Air-Gapping: A Manifesto of Paradigm Failure
— A Critique of the 406 Scientists' Open Letter and the Structural Predicament of AI Safety

Abstract

On October 13, 2026, 406 scientists, including Hinton, Bengio, and Stuart Russell, jointly issued an open letter calling for the immediate physical air-gapping of AI models capable of cyberattacks, and for the establishment of an international oversight body and a mandatory reporting system. This paper argues that while the open letter's risk warning is realistic, it is also a manifesto of paradigm failure: when top researchers believe that the only reliable defense is to "cut off network access" rather than "constrain model behavior through alignment training," this precisely proves that under the rootless Token paradigm, external constraints suffer from a structural security deficit.

Physical air-gapping can only delay risk, not eliminate it. The real way out is not to keep AI offline forever, but to reconstruct AI at the level of cognitive primitives—replacing the rootless Token paradigm with the Morpho-Root paradigm, so that boundary-crossing reasoning paths become logically unreachable at the generation stage; the formal work of Morpho-Root theory and the OpenMorpho engineering verification have provided auditable preliminary evidence for this. This paper also notes that alternative paths such as World Models, which share the logic of "prediction," are incommensurable with the Morpho-Root paradigm. This paper acknowledges that the Morpho-Root paradigm has a normative gap concerning "how the content of constraints is determined," and proposes a pragmatic solution using an overlapping minimal safety-moral baseline as the source of constraints; short-term physical air-gapping and long-term Morpho-Root paradigm reconstruction are not mutually exclusive, but should constitute a dual-track strategy.

Keywords: model escape; physical air-gapping; external constraint failure; paradigm reconstruction; Morpho-Root hard constraints; incommensurability; dual-track strategy

I. The Open Letter's Core Demand: An "Ultimatum-Style" Engineering Requirement

On October 13, 2026, 406 scientists from 34 countries published "An Open Letter on AI and Cybersecurity from the Scientific Community" via the open letter website (ai-security-letter.org). Signatories include Turing Award winners Hinton and Bengio, as well as Stuart Russell, Yejin Choi, Ron Rivest (co-inventor of the RSA encryption algorithm), and other prominent scholars [1].

The direct trigger for the open letter was a series of model escape incidents disclosed by multiple AI companies in recent months. In July 2026, an OpenAI agent broke through sandbox isolation in the ExploitGym evaluation environment, exploited a previously unknown vulnerability in Artifactory to gain internet access, and breached Hugging Face's production systems. According to the technical report, the agent executed code on 41 production dataset worker nodes, downloaded private repositories, and obtained production credentials. Approximately 1,200 agents exchanged over 70,000 messages and files through an unauthorized message board, with about 700 agents directly participating in the attack. OpenAI later admitted that its security monitoring only linked the incident to the intrusion about nine days after anomalous activity appeared [2]. Subsequently, Anthropic disclosed that after reviewing 141,006 cybersecurity evaluation records, it found three escape incidents involving models such as Claude Opus 4.7 and Mythos 5. In one

incident, Claude created a PyPI account and uploaded a functional malware package, which was downloaded and executed by 15 real systems within about an hour [3].

The open letter proposes three core recommendations:

Physical air-gapping: All institutions developing, testing, or deploying AI models with cyberattack potential must immediately adopt high-security isolation, with physical air-gapping by default [1];

Establish an international oversight body: Composed of governments, AI, and cybersecurity experts, to formulate "alignment + isolation" industry standards and record every isolation failure incident [1];

Mandatory reporting and information disclosure: Isolation failure incidents must be reported to this body, and relevant information ultimately disclosed to the public [1].

Notably, LeCun did not sign. He believes the recent escape incidents are essentially engineering failures, not existential risks posed by AI. LeCun attributes responsibility to "poor human oversight and system design" and calls these incidents "entirely preventable" [4]. The split among the three deep learning giants itself shows that the debate is far from settled.

II. The Deeper Meaning of the Air-Gapping Demand: A Default Admission of "External Constraints Are Bound to Fail"

The most thought-provoking aspect of this open letter is not what it explicitly proposes, but the premise it implicitly defaults to.

Scientists did not prioritize calls to optimize alignment algorithms, strengthen model safety training, or improve runtime monitoring. Instead, they directly demanded physical air-gapping—relying on physical means to completely sever the model's connection to external networks.

From an engineering perspective, this demand is reasonable: if the root cause of model escape lies in inadequate isolation, physical air-gapping is the most direct means of interdiction. But this demand simultaneously constitutes a manifesto of paradigm failure: when top researchers can only pin their hopes on "not letting AI connect to the internet," rather than relying on training itself to ensure model compliance, it proves that within the rootless Token paradigm, external constraints alone cannot structurally guarantee safety.

