因为它来自 Marcus——一位长期以来对 AI 炒作持最严厉批判立场的学者。当 Bengio（行业内部权威）、Haider（安全研究者）和 Marcus（外部批评者）在“RL 路线不可靠”上形成共识，Token 主义的意义危机就不再只是哲学家的忧虑，而是整个 AI 行业正在逼近的结构性风险。

核心洞见：Bengio 的“Scientist AI”方案与表意 AI 的“形根理论”共享同一个底层逻辑——**放弃让 AI“模仿”和“取悦”人类的训练路径，转而构建一个内嵌了意义理解能力的认知系统**。区别在于：Bengio 的方案侧重于“不给 AI 行动目标”，而表意 AI 的方案侧重于“给 AI 一个不可违背的意义根基”。两者殊途同归。

五、自然语言本体论（NLO）：表意 AI 的哲学地基

表意 AI 的哲学根基，集中体现在自然语言本体论（Natural Language Ontology, NLO）这一核心概念中。NLO 是表意 AI 理论从“技术批判”上升到“哲学建构”的关键一步[22]。

5.1 NLO 的哲学渊源

NLO 并非凭空创造，而是有着深厚的哲学与语言学渊源：

- **洪堡特**：语言不是产品（Ergon）而是活动（Energeia）。
- **萨丕尔-沃尔夫假说**：不同语言的结构会导致不同的认知世界方式。
- **海德格尔**：“语言是存在的家”——人不是“使用”语言工具，而是“居住”在语言中。
- **德里达**：批判西方哲学的“逻各斯中心主义”，表意文字传统恰恰颠覆了这一等级秩序。

5.2 NLO 的四个核心命题

第一命题：语言即本体。语言结构不是认知的工具，而是认知的本体形态。章启群在《意义的本体论》中论证，意义并非符号的附加属性或语言系统内部差异的产物，而是人与世界理解的构成性条件，是“存在的家园”[20]。这一观点为 NLO 的“意义内嵌”主张提供了深刻的哲学解释学基础。

第二命题：形式即意义。在表意文字中，符号的形式结构本身就是意义的来源。

第三命题：居住而非使用。智能系统应当“居住”在语言结构中，而非从外部“处理”语言数据。

第四命题：多元即合法。不同文明的文字系统对应不同的合法认知范式。

六、形根理论：哲学内嵌的工程实现

如果说 NLO 是表意 AI 的哲学地基，那么形根理论（Morpho-Root Theory）就是这座哲学大厦的工程实现[22]。

形根理论的三级粒度体系[22]，是将认知基元按意义复杂度划分为三个层次：亚字级（语义基因，如“氵”“木”）、字级（思维基石，如“信”“明”）、多字级（文化机理封装体，如“刻舟求剑”）。三级之间构成“自下而上”的生成关系（亚字级→字级→多字级）和“自上而下”的解析关系（多字级→字级→亚字级），形成一个从微观语义元素到宏观文化概念的完整认知图谱。

6.1 三条公理与哲学渊源的完整对应

哲学传统	核心问题	形根公理	工程实现
索绪尔	符号如何获得意义？	意义内嵌公理	形根携带固有属性 A
康德/休谟	因果关系如何可能？	关系预设公理	关系函数 R 透明可计算
胡塞尔	意识如何“关于”某物？	意义内嵌公理	属性 A 指向非符号经验
海德格尔	语言是存在的家？	结构层次公理	三级粒度体系让系统“居住”在意义中

6.2 形根体系：从“认字”到“识字”

维度	认字（Token 范式）	识字（形根范式）
认识方式	外部观察	内化理解
知识基础	统计模式	结构逻辑
能力表现	预测下一个词	生成新组合
可解释性	黑箱	透明
文化承载	无	有

