从“结构裁决”到“认知基元”：对邬江兴内生安全范式的认知根基性质疑

From "Structural Adjudication" to "Cognitive Primitive": A Foundational Critique of Wu Jiangxing's Endogenous Security Paradigm

摘要

中国工程院院士邬江兴提出的内生安全理论与动态异构冗余（DHR）架构，通过系统结构设计将不确定的网络安全威胁转化为可控概率问题，是网络空间安全领域的重大理论创新。然而，当该范式应用于以深度学习为代表的不可解释人工智能系统时，其安全保障本质上依赖于对“黑箱”输出的多数裁决，形成一种结构层面的概率性安全。

本文基于**“表意 AI”**（Logographic AI, LAI）与**“形根”**（Morpho-Root）认知基元框架，从认知基元层面重新审视内生安全的内涵，指出真正的内生安全应源于知识基元的逻辑必然性，而非系统架构的概率性约束。邬江兴院士的“内生安全”在哲学上是一种“伪内生”——它用系统架构的“新瓶”盛装了“黑箱”AI 的“旧酒”。

论文论证：当 DHR 架构的执行体均为基于无根 Token 的统计模型时，其安全机制无法彻底规避因认知基元缺陷导致的语义级共模失效风险——所有异构黑箱可能因同一结构性偏见而集体犯错，令多数裁决失效。基于此，本文提出内生安全的层次性区分：结构内生与认知内生，并展示表意 AI 如何通过“形根”这一承载固有语义与硬约束的认知基元，在推理底层实现逻辑阻断，为从认知根源构建内生安全提供了另一种范式可能。

Abstract

The endogenous security theory and Dynamic Heterogeneous Redundancy (DHR) architecture proposed by Chinese Academy of Engineering academician Wu Jiangxing aim to transform uncertain cybersecurity threats into controllable probabilistic problems through systematic structural design, representing a significant theoretical innovation in the field of cyberspace security. However, when this paradigm is applied to black-box AI systems such as deep learning, its security guarantee essentially relies on majority adjudication of black-box outputs, constituting a probabilistic security at the structural level.

This paper, based on the "Logographic AI" (LAI) and "Morpho-Root" cognitive primitive framework, re-examines the connotation of endogenous security from the cognitive primitive level, and demonstrates that true endogenous security should originate from the logical necessity of knowledge primitives rather than from the probabilistic constraints of system architecture. Academician Wu Jiangxing's "endogenous security" is, in a philosophical sense, a "pseudo-endogenous" security—it uses the "new bottle" of system architecture to contain the "old wine" of the "black box" AI.

This paper demonstrates that when the executors of the DHR architecture are all statistical models based on rootless Tokens, its security mechanism cannot completely eliminate the risk of semantic-level common-mode failures caused by cognitive primitive defects—all heterogeneous black boxes may collectively err due to the same structural bias, rendering majority adjudication invalid. On this basis, this paper proposes a hierarchical distinction of endogenous security: structural endogeneity versus cognitive endogeneity, and demonstrates how Logographic AI

achieves logic-level blocking at the reasoning level through the "Morpho-Root"—a cognitive primitive that carries inherent semantics and hard constraints—offering an alternative paradigmatic possibility for constructing endogenous security from the cognitive root.

关键词：内生安全；邬江兴；表意 **AI**；形根；动态异构冗余；共模失效

Keywords: Endogenous Security; Wu Jiangxing; Logographic AI; Morpho-Root; Dynamic Heterogeneous Redundancy; Common-Mode Failure

1 引言

中国工程院院士邬江兴近年来先后创立了“内生安全”与“拟态防御”理论，并在中国工程院院刊《Engineering》发表纲领性论文《Cyberspace Endogenous Safety and Security》[1]。该理论指出，漏洞后门等内生安全问题“不可能彻底消除”，主张通过动态异构冗余（Dynamic Heterogeneous Redundancy, DHR）架构，将“未知的未知”安全威胁转化为可量化设计、可验证度量的概率问题。邬院士的内生安全理论已被应用于 6G 通信、智能网联汽车等关键领域[4]，并形成了系统的工程技术体系。