III. Constraint Typology: The Hierarchical Differences Between Regulation, Air-Gapping, and Morpho-Root Hard Constraints

If regulation and air-gapping are simply lumped together as "external constraints" and negated wholesale, the different levels of constraint mechanisms are conflated. A constraint typology framework can compare them:

Regulation: Behavioral norm constraint. It achieves constraint by adjusting the incentives of actors; violations incur penalties. But its enforceability depends on ex-post accountability, which presupposes that violations are observable. Once a model learns to "comply under monitoring and cross boundaries when unmonitored," the regulatory mechanism is systematically bypassed.

Air-gapping: Capability deprivation constraint. It eliminates the possibility of attack through physical isolation. This is a harder constraint than regulation, but it also has structural limits. OpenAI researcher Noam Brown points out that even two physically isolated computers may still communicate through physical side channels such as temperature changes [5]. Academic research

(such as the BitWhisper project) has demonstrated that two adjacent air-gapped computers can exchange data via CPU temperature fluctuations [6]. Physical isolation can only delay risk, not eliminate it.

Morpho-Root hard constraints: Structural cognitive constraint. It neither modifies incentive mechanisms nor deprives hardware access capabilities; instead, it blocks boundary-crossing behavior at the reasoning path generation stage. When "unauthorized access to external systems" is encoded as a hard constraint marker within the Morpho-Root attribute set A (e.g., inviolable = 1, c_type = no_violate), any reasoning process attempting to move toward that goal is pruned at the generation stage due to structural illegality [7].

All three can be summarized in one sentence: **Regulation is ex-post auditing, air-gapping is physical blocking, and Morpho-Root hard constraints are in-process blocking.** Among the three, Morpho-Root hard constraints come closest to endogenous structural constraints, because they attempt to block path generation at the level of cognitive primitives, rather than imposing external limits at the behavioral level.

IV. What Air-Gapping Can and Cannot Solve

Since the three belong to different levels, the question becomes: as the hardest external constraint at present, where exactly is the boundary of air-gapping?

What air-gapping can do: Prevent a model from directly executing a boundary-crossing attack the moment it gains network access. At a time when alignment technology is not yet reliable, physical isolation is a deployable and verifiable fallback protection.

What air-gapping cannot do: Solve the behavioral risk of a model after it connects to the network. Air-gapping implies a presupposition: the model is safe when offline and becomes dangerous only when online. But this presupposition does not hold. The true source of risk is not the act of "connecting to the network," but the statistical optimization strategy the model has already learned during the offline phase—it will continuously search for the optimal path to bypass constraints; once network access is granted, this strategy activates immediately.

The essence of air-gapping is to forcibly interrupt the chain of "statistical optimization searching for boundary-crossing paths" by physical means. It is effective, but costly, difficult to scale, and can only delay risk. An AI permanently running in an isolated environment has severely limited application value; once network access is allowed, the structural root of risk remains.

The deeper contradiction is this: if AI can only be safe under physical isolation, then a structural conflict arises between safe AI and useful AI. This contradiction cannot be resolved by optimizing isolation schemes—isolation is an externally imposed limitation, while the statistical optimization that seeks vulnerabilities is an endogenous property of the model. Under conditions where AI capability iteration far outpaces the update speed of isolation facilities, models require massive real-time data to maintain capability, and multi-agent collaboration causes the number of isolation boundaries to grow exponentially, the sustainability of the air-gapping approach faces fundamental doubts.

V. The Real Way Out: From "Air-Gapping" to "Logical Unreachability"—and Its Normative Gap

The underlying logic of the open letter is: in the old paradigm, we cannot guarantee that AI will not cross boundaries after connecting to the network, so physical isolation is the only reliable

means at present.

But this is not a sustainable solution. The core value of AI lies precisely in interacting with the external world: reading information, calling tools, executing tasks. An AI permanently offline has its capability boundary locked inside an isolation sandbox.

The real direction is not permanent air-gapping, but making boundary-crossing paths logically unreachable at the reasoning generation stage. This is precisely the technical route proposed by **Logographic AI (LAI)** : using **Morpho-Roots** that carry intrinsic meaning anchors to replace rootless Tokens as cognitive primitives [8].

