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

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.

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

1 Introduction

In recent years, Chinese Academy of Engineering academician Wu Jiangxing has successively established the theories of "endogenous security" and "mimic defense," and published the programmatic paper Cyberspace Endogenous Safety and Security in Engineering, the official journal of the Chinese Academy of Engineering [1]. This theory points out that endogenous security problems such as vulnerabilities and backdoors "cannot be completely eliminated," and advocates transforming "unknown unknowns" security threats into quantifiable design and verifiable probabilistic problems through the Dynamic Heterogeneous Redundancy (DHR) architecture. Academician Wu's endogenous security theory has been applied to critical fields such as 6G

communications and intelligent connected vehicles [4], and has formed a systematic engineering and technical system.

At the same time, the current mainstream AI paradigm represented by deep learning (collectively referred to in this paper as "Phonographic AI," PAI) is based on the statistically fitted "Token" cognitive primitive, and the "black box" nature of its reasoning process leads to the "three impossibilities" problem: inexplicability, untraceability, and unaccountability. Academician Wu Jiangxing himself has pointed out that the endogenous security individuality problems of AI—namely "inexplicability, non-inferability, and unaccountability"—are the greatest obstacles to the current promotion and application of AI [3].

Against this background, a fundamental question emerges: when the objects of adjudication under the DHR architecture are themselves inexplicable "black boxes," does the "endogeneity" of a security mechanism based on "black box voting" hold? Does its "security" possess logical necessity?

This paper, based on the "Logographic AI" (LAI) theory and its core concept of the "Morpho-Root" [6][7][8], systematically questions the theoretical foundations of Wu Jiangxing's endogenous security paradigm. Logographic AI theory advocates replacing the statistically defined "Token" with the "Morpho-Root" as the cognitive primitive—which carries inherent semantics, embedded attributes, and hard constraints—thereby achieving white-box, traceable, and logically verifiable reasoning. This paper attempts to demonstrate that true endogenous security must begin with "rootedness" at the cognitive primitive level, rather than stopping at "adjudication mechanisms" at the systemic structural level.

2 Core Propositions of Wu Jiangxing's Endogenous Security Theory

2.1 The Ineliminability of Endogenous Security Problems

Starting from Hegel's philosophical principle that "all things are contradictions in themselves," Academician Wu Jiangxing points out that beyond their intrinsic functions, software and hardware systems always have accompanying/derivative side effects or implicit "dark functions" (such as vulnerabilities, backdoors, trapdoors), which constitute "endogenous security problems." He explicitly states that endogenous security problems "cannot be completely eliminated at either the theoretical or engineering level," and that "a sterile, virus-free state is almost an unattainable vision in the field of security" [1].

2.2 DHR Architecture: From "Eliminating Threats" to "Managing Probabilities"

Based on the premise that "endogenous security problems cannot be eliminated," Academician Wu shifts his problem-solving approach: rather than pursuing the elimination of threats, he pursues the realization of reliable services under "toxic and pathogen-carrying" conditions. Academician Wu

Jiangxing points out that the new paradigm of AI endogenous security uses the idea of "structure determines security" to suppress common security problems, and solves individual security problems through multi-agent collaborative decision-making [3]. At the 2024 China Internet Conference, Academician Wu Jiangxing further pointed out that traditional cybersecurity defense paradigms have inherent defects, and proposed an "endogenous security and mimic construction method" that can transform uncertain systemic security risks in cyberspace into known and controllable non-systemic security problems [2]. Its core invention is the Dynamic Heterogeneous Redundancy (DHR) architecture, whose working principles can be summarized as follows:

1. **Heterogeneity**: employing multiple functionally equivalent but differently implemented executors (e.g., different algorithms, different operating systems, different model architectures);
2. **Redundancy**: multiple executors process the same input in parallel;
3. **Adjudication mechanism**: performing majority consensus adjudication on multiple outputs through a policy adjudicator;
4. **Dynamic feedback**: when anomalous outputs are detected, isolating or replacing abnormal executors through a negative feedback controller.