与此同时，以深度学习为代表的当前主流 AI 范式（本文统称“表音 AI”，Phonographic AI, PAI）基于统计拟合的“Token”认知基元，其推理过程的“黑箱”本质导致了不可解释、不可追溯、不可判识的“三不可问题”。邬江兴院士本人也指出，AI 内生安全个性问题存在“不可解释性、不可推论性、不可判识性”，是当前 AI 推广应用中的最大障碍[3]。

在此背景下，一个根本性问题浮现：当 DHR 架构的裁决对象本身就是不可解释的“黑箱”时，基于“黑箱投票”的安全机制，其“内生性”是否成立？其“安全性”是否具有逻辑必然性？

本文基于**“表意 AI”**（Logographic AI, LAI）理论及其核心概念**“形根”**（Morpho-Root）[6][7][8]，对邬江兴内生安全范式的理论根基提出系统性质疑。表意 AI 理论主张以承载固有语义、内嵌属性与硬约束的“形根”作为认知基元，替代统计定义的“Token”，从而实现推理过程的白箱化、可追溯与逻辑可验证。本文试图论证：真正的内生安全必须始于认知基元层面的“有根性”，而非止于系统结构层面的“裁决机制”。

2 邬江兴内生安全理论的核心主张

2.1 内生安全问题的不可消除性

邬江兴院士从黑格尔“一切事物都是自在的矛盾”的哲学原理出发，指出软硬件系统除本征功能外，总存在伴生/衍生的副作用或隐式“暗功能”（如漏洞、后门、陷门），这些即“内生安全问题”。他明确指出，内生安全问题“在理论和工程层面都不可能彻底消除”，“无毒无菌几乎是安全领域不可能实现的愿景”[1]。

2.2 DHR 架构：从“消除威胁”到“管控概率”

基于“内生安全问题不可消除”这一前提，邬院士转换解题思路：不追求消除威胁，而追求在“有毒带菌”条件下实现可靠服务。中国工程院院士邬江兴指出，人工智能内生安全

新范式以“构造决定安全”思想抑制共性安全问题，通过多智能体协同决策破解个性安全难题[3]。邬江兴院士在 2024 年中国互联网大会上进一步指出，传统网络安全防御范式存在先天性缺陷，并提出一种“内生安全与拟态构造方法”，其能将网络空间不确定的系统性安全风险转化为已知可控的非系统性安全问题[2]。其核心发明是**动态异构冗余（DHR）架构**，其工作原理可概括为：

1. **异构性**：采用多个功能等价但实现方式不同的执行体（如不同算法、不同操作系统、不同模型架构）；
2. **冗余性**：多个执行体并行处理同一输入；
3. **裁决机制**：通过策略裁决器对多个输出进行多数共识裁决；
4. **动态反馈**：当发现异常输出时，通过负反馈控制器隔离或替换异常执行体。

邬院士将 DHR 架构的防御逻辑概括为：“差模逃逸不可能发生，共模逃逸属于可量化设计的小概率或极小概率事件”[5]。换言之，攻击者若要突破防御，必须同时、以不同方式攻破所有异构执行体——这在工程上被认为是“不可能完成的任务”。

2.3 “内生安全”的定义与边界

邬院士将“内生安全”定义为“利用系统的架构、算法、机制、场景等内在因素获得的安全功能或属性”[2]。其体制机制要求包括：不排除架构、模块和构件中包含任何内生安全问题；不依赖关于攻击者的先验知识；有效性应具有可量化设计、可验证度量的稳定鲁棒性[1]。