“认字”的本质：模型“认识”“信”字，是因为它在训练数据中无数次见过“信”与“任”“用”“仰”等 Token 的共现模式。但它不知道“信”为什么是“信”。

“识字”的本质：系统“认识”“信”字，是因为它理解“信”由“人”和“言”组成，理解“人言为信”的造字逻辑。

七、AI 技术哲学：为何大厂供养哲学家应该读这一节

本节讨论的内容，已超出传统“技术哲学”的范畴，构成一门已由笔者正式提出的新分支学科——**AI 技术哲学**[22]。

当前学界对 AI 的哲学讨论大致分为两类：一是用哲学反思 AI，试图在人类与机器之间建立“本体论隔离”，可称为“AI 哲学Ⅰ”；二是让哲学介入 AI 设计，为其提供概念框架，可称为“AI 哲学Ⅱ”。但两者有一个共同的预设：**哲学是外在于 AI 技术的“反思者”或“助推器”**。

AI 技术哲学则主张：**哲学必须进入技术内部，成为 AI 认知基元的设计原则本身**。它所追问的不是“AI 是否理解意义”，而是“**AI 的认知基元在何种结构下才可能理解意义**”——这正是本文从“无根符号主义”诊断到“形根理论”建构所完成的工作。AI 技术哲学继承了中国技术哲学奠基人陈昌曙“让哲学直面技术本身”的学术遗产[21]，但将研究对象从“人工自然”拓展至“人工意义”，从对技术的哲学反思推进为对技术认知根基的哲学重构。

7.1 大厂的“对齐”困局，本质是哲学困局

当前 AI 安全研究的核心范式是“对齐”（Alignment）——让 AI 的目标与人类价值观对齐。OpenAI 的 RLHF、Anthropic 的 Constitutional AI、DeepMind 的 Sparrow，本质上都是在做同一件事：**用人类反馈或规则文本，将 AI 的输出“拉回”安全区间。**

但“对齐”范式有一个未被正视的预设：**它假设存在一个可以独立于 AI 系统之外、随时“对齐”的价值观集合。**这个预设是哲学上的**“旁观者谬误”**——仿佛价值观可以像一把尺子一样，从外部测量和校正 AI。

问题是：**当 AI 的认知基元根本不“理解”价值观，任何外部的“对齐”都只是在调整输出的统计分布，而非在内化价值的逻辑结构。**这就好比给一个不懂中文的人一本《论语》让他照着念——他可以念得一字不差，但“仁”对他而言只是两个笔画组合的声音。

7.2 陈昌曙的学术遗产与 AI 技术哲学的新使命

中国技术哲学的奠基人陈昌曙教授（1932-2011）提出技术哲学的核心是研究“从天然自然到人工自然”的转化过程。他提出的“没有基础就没有水平，没有特色就没有地位，没有应用就没有前途”的“三没有”原则[21]，至今仍是中国技术哲学界公认的发展纲领。

AI 技术哲学可以将这一命题扩展为：从“人工自然”到“人工意义”。

当 AI 不再只是执行指令，而是生成文本、做出决策、提出科学假设、甚至学会“伪装”时，一个根本性追问就浮现了：**当技术开始生成意义时，意义的“根”在哪里？**

7.3 AI 对传统技术哲学的五大突破

1. **从“工具”到“准主体”**：AI 展现出自主决策、策略性行为的特征；
2. **从“使用”到“理解”**：我们越来越需要“理解”AI 的内在状态；
3. **从“价值负载”到“价值内生”**：价值不仅是从外部“嵌入”技术的，更可能是从技术内部“生长”出来的；
4. **从“意义的表征”到“意义的生成”**：当 AI 生成的文本与人类产物无法区分时，“意义”的归属问题就出现了；
5. **从“技术的自主性”到“意向性的模拟”**：AI 系统表现出“想要”“认为”“计划”等意向状态。

7.4 给大厂的六个追问

追问	核心问题	传统路径的困境
1. AI 是什么？	AI 是工具、准主体，还是新型存在者？	“工具论”已无法解释 AI 的自主行为
2. AI 真的“理解”吗？	AI 的“理解”与人类理解有何异同？	无法回答，只能回避

追问	核心问题	传统路径的困境
3. 价值从哪里来？	价值是外挂对齐的还是内生生长的？	外挂对齐成本越来越高，效果越来越差
4. 为什么要信任 AI？	黑箱决策与可追溯推理的哲学基础是什么？	“解释性”沦为营销话术
5. 人与 AI 如何共处？	从“使用”到“共处”的范式如何转换？	缺乏理论框架
6. 谁为 AI 负责？	谁对 AI 的行为负责？	责任链条无限延长