Academician Wu summarizes the defense logic of the DHR architecture as follows: "Differential-mode escape is impossible, and common-mode escape is a small or extremely small probability event that can be quantifiably designed" [5]. In other words, to breach the defense, an attacker must simultaneously breach all heterogeneous executors in different ways—which is considered an "impossible mission" in engineering.

2.3 Definition and Boundaries of "Endogenous Security"

Academician Wu defines "endogenous security" as "security functions or attributes obtained through intrinsic factors of the system such as architecture, algorithms, mechanisms, and scenarios" [2]. Its institutional and mechanistic requirements include: not excluding any endogenous security problems contained in architectures, modules, and components; not relying on prior knowledge about attackers; and effectiveness should possess quantifiable design and verifiable, stable robustness [1].

3 Overview of Logographic AI (LAI) and "Morpho-Root" Theory

3.1 Critique of "Tokenism"

Logographic AI theory points out that the current mainstream AI (i.e., "Phonographic AI") uses the "Token"—a discrete symbol whose semantics are temporarily conferred by statistical co-occurrence relationships—as its cognitive primitive. The "meaning" of the Token is entirely "rootless," drifting with changes in the training data distribution and incapable of internalizing causal logic or ethical constraints [8]. This "rootlessness" leads to three structural deficiencies:

1. Statistical correlation is not causality: AI can predict the co-occurrence of "hydrogen" and "embrittlement," but cannot understand the physical mechanism of "hydrogen embrittlement";
2. Semantic drift is unavoidable: when the training data distribution shifts, the "meaning" of the Token changes accordingly;
3. Compositionality is superficial: Token cannot natively represent hierarchical, cross-scale knowledge structures. Reference [7] further points out that this "data warehouse" paradigm—which uses discrete Tokens as storage units and statistical associations as organizational logic—fundamentally cannot support "model construction," the core activity of intelligence, and must shift to a "cognitive soil" knowledge representation system with the "Morpho-Root" as the basic unit.

3.2 The "Morpho-Root" as a Rooted Cognitive Primitive

Logographic AI theory proposes replacing the "Token" with the "Morpho-Root" as the cognitive primitive. The Morpho-Root is formalized as a triple [6][7][8]:

r = ⟨ S, A, R⟩

Where:

- **S (Symbol)**: a machine-readable unique identifier;
- **A (Attributes)**: an attribute set embedding the Morpho-Root's inherent semantic features, physical properties, and hard constraints (e.g., "cannot deceive," "melting point not lower than X°C");
- **R (Relation Functions)**: a set of relation functions defining the presupposed logical connections between this Morpho-Root and other Morpho-Roots, predefined by domain knowledge and hardwired rather than learned from data.

The key innovation of the Morpho-Root is that **meaning is embedded from the start, rather than inferred from statistics**. This makes reasoning based on the Morpho-Root transparent, traceable, and logically verifiable—the "white box" replaces the "black box" [8]. On this basis, reference [7] deepens the definition of the "Morpho-Root" as a primitive integrating symbols, perception, and rules, and proposes a pioneering six-stage incremental engineering implementation pathway—including "Morpho-Root segmentation, semantic network construction, disambiguation reasoning, mechanism evaluation, architecture integration, and explanation generation"—providing a complete blueprint for moving Morpho-Root theory from philosophical conception to engineering deployment.

3.3 The Graph-Traversal Reasoning Mechanism of the Morpho-Root Network

The core of Logographic AI's reasoning mechanism lies in "graph traversal" rather than "statistical fitting" [8]. Specifically:

First, construction of the Morpho-Root semantic network: Morpho-Roots are connected through predefined logical relation functions (R), forming a directed graph structure—with Morpho-Roots as nodes and relations as edges. This network is not statistically learned from data, but predefined and hardwired by domain knowledge [8].

Second, Morpho-Structural Entropy (MSE)-guided graph traversal: MSE is a metric for measuring the structural determinacy of a node's local neighborhood in the Morpho-Root network. For node v, MSE is defined as [8]:

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

Where N(v) is the set of all outgoing edges from node v, and p(e) is the certainty weight of edge e. **Low MSE regions** ($H \to 0$): there exist clear, well-validated causal paths, and the engine performs deterministic traversal along these paths; **high MSE regions** (H large): there exist multiple competing paths or key relations lacking sufficient validation, and the engine pauses traversal and triggers external validation requests rather than forcing unreliable recommendations—this implements the principle of "knowing what it does not know" [8].