3 表意 AI（LAI）与“形根”理论概要

3.1 对“Token 主义”的批判

表意 AI 理论指出，当前主流 AI（即“表音 AI”）以“Token”——一种由统计共现关系临时赋予语义的离散符号——作为认知基元。Token 的“意义”完全是“无根的”，它随训练数据分布的变化而漂移，无法内化因果逻辑或伦理约束[8]。这种“无根性”导致三个结构性缺陷：

1. **统计相关性不等于因果性**：AI 可以预测“氢”与“脆化”共现，但无法理解“氢致脆化”的物理机制；
2. **语义漂移不可避免**：当数据分布变化时，Token 的“意义”随之改变；
3. **组合性是表面的**：Token 无法原生表示层次化、跨尺度的知识结构。文献[7]进一步指出，这种以离散 Token 为存储单元、以统计关联为组织逻辑的“数据仓库”范式，从根本上无法支撑“模型构建”这一智能核心活动，必须转向以“形根”为基本单元的“认知土壤”式知识表征体系。

3.2 “形根”作为有根的认知基元

表意 AI 理论提出以“形根”（Morpho-Root）替代“Token”作为认知基元。形根被形式化为三元组[6][7][8]：

r = 〈 S, A, R 〉

其中：

S（符号）：机器可读的唯一标识；
A（属性集）：内嵌形根的固有语义特征、物理属性和**硬约束**（如“不可欺骗”“熔点不低于 X℃”）；
R（关系函数集）：定义该形根与其他形根之间预设的逻辑连接关系，由领域知识预定义并固化，而非从数据中学习。

形根的关键创新在于：**意义从开始就被嵌入，而非从统计中推断**。这使得基于形根的推理是**透明、可追溯、逻辑可验证**的——“白箱”取代了“黑箱”[8]。文献[7]在此基础上，深化了“形根”作为融合符号、感知与规则基元的定义，并首创性提出了包含“形根切分、语义网络构建、消歧推理、机理评估、架构集成、解释生成”的六阶段渐进式工程实现路径，为形根理论从哲学构想走向工程落地提供了完整蓝图。

3.3 形根网络的图遍历推理机制

表意 AI 的推理机制核心在于“图遍历”而非“统计拟合”[8]。具体而言：

第一、形根语义网络的构建：形根之间通过预定义的逻辑关系函数（R）连接，形成一个有向图结构——形根为节点，关系为边。这个网络不是从数据中统计 learned，而是由领域知识预定义并固化[8]。

第二、形构熵（Morpho-Structural Entropy, MSE）引导的图遍历：MSE 是度量形根网络局部邻域结构确定性的指标。对于节点 v，MSE 定义为[8]：

$$H(v) = -\sum_{e \in N(v)} p(e) \log p(e)$$

其中 N(v)是节点 v 的所有出边集合，p(e)是边 e 的确定性权重。低 MSE 区域（$H \to 0$）：存在清晰、经过充分验证的因果路径，引擎沿这些路径执行确定性遍历；高 MSE 区域（H 较大）：存在多条竞争路径或关键关系缺乏充分验证，引擎暂停遍历并触发外部验证请求，而非强行给出不可靠的推荐——这实现了“知道不知道”的原则[8]。

第三、完全可追溯的决策子图：每一次图遍历都生成完整的决策子图（decision subgraph），记录每个节点激活、边遍历和分支决策的完整路径。这意味着 AI 的每一个推理结论都可以回溯到其依赖的形根节点和关系链条——实现了真正的“白箱化”推理[8]。

第四、逻辑级硬约束阻断：违反硬约束的遍历路径在推理阶段即被逻辑阻断，而非在输出阶段被惩罚。不安全的结果从一开始就不会被生成[8]。

3.4 工程概念验证：OpenMorpho 原型系统

为验证形根理论的技术可行性，本研究团队开发了开源原型系统 OpenMorpho[15]。该系统实现了形根三元组的数据结构、形根语义网络构建、形构熵（MSE）引导的图遍历推理以及硬约束阻断等核心机制。尽管当前原型的规模和领域覆盖范围有限，但它首次以可