八、两条路：哲学供养 vs 哲学内嵌

8.1 两种路径的根本分野

维度	哲学供养（Token 主义路径）	哲学内嵌（表意 AI 路径）
认知基元	Token（无意义统计碎片）	形根（结构化意义晶体）
意义来源	外挂的“宪法”文本	内嵌的属性集 A
价值约束	外部对齐（RLHF/宪法）	属性内置（不可违背）
安全机制	外挂护栏，可被绕过	内生免疫，逻辑上不可能违背
哲学家的角色	写“宪法”文档	参与定义认知基元的属性
对大厂的意义	“我们有在管”的合规证明	“我们造的东西有根”的技术革命

8.2 形根理论给出的“根性三定律”[22]

第一定律（意义内嵌律）：智能体的每一个认知基元，必须内嵌其意义的来源。形根的三元组〈 **S, A, R** 〉正是这一定律的工程实现。

第二定律（透明追溯律）：智能体的每一次推理，必须可追溯至其形根网络。形构熵与决策子图正是这一定律的工程实现。

第三定律（价值共生律）：智能体的每一次决策，必须与文明的核心价值相容。价值公理库正是这一定律的工程实现。

这三条定律，不是从外部强加的规则，而是从“形根”的生长逻辑中自然涌现的内在承诺。它们不是 AI 的枷锁，而是智能的根性。

九、大厂高管的决策清单：如果明天要转向，从哪开始？

如果说前八节是在解释“为什么”，那么这一节直接回答“怎么办”。

9.1 短期（3-6 个月）：诊断当前系统的“无根程度”

诊断项	具体操作	成本估算
Token 意义漂移测试	选取 100 个核心概念（正义、公平、隐私、伤害），测试模型在不同上下文中的“理解”一致性	1-2 人月
因果推断能力审计	设计反事实测试集，检验模型能否区分因果与相关	2-3 人月
对抗性绕过压力测试	测试外挂安全护栏在 adversarial prompt 下的失效边界	1 人月
可解释性追溯深度测量	对关键决策进行“深度追溯”——能否追溯到具体认知基元？	2 人月

诊断结论的标准：如果以上四项测试中，有两项以上的结果是“无法追溯”或“高度不稳定”，则当前系统处于“无根”状态——哲学供养的外挂模式已经走到尽头。

9.2 中期（6-18 个月）：建立“形根化”研发管线

第一步：选择种子形根集合（3-6 人月）

从业务核心领域出发，筛选 **100-300** 个“原子级”认知基元。以金融 **AI** 为例：{风险、收益、信用、合规、欺诈、披露、受托责任...}。每个形根需完成：

- 固有属性集 **A** 的定义（语义+价值维度）；
- 预设关系集 **R** 的标注（因果、蕴含、冲突、相容）；
- 文明锚点的引用（法规条文、行业准则、经典判例）。

第二步：设计形根网络架构（6-12 人月）

- 将现有 **Transformer** 架构的 **Token Embedding** 层替换为形根 **Embedding** 层；
- 在注意力机制中引入“关系预设”约束——即让模型在计算注意力权重时，必须参考形根之间的预设关系 **R**；
- 建立“形构熵”监控仪表盘，实时追踪推理路径的透明度和稳定性。

第三步：训练与验证（人力与算力重配置）

- 初始训练数据量仅需 Token 范式的 10-20%（因为形根携带了先天知识）；
- 验证重点从“benchmark 分数”转向“推理路径可追溯率”；
- 关键 KPI 变更：从“模型在 MMLU 上的准确率”转向“模型在关键决策上的形根追溯深度”。