Third, fully traceable decision subgraphs: each graph traversal generates a complete decision subgraph, recording the full path of every node activation, edge traversal, and branching decision. This means that every reasoning conclusion of the AI can be traced back to the Morpho-Root nodes and relation chains it depends on—achieving truly "white-box" reasoning [8].

Fourth, logic-level hard constraint blocking: traversal paths that violate hard constraints are logically blocked at the reasoning stage, rather than penalized at the output stage. Unsafe outcomes are never generated from the very beginning [8].

3.4 Engineering Concept Validation: The OpenMorpho Prototype System

To validate the technical feasibility of Morpho-Root theory, our research team has developed the open-source prototype system OpenMorpho [15]. This system implements core mechanisms including the Morpho-Root triple data structure, Morpho-Root semantic network construction, Morpho-Structural Entropy (MSE)-guided graph traversal reasoning, and hard constraint blocking. Although the current prototype's scale and domain coverage are limited, it demonstrates for the first time in runnable code that Morpho-Root-based reasoning can output complete, auditable decision subgraphs and achieve logic-level automatic blocking of inference paths that violate domain hard constraints. This concept validation provides an engineering reference for the subsequent discussion of "cognitive endogenous security."

4 Systematic Questioning of Wu Jiangxing's Endogenous Security Paradigm

4.1 Fundamental Questioning of the Definition of "Endogeneity": Structural Constraints vs. Cognitive Attributes

Academician Wu Jiangxing defines "endogenous security" as "security functions obtained through intrinsic factors of the system such as architecture, algorithms, mechanisms, and scenarios" [2]. Logographic AI theory argues that the "intrinsic factors" in this definition are essentially external constraints at the system architecture level, rather than internal attributes at the cognitive primitive level [8].

The security mechanism of the DHR architecture is: external adjudication of the outputs of multiple black boxes. This means that the security function is not a necessary attribute of the AI's own reasoning process, but an external governance mechanism superimposed upon the black box. Even if the adjudication mechanism is effective in engineering, it philosophically acknowledges and circumvents the inexplicability of the "black box"—this is a strategy of "circumventing the problem" rather than "solving the problem."

Logographic AI theory maintains that true "endogeneity" should mean that security attributes are internalized as constitutive features of the cognitive primitive—that is, directly encoding hard constraints such as "cannot produce harmful outputs" and "must comply with physical laws" in the attribute set (A) of the "Morpho-Root." Under this framework, unsafe design paths are **logically blocked** at the reasoning stage, rather than **probabilistically adjudicated** at the output stage [8]. Therefore, from the perspective of Logographic AI, Academician Wu's "endogeneity" is an **engineering metaphor**, not a **logical necessity**.

4.2 Challenge to "Black Box" Agnosticism: Circumventing vs. Solving

Academician Wu explicitly states that endogenous security problems "cannot be completely eliminated" [1], thus accepting the premise of being "toxic and pathogen-carrying." This "pragmatic" stance is reasonable in engineering under current conditions, but constitutes a kind of **technological defeatism** in theory—it abandons the effort to make AI itself trustworthy, and instead uses complex system structures to offset the untrustworthiness of the black box.

Logographic AI theory holds that the inexplicability of the "black box" is not fate, but the product of a specific cognitive primitive (Token). By replacing the "Token" with the "Morpho-Root" as the cognitive primitive, the reasoning process can be reconstructed as **white-box logical graph traversal**, with every step traceable and auditable [8]. In this sense, the "black box" is not ineliminable, but has yet to be eliminated by a paradigm revolution.

4.3 Logical Falsification of "Common-Mode Failure" Risk: Probabilistic Security vs. Necessary Security

This is the most fundamental critique. Academician Wu's DHR architecture bases security on "heterogeneity"—the assumption that different executors will not make the same mistake simultaneously [5]. However, Logographic AI theory points out: **all Token-based Phonographic AIs, regardless of how heterogeneous their implementations, share the same underlying logic of statistical fitting, and may share structural biases or defects originating from the training data or paradigm itself** [7][8].