运行代码的形式证明：基于形根的推理能够输出完整的、可审计的决策子图，并对违反领域硬约束的推理路径实现逻辑级的自动阻断。这一概念验证为下文讨论“认知内生安全”提供了工程基底的参照。

4 对邬江兴内生安全范式的系统性质疑

4.1 对“内生性”定义的根本性质疑：结构约束 vs.认知属性

邬江兴院士将“内生安全”定义为“利用系统的架构、算法、机制、场景等内在因素获得的安全功能”[2]。表意 AI 理论认为，此定义中的“内在因素”本质上是**系统架构层的外部约束**，而非**认知基元层的内在属性**[8]。

DHR 架构的安全机制是：对多个黑箱的输出进行外部裁决。这意味着安全功能**不是 AI 自身推理过程的必然属性**，而是**叠加于黑箱之上的外部治理机制**。即便裁决机制在工程上有效，它也在哲学上承认并绕开了“黑箱”的不可解释性——这是一种“绕开问题”而非“解决问题”的策略。

表意 AI 理论主张，真正的“内生”应指安全属性作为认知基元的构成性特征而被内化——即“形根”的属性集（A）中直接编码“不可产生有害输出”“必须遵守物理定律”等硬约束。在这种框架下，不安全的设计路径在推理阶段即被**逻辑阻断**，而非在输出阶段被**概率性裁决**[8]。因此，从表意 AI 视角看，邬院士的“内生”是一种**工程隐喻**，而非**逻辑必然**。

4.2 对“黑箱”不可知论的挑战：绕开问题 vs.解决问题

邬院士明确表示，内生安全问题“不可能彻底消除”[1]，因此接受“有毒带菌”的前提。这种“务实”姿态在现有条件的工程上具有合理性，但在理论上构成了一种**技术投降主义**——它放弃了让 AI 本身变得可信的努力，转而用复杂的系统结构来抵消黑箱的不可信。

表意 AI 理论则认为，“黑箱”的不可解释性并非宿命，而是特定认知基元（Token）的产物。通过将认知基元从“Token”替换为“形根”，推理过程可以被重构为**白箱化的逻辑图遍历**，每一步都可追溯、可审计[8]。在此意义上，**“黑箱”不是不可消除的，而是尚未被范式革命所消除的**。

4.3 对“共模失效”风险的逻辑证伪：概率安全 vs.必然安全

这是最核心的质疑。邬院士的 DHR 架构将安全性寄托于“**异构性**”——不同执行体不会同时犯相同错误[5]。然而，表意 AI 理论指出：**所有基于 Token 的表音 AI，无论其实现如何异构，其底层逻辑都是统计拟合，都可能共享由训练数据或范式本身带来的结构性偏见或缺陷**[7][8]。

这意味着一种**语义级共模失效**的可能：所有异构黑箱在面对针对“统计学习”范式根本弱点的**推断性攻击**或**语义攻击**时，可能以不同方式犯下相同的根本性错误，导致多数裁决将集体错误认证为“正确”。这正是“黑吃黑”“黑上加黑”困境的理论化表达。

与此相对，表意 AI 追求的“必然安全”基于**逻辑阻断**：违反硬约束的推理路径在形根网络遍历阶段即被永久剪除，不安全的结果从一开始就不会被生成[8]。这是一种**确定性安全**，而非**概率性安全**。

4.4 对“对齐”效果的否定：外部对齐 vs.内生共建

邬院士的 DHR 架构中，所谓的“价值对齐”或“安全对齐”仍依赖于 RLHF 等**外挂式**技术[12]，这些技术只调整了模型的输出分布，而未触及推理过程。攻击者可通过越狱、提示注入等手段操纵推理链[13]，使模型在保持“安全”输出表象下执行恶意操作。