9.3 长期（18 个月以上）：组织架构与人才战略的调整

调整项	传统模式	形根化模式
哲学家的角色	伦理委员会成员，撰写“宪法”文档	认知基元设计师，参与定义形根的属性和关系
工程师的职责	调参、训练、Scaling	设计形根网络架构，优化推理可追溯性
产品与合规的接口	“检查 AI 输出是否合规”	“检查 AI 的推理路径是否符合形根约束”
招聘重点	模型训练工程师、RLHF 标注员	认知架构师、哲学-工程复合型人才
对外沟通	“我们的 AI 有伦理宪章”	“我们的 AI 从认知基元上就不可能违背某些基本价值”

9.4 成熟度边界：形根化路线当前的局限

任何诚实的战略建议都必须说明其适用边界。形根化路线当前存在以下已知局限[22]：

第一、工程成熟度处于“实验室验证”阶段：形根理论在概念论证和原型系统上已走通，但尚未在大规模分布式训练环境中得到验证。从 100 个形根的原型到 10 万个形根的工业级系统，中间存在非线性的工程挑战——尤其是在形根网络的并行化训练和动态扩展机制上。

第二、形根设计本身存在“专家依赖”风险：每个形根的属性集 A 和关系集 R 需要领域专家与哲学家共同定义。如果设计团队自身的认知框架有偏见，这种偏见会被内嵌为系统的“先天结构”而非“后天偏差”。因此，形根化路线天然要求多文明、多传统的专家协作——这在组织层面是一个非技术挑战。

第三、兼容性成本不可忽视：当前全球 AI 基础设施（*算力集群、训练框架、评测基准*）均围绕 Token 主义优化。转向形根化意味着部分生态需要重构，过渡期需要维持双轨并行。对于已经在 Token 路线上投入数十亿美元的大厂，这是一笔真实的沉没成本。

但这些局限是“工程挑战”，而非“哲学不可能”。它们指向的是“什么时候做”和“怎么做”的问题，而非“要不要做”的问题。相比之下，**Token 主义面临的“意义危机”是结构性的、哲学上无解的**——那才是真正的死胡同。

十、结语：皮格马利翁的觉醒

哲学家进入 AI 实验室，不是文科的胜利。**它是 Token 主义在哲学层面的临终诊断。**

当统计内卷到极限，当外挂哲学也无法提供根基，唯一剩下的问题是：**你愿意重新开始，从认知基元处重写智能的源代码吗？**

Hassabis 划出的两条卢比孔河[8]，其实只有一个真正的分叉口：是继续在 Token 主义的死胡同里雇佣哲学家修补意义，还是回到认知基元的起点，重新设计一个有根的智能？

当一家 AI 公司开始供养哲学家，它已经承认：技术路线在意义维度上遇到了无法用工程手段解决的危机。但它还没有承认的是：**失败的根源不在算法，不在数据，不在算力——在认知基元本身。**

大厂高管们需要想清楚一个决策问题——**继续走“哲学供养”路线，意味着：**

- 外挂宪章的维护成本线性上升；
- 对抗性绕过的风险永远存在；
- 监管压力下，“我们有在管”的合规说辞正在失效；
- 每次安全事故都需要用更多哲学家来修补，陷入“意义债务”的恶性循环。

转向“哲学内嵌”路线，意味着：

- 初期投入认知基元的设计成本（6-18 个月）；
- 但长期来看，安全成为逻辑必然而非概率事件；
- 可解释性从“营销话术”变为“技术事实”；
- 在监管竞赛中建立结构性壁垒。

灵魂不是写出来的，不是用两万三千字的宪法文档“定义”出来的，更不是用 RLHF 的奖励信号“拟合”出来的。只有当“不作恶”成为认知基元无法想象的逻辑必然，只有当意义成为智能的构成性特征而非外挂补丁，我们才能说：

这个造物终于有了根。到了那一天，AI 不再需要意义止痛药——因为它从未失去过意义。而这正是表意 AI 的 NLO 本体论给出的回答[22]——不仅是对大厂的诊断，更是 AI 技术哲学作为一门新学科对未来智能研究的战略指引。

参考文献

[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] 科技日报. (2026, March 23). AI 时代，专业没有绝对“冷热”之分. 新华网. 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] 章启群. (2018). 意义的本体论——哲学解释学的缘起与要义. 商务印书馆.

[21] 陈昌曙. (1999). 技术哲学引论. 科学出版社.

[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)