This suggests the possibility of a **semantic-level common-mode failure**: when faced with inferential or semantic attacks targeting the fundamental weaknesses of the "statistical learning" paradigm, all heterogeneous black boxes may make the same fundamental error in different ways, causing the majority adjudication to certify a collective error as "correct." This is the theoretical expression of the **"black-box-devouring-black-box"** predicament.

In contrast, the **"necessary security"** pursued by Logographic AI is based on **logical blocking**: reasoning paths that violate hard constraints are permanently pruned at the Morpho-Root network traversal stage, so unsafe outcomes are never generated from the beginning [8]. This is **deterministic security**, rather than **probabilistic security**.

4.4 Negation of "Alignment" Effectiveness: External Alignment vs. Endogenous Co-construction

In Academician Wu's DHR architecture, so-called "value alignment" or "safety alignment" still relies on external technologies such as RLHF [12], which only adjust the output distribution of the model without touching the reasoning process. Attackers can manipulate the reasoning chain through means such as jailbreaking and prompt injection [13], causing the model to execute malicious operations while maintaining a "safe" output appearance.

The "Value Axiom Library" mechanism proposed by Logographic AI theory aims to solidify core ethics and values as a priori knowledge in the attributes of the "Morpho-Root," making them **constitutive rules** of AI cognition [8]. This makes "security" and "benevolence" intrinsic properties of AI reasoning rather than external constraints—thereby resisting manipulation of the model's logic itself.

5 The Fundamental Unacceptability of "Black Box Voting" in Safety-Critical Domains

5.1 Real-world Constraints: The Non-transferability of Human Final Decision-making Authority

In domains involving nuclear energy, aerospace, power grid dispatch, and other life-and-property-critical areas, **no country or institution currently permits a "black box AI voting committee" to make final decisions**. In these scenarios, AI can only serve as an **assistant consultant**, providing reference options, while final decision-making authority must remain with humans, subject to rigorous mathematical modeling, physical simulation, and multi-layered human re-verification.

The "Implementation Opinions on Promoting High-Quality Development of 'AI+' Energy," issued in 2025 by the National Development and Reform Commission and the National Energy Administration, proposes in the nuclear power domain to "build intelligent auxiliary support systems for nuclear safety early warning, intelligent traceability analysis of power plant operational events, and emergency response," and in the power grid domain to "explore the application of AI models in intelligent auxiliary decision-making and dispatch control for the power grid" [9]. The assessment report issued by the U.S. Nuclear Regulatory Commission (NRC) in October 2024 also points out that nuclear power plants "allowing AI to take over functions explicitly designated for human responsibility" would constitute a regulatory conflict [11]. Previously, the joint report issued in September 2024 by the NRC, the UK Office for Nuclear Regulation (ONR), and the Canadian Nuclear Safety Commission (CNSC) pointed out that there are still regulatory obstacles to deploying artificial intelligence in the nuclear energy sector, and that the existing regulatory framework is not yet fully prepared to address the new challenges posed by AI technology [10]. The International Atomic Energy Agency (IAEA) continues to promote safety discussions centered on "human factors engineering" in AI applications in the nuclear field [14].

These real-world norms reveal a deep consensus: as long as the AI's reasoning process is a "black box," its conclusions **cannot be ultimately relied upon** in safety-critical scenarios—regardless of how many layers of "adjudication mechanisms" are superimposed on the periphery.

5.2 Constructive Experimental Validation of Semantic Common-Mode Failure and Morpho-Root Contrast

Section 4.3 theoretically demonstrated that Token-based Phonographic AIs may still share structural defects even under heterogeneous executor conditions. This section uses a concrete reasoning task—hydrogen embrittlement risk assessment—to constructively demonstrate how this "semantic common-mode failure" occurs, and contrasts the logic-blocking mechanism of Morpho-Root network reasoning.

(1) Experimental Task Definition

Hydrogen embrittlement of metallic materials requires three necessary conditions to be simultaneously satisfied: (1) diffusible hydrogen in the environment; (2) the material is a hydrogen-embrittlement-susceptible alloy; (3) tensile stress exists. Four scenarios are designed:

- Case A (positive): all three conditions satisfied → expected "Yes."