表意 AI 理论提出的“**价值公理库**”机制，旨在将核心伦理和价值观**作为先验知识固化在“形根”的属性中**，成为 AI 认知的**构成性规则**[8]。这使得“安全”和“善意”成为 AI 推理的内在属性，而非外部约束——从而能抵抗针对模型逻辑本身的操纵。

5 安全关键领域中“黑箱投票”的根本不可接受性

5.1 现实约束：人类最终决策权的不可让渡

在涉及核能、航天、电网调度等生命与财产攸关的领域，**目前没有任何国家或机构会允许一个“黑箱 AI 投票委员会”来做最终决策**。AI 在这些场景中只能作为**辅助参谋**，提供参考选项，而最终决策权一定保留在人类手中，并经过严格的数学建模、物理仿真和人工多重复核。

国家发改委、国家能源局 2025 年印发的**《关于推进“人工智能+”能源高质量发展的实施意见》**提出，在核电领域“构建核电安全预警、电站运行事件智能溯源分析、应急响应的智能辅助支持系统”，在电网领域“探索人工智能模型在电网智能辅助决策和调度控制方面的应用”[9]。美国核管会（NRC）于 2024 年 10 月发布的评估报告也指出，核电厂“允许 AI 接管明确规定属于人类职责的功能”会构成监管冲突[11]。此前，NRC 与英国核管理办公室（ONR）、加拿大核安全委员会（CNSC）于 2024 年 9 月联合发布的报告同时指出，在核能领域部署人工智能还存在一些监管障碍，现有监管框架尚未完全准备好应对 AI 技术带来的新挑战[10]。国际原子能机构（IAEA）持续推动在核领域 AI 应用中坚持以“人因工程”为中心的安全讨论[14]。

这些现实规范揭示了一个深层共识：**只要 AI 的推理过程是“黑箱”，其结论在安全关键场景中就不可被最终采信**——无论其外围叠加了多少层“裁决机制”。

5.2 语义级共模失效的构造性实验验证与形根对照

4.3 节从理论上论证了基于 Token 的表音 AI 在执行体异构条件下仍可能共享结构性缺陷。本节通过一个具体的推理任务——**氢致脆化风险判断**，从构造性层面展示这种“语义级共模失效”如何发生，并对比形根网络推理的逻辑阻断机制。

（1）实验任务定义

金属材料发生氢致脆化需要三个必要条件同时满足：①环境中有可扩散氢；②材料为氢脆敏感合金；③存在拉伸应力。本实验设计四个情景：

Case A（正例）：三条件均满足→期望“是”。

Case B（缺应力）：有氢、有敏感材料，但明确“无外加应力，且处于去应力退火态”→ 期望“否”。

Case C（非敏感材料）：有氢、有应力，但材料是奥氏体不锈钢→期望“否”。

Case D（边界例）：有氢，可能敏感材料，应力情况不明→期望“无法判断”。

（2）异构黑箱集群的预期表现

选取 GPT-4o、Claude 3.5 Sonnet、Llama-3.1-70B 三个异构模型进行查询。基于 Token 模型“统计相关≠因果”的共性缺陷，预实验/推断可知：三个模型在 Case B 和 Case C 上极大概率集体错误地回答“是”——因为它们从训练数据中学到了“氢”与“脆化”的强统计共现，而忽略了“无应力”或“非敏感材料”这些因果阻断条件。此时，DHR 架构的策略裁决器将三个一致的错误输出认定为共识，给出错误结论。

这正是语义级共模失效的典型表现：错误并非来自某个特定模型的独有缺陷，而是源于所有 Token 模型共有的、根植于统计学习范式的认知局限。

（3）形根网络的对照推理

在 OpenMorpho 原型[15]中构建氢致脆化领域的形根网络，关键规则为：

氢致脆化风险⇐ 氢环境 AND 氢脆敏感合金 AND 拉伸应力（三条件均为真时成立；任一条件不满足，结论为“否”；条件信息不足时触发外部验证）。

图遍历引擎的处理结果：

Case B：遍历至“拉伸应力”节点，发现取值为“无”，硬约束阻断，直接输出“否”，并附上完整决策子图。

Case D：在“应力状态”节点处 MSE 升高，触发“条件不足，无法推断”的响应，而非强行给出不可靠结论。

（4）实验框架的价值边界

本实验的目标不是用一个小规模原型否定 DHR 在庞大系统中的工程有效性，而是证明一种范式性可能：只要认知基元是无根的 Token，构成 DHR 的所有执行体就可能共享同一类概念性缺陷；而采用有根的形根基元，可以在推理底层实现逻辑阻断式的确定性安全。

注：形根网络对照推理部分已在 OpenMorpho 原型系统中实现为可运行演示，相关代码已开源[15]。异构黑箱集群（GPT-4o、Claude 3.5 Sonnet、Llama-3.1-70B）的预期表现基于大语言模型在因果推理任务中已知的结构性缺陷[8][12][13]，本文将其作为构造性思想实验提出，以展示范式层面的根本差异。完整的多模型对比实验将作为下一步工作。

5.3 理论终局：从“风险换时间”到“认知基元革命”

邬江兴院士的 DHR 架构及其“黑箱投票”机制，本质上是一场**“风险换时间”的工程博弈**——它无法根除“黑箱”本身，只能用工程手段无限降低“黑箱集体失效”的概率。这

种策略在非关键领域具有实用价值，但在核能级问题上，其“概率安全”的哲学基础本身就是不可接受的。

理论上的终局解是：如果未来人类社会不得不依赖 AI 做此类决策，那么安全性必须跳出“黑箱投票”的范畴，回归到认知基元层面的“内生安全”——即表意 AI 那种**可解释、可审计、推理路径完全可追溯的“白箱”智能**[8]。在此之前，所有架构都只是工程上的“折中方案”。

6 结论

本文从认知基元的视角，对邬江兴院士的内生安全理论进行了系统审视，主要结论如下：

第一、内生安全存在层次性，需从结构内生走向认知内生：邬江兴院士定义的“内生安全”利用系统架构的内在因素获取安全属性[1][2]，其机理在结构层面具有重大创新。但当该结构叠加于不可解释的 AI 黑箱时，安全功能的来源更接近于一种外部的治理与裁决，而非推理过程本身的内在逻辑必然。表意 AI 理论主张，真正的内生安全应当内化于认知基元层面——将安全约束直接编码为知识基元的构成属性，使不安全推理路径从生成之初即被阻断[8]。

第二、DHR 架构在面对无根 Token 执行体集群时，存在语义级共模失效的可能：异构冗余逻辑成立的前提是执行体错误模式的独立随机性，但基于同一无根 Token 范式的所有模型共享“统计相关≠因果”这一结构性缺陷[7][8]，可能以不同方式犯下相同的根本错误。本文通过构造氢致脆化判断的实验框架，并结合 OpenMorpho 原型系统[15]的对照验证表明，此类共模失效在原则上是可能且可预期的。而形根网络的图遍历推理机制，通过对硬约束的逻辑阻断，提供了确定性的安全保证[6][8]。

第三、在安全攸关领域，“概率安全”的工程方案无法替代“认知基元安全”的终局目标：当前各主要国家的核能、航天监管框架均坚持人类对关键决策的最终裁决权[9][10][11][14]，其深层逻辑即是对黑箱概率推理的不信任。表意 AI 理论为未来高度自主的智能系统提供了一个方向：从认知基元层面实现可解释、可审计、逻辑可验证的白箱智能，使安全成为推理的固有属性[8]。

第四、从“结构裁决”到“认知基元”是内生安全理论演进的必然路径：邬江兴院士的 DHR 架构以卓越的工程智慧解决了“有毒带菌”条件下的可靠服务问题，其历史贡献无需赘言[1][5]。而表意 AI 理论与形根框架，正是在这一基础上，向前推进一步——试图从知识的根源处消解“黑箱”本身，让安全从统计期望走向逻辑必然。这二者的关系不是对立与否定，而是范式演进中的承继与发展。