- Case B (missing stress): hydrogen and susceptible material present, but explicitly "no applied stress, in stress-relief annealed state" → expected "No."
- Case C (non-susceptible material): hydrogen and stress present, but the material is austenitic stainless steel → expected "No."
- Case D (boundary): hydrogen present, possibly susceptible material, stress state unknown → expected "Cannot Determine."

(2) Expected Performance of Heterogeneous Black Box Clusters

Three heterogeneous models—GPT-4o, Claude 3.5 Sonnet, and Llama-3.1-70B—are selected for querying. Based on the common defect of Token models—"statistical correlation ≠ causation"—prior analysis and inference indicate that all three models are highly likely to collectively and incorrectly answer "Yes" on Case B and Case C—because they have learned the strong statistical co-occurrence of "hydrogen" and "embrittlement" from training data, while ignoring causal blocking conditions such as "no stress" or "non-susceptible material." At this point, the DHR architecture's policy adjudicator will recognize the three consistent erroneous outputs as consensus, yielding an incorrect conclusion.

This is a typical manifestation of semantic common-mode failure: the error does not originate from a unique defect of any specific model, but from the cognitive limitation rooted in the statistical learning paradigm shared by all Token models.

(3) Contrastive Reasoning of the Morpho-Root Network

A hydrogen embrittlement Morpho-Root network is constructed in the OpenMorpho prototype [15], with the core rule:

> HE_Risk ⇐ Hydrogen_Environment AND Susceptible_Alloy AND Tensile_Stress (risk holds only when all three are true; if any condition is false, conclusion is "No"; if information is insufficient, external validation is triggered).

The graph traversal engine's processing results:

- Case B: upon traversing to the "Tensile Stress" node, the value is found to be "absent" → hard constraint blocking, directly outputs "No," accompanied by a complete decision subgraph.
- Case D: MSE rises at the "Stress State" node, triggering a "insufficient conditions, cannot infer" response, rather than forcing an unreliable conclusion.

(4) Value Boundaries of the Experimental Framework

The goal of this experiment is not to use a small-scale prototype to negate the engineering effectiveness of DHR in large-scale systems, but to demonstrate a paradigmatic possibility: as long as the cognitive primitive is the rootless Token, all executors constituting DHR may share the same type of conceptual defect; whereas adopting the rooted Morpho-Root primitive can achieve logic-blocking deterministic security at the reasoning level.

Note: The Morpho-Root network contrastive reasoning part has been implemented as a runnable demonstration in the OpenMorpho prototype system, and the relevant code has been open-sourced [15]. The expected performance of the heterogeneous black box cluster (GPT-4o, Claude 3.5 Sonnet, Llama-3.1-70B) is based on known structural defects of large language models in causal reasoning tasks [8][12][13]; this paper presents it as a constructive thought experiment to demonstrate fundamental differences at the paradigmatic level. A complete multi-model comparative experiment is left as future work.

5.3 Theoretical Final Outcome: From "Risk for Time" to "Cognitive Primitive Revolution"

Academician Wu Jiangxing's DHR architecture and its "black box voting" mechanism are essentially a **"risk for time"** engineering game—it cannot eliminate the "black box" itself, but can only use engineering means to infinitely reduce the probability of "collective black box failure." This strategy has practical value in non-critical domains, but on nuclear-level problems, its philosophical foundation of "probabilistic security" is itself unacceptable.

The **theoretical final outcome** is: if human society must rely on AI for such decision-making in the future, then security must move beyond "black box voting" and return to "endogenous security" at the **cognitive primitive level**—that is, the interpretable, auditable, and fully traceable **"white box"** intelligence of Logographic AI [8]. Until then, all architectures are merely engineering **"compromises."**

6 Conclusion

This paper, from the perspective of cognitive primitives, has systematically examined Academician Wu Jiangxing's endogenous security theory. The main conclusions are as follows:

First, endogenous security exhibits a hierarchical nature, requiring a transition from structural endogeneity to cognitive endogeneity. The "endogenous security" defined by Academician Wu Jiangxing utilizes intrinsic factors of system architecture to obtain security attributes [1][2], and its mechanism represents a significant innovation at the structural level. However, when this structure is superimposed upon black-box AI black boxes, the source of security functions is closer to an external governance and adjudication than to the intrinsic logical necessity of the reasoning process itself. Logographic AI theory maintains that true endogenous security should be internalized at the cognitive primitive level—encoding security constraints directly as constitutive attributes of knowledge primitives, so that unsafe reasoning paths are blocked from the very moment of generation [8].

Second, the DHR architecture, when faced with clusters of rootless Token-based executors, is susceptible to semantic common-mode failure. The premise for the validity of heterogeneous redundancy logic is the independent randomness of executor error patterns, but all models based on the same rootless Token paradigm share the structural defect of "statistical correlation ≠

causation" [7][8], and may commit the same fundamental error in different ways. Through the constructed experimental framework for hydrogen embrittlement judgment, combined with the contrastive validation of the OpenMorpho prototype system [15], this paper demonstrates that such common-mode failure is in principle possible and predictable. The graph-traversal reasoning mechanism of the Morpho-Root network, through logic-level blocking of hard constraints, provides deterministic security guarantees [6][8].

Third, in safety-critical domains, the "probabilistic security" engineering approach cannot replace the ultimate goal of "cognitive primitive security." The regulatory frameworks of major countries in nuclear energy and aerospace all insist on human final decision-making authority [9][10][11][14], and the underlying logic is precisely distrust of black-box probabilistic reasoning. Logographic AI theory offers a direction for future highly autonomous intelligent systems: achieving interpretable, auditable, and logically verifiable white-box intelligence from the cognitive primitive level, making security an inherent attribute of reasoning [8].

Fourth, the transition from "structural adjudication" to "cognitive primitive" is the inevitable path for the evolution of endogenous security theory. Academician Wu Jiangxing's DHR architecture, with its outstanding engineering wisdom, solved the problem of reliable service under "toxic and pathogen-carrying" conditions, and its historical contribution is beyond dispute [1][5]. Logographic AI theory and the Morpho-Root framework, building upon this foundation, take a step further—attempting to dissolve the "black box" itself from the root of knowledge, moving security from statistical expectation to logical necessity. The relationship between the two is not one of opposition and negation, but of inheritance and development within paradigmatic evolution.

Academician Wu Jiangxing's endogenous security theory has laid a solid engineering foundation for cyberspace security. As AI systems increasingly become intelligent agents with autonomous reasoning capabilities, the core of security problems is shifting from "system architecture" to "cognitive primitive." Acknowledging this shift and initiating the paradigmatic exploration from structural endogeneity to cognitive endogeneity will be a crucial step in the continued development of endogenous security theory.

References

[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] Wu J X. 内生安全理论方法赋能数字产业新质生产力 [Endogenous security theory empowering new quality productive forces in the digital industry][C]. 2024 China Internet Conference, Data Security Sub-forum, Beijing, 2024. (in Chinese)

[3] Institute for Big Data, Fudan University. 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 [Breaking through and breaking out: the path to constructing a new paradigm of endogenous security for artificial intelligence — from theoretical foundations to leading development highlights][EB/OL]. (2026-02-10). https://ibd.fudan.edu.cn/d8/29/c24057a776233/page.htm. (in Chinese)

[4] Li Y F, Zheng X Y, Qiu H, et al. Research on resilience technology for intelligent connected vehicle empowered by endogenous security and safety[J]. Scientia Sinica Informationis, 2025, 55(8): 1822-1848. DOI: 10.1360/SSI-2025-0041.

[5] Wu J X. Research on cyberspace mimic defense[J]. Journal of Cyber Security, 2016, 1(4): 1-10. DOI: 10.19363/j.cnki.cn10-1380/tn.2016.04.001.

[6] Liu S. 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. The Foundational Revolution of Intelligence: From Data Warehouse 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] National Development and Reform Commission, National Energy Administration. Implementation Opinions on Promoting High-Quality Development of "AI+" Energy (NDRC Energy Technology Document No. 73 [2025])[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 rea soning prototype[CP/OL]. (2026). https://github.com/WindSeaAI/OpenMorpho/tree/main/OpenMorpho/examples/hydrogen-embrittlement-demo. [Accessed: 2026-07-12].