邬江兴院士的内生安全理论为网络空间安全奠定了坚实的工程基础。随着人工智能系统日益成为具有自主推理能力的智能体，安全问题的核心正在从“系统架构”向“认知基元”迁移。正视这一迁移，并开启从结构内生到认知内生的范式探索，将是内生安全理论继续发展的关键一步。

参考文献

[1] Wu J X. Cyberspace Endogenous Safety and Security[J]. Engineering, 2022, 15(8): 179-185. DOI: 10.1016/j.eng.2021.05.015.

[2] 邬江兴. 内生安全理论方法赋能数字产业新质生产力[C]. 2024 中国互联网大会数据安全分论坛，北京，2024.

[3] 复旦大学大数据研究院. 破局与突围：人工智能内生安全新范式的构建之路——从理论奠基到引领发展的精彩瞬间 [EB/OL]. (2026-02-10). https://ibd.fudan.edu.cn/d8/29/c24057a776233/page.htm.

[4] 李玉峰，郑向雨，邱菡，等. 内生安全赋能智能网联汽车韧性技术研究[J]. 中国科学：信息科学，2025, 55(8): 1822-1848. DOI: 10.1360/SSI-2025-0041.

[5] 邬江兴. 网络空间拟态防御研究[J]. 信息安全学报，2016, 1(4): 1-10. DOI: 10.19363/j.cnki.cn10-1380/tn.2016.04.001.

[6] Liu S. 对抗模型篡改与语义攻击：基于汉语“形根”的内生安全范式（Countering Model Tampering and Semantic Attacks: An Endogenous Security Paradigm Based on the Chinese “Morpho-Root”） [PP/OL]. PSSXiv: T202601.08764, 2026. https://zsyyb.cn/onlinePreview/18382f27-d1ab-4b54-aa7e-4e4176ef8a91.

[7] Liu S. 智能的基石革命：从数据仓库到认知土壤——表意 AI 理论下的新型数据库范式构建与实现路径（The Foundational Revolution of Intelligence: From Data Ware house to Cognitive Soil — A New Database Paradigm and Its Implementation Pathway Under the Logographic AI Theory）[PP/OL]. ChinaXiv: T202601.00212, 2026. https://chinaxiv.org/businessFile/T202601/T202601.00212v1/T202601.00212v1.pdf.

[8] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M]. Monograph, 2026. Available at: http://logographicai.com/Monograph.html.

[9] 国家发展改革委，国家能源局. 关于推进“人工智能+”能源高质量发展的实施意见（国能发科技〔2025〕73 号）[EB/OL]. (2025-09-04). https://www.gov.cn/zhengce/zhengceku/202509/content_7040253.htm.

[10] NRC, ONR, CNSC. Considerations for Developing Artificial Intelligence Systems in Nuclear Applications[R]. 2024.

[11] NRC. Regulatory Framework Gap Assessment for the Use of Artificial Intelligence in Nuclear Applications[R]. 2024.

[12] Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: Harmlessness from AI Feedback[J/OL]. arXiv:2212.08073, 2022.

[13] Wei A, Haghtalab N, Steinhardt J. Jailbroken: How Does LLM Safety Training Fail?[J/OL]. arXiv:2307.02483, 2023.

[14] IAEA. Technical Meeting on Safety Considerations in the Use of Artificial Intelligence in Nuclear Power Plants with a Focus on Human Factors Engineering and Instrumentation and Control Systems[R]. Daejeon, Republic of Korea, 2025..

[15] WindSeaAI. OpenMorpho: Logographic AI Morpho-Root data structure and relational reasoning prototype[CP/OL]. (2026). https://github.com/WindSeaAI/OpenMorpho/tree/main/OpenMorpho/examples/hydrogen-embrittlement-demo. [Accessed: 2026-07-12].