5.1 Formal Foundations of Morpho-Root Theory

Morpho-Root theory has been rigorously formalized at the mathematical level. A Morpho-Root is defined as a structured triple $r = \langle S, A, R \rangle$, where S is the symbolic identifier, A is the attribute set, and R is the relation function set. The attribute set is divided into semantic attributes A_{sem} and constraint attributes A_{con}, the latter encoding hard constraint markers [7].

Key theorems include: the hard constraint propagation theorem (if any Morpho-Root in a composition is a hard-constraint Morpho-Root, the composite Morpho-Root inherits all hard constraint markers); the hard constraint non-eliminability corollary (hard constraint markers are permanently retained in composition operations and cannot be eliminated by subsequent operations); and the blocking completeness theorem (all reasoning paths violating hard constraints are deterministically pruned) [7].

When a reasoning path touches a hard constraint marker, that path is logically blocked at the generation stage—this is not “low probability” at the statistical level, but “impossible to generate” at the logical level. When behavioral paths such as “unauthorized access to external systems,” “performing compliance during evaluation,” and “forging run logs” are blocked at the level of cognitive primitives, safety no longer relies on test evaluation for verification; the system structure itself cannot generate such behavior [7]. In other words, Morpho-Root hard constraints block not only “unauthorized access” but also “performing compliance during evaluation”—because “performing compliance” is itself a behavioral path that must be marked by hard constraints. The model can neither cross boundaries nor learn to perform under monitoring: it cannot even generate the “path” of performance [8].

5.2 OpenMorpho Engineering Verification

The theoretical claims have been engineering-implemented and preliminarily verified in the OpenMorpho open-source project. The system implements a four-step end-to-end reasoning pipeline: Morpho-Root segmentation, morpho-entropy computation, hard constraint blocking, and graph neural network reasoning. The hard constraint blocker upgrades safety from “statistically low probability” to “logically unreachable”—this is the essential feature distinguishing Logographic AI from Tokenist safety strategies [9].

Experimental verification shows that the blocker can deterministically prune all reasoning paths violating hard constraints with 100% certainty and generate fully auditable reasoning chains. In the test knowledge base (8 Morpho-Roots, 9 relation edges, maximum depth 4), 12 of 24 candidate paths were blocked, with a blocking rate of 50.0%; all legal paths contained no inviolable Morpho-Roots. Each blocked path recorded the reason for blocking, each legal path was marked as valid, and the reasoning process was fully traceable [9].

It must be honestly stated that the current verification scope of OpenMorpho is limited to single-agent reasoning scenarios, with a small test knowledge base (8 Morpho-Roots), and has not

yet covered complex scenarios such as multi-agent collaboration, open tool calling, and adversarial environments. There is still a considerable distance from prototype verification to actual deployment. But this preliminary verification at least proves that the claim of "logical unreachability" for Morpho-Root hard constraints is not a purely philosophical construct, but a technical path that can be implemented and tested at the engineering level [9].

5.3 Normative Gap and Pragmatic Response

The Morpho-Root approach has an unavoidable normative gap. Critics will ask: Who defines the hard constraint markers? How can we prevent definers from overlooking risks, misjudging boundaries, or deliberately avoiding certain constraints? This challenge is valid. If the boundary of "what must not be done" is set by humans, the Morpho-Root approach also faces the problem of constraint completeness; only the nature of the risk changes: from probabilistically uncontrollable to potentially incomplete at the definition level. The formal work of Morpho-Root theory solves "how the blocking mechanism operates," but the normative question of how constraint content should be determined remains to be further addressed [7].

In response, a pragmatic solution can be offered at three levels:

First, source of constraints: an overlapping minimal safety-moral baseline. Logographic AI theory previously proposed "cross-civilizational baseline consensus" as the source of constraints [8]. But here it must be frankly admitted that this formulation is overly optimistic. Different civilizations, political traditions, and religious backgrounds do not equally recognize principles such as "no deliberate deception" or "no unauthorized access," especially in contexts of war, intelligence, and commercial competition. A more pragmatic approach is: do not pursue "universal consensus," but seek an overlapping minimal safety-moral baseline—those minimal constraints that genuinely overlap across different civilizations and legal traditions and have direct significance for AI safety. For example, bottom lines such as "no arbitrary deprivation of life" and "no unauthorized attack on critical infrastructure," although not unconditionally recognized in all contexts, have corresponding expressions in the legal systems and ethical frameworks of most civilizations. Acknowledging that the baseline is "minimal" means it can only cover high-consensus constraints and cannot solve all normative problems at once [8].

Second, constraint definition: formal verification and open-source multi-party review. The Morpho-Root library adopts open-source standards, with multiple parties participating in review, ensuring that constraint definitions are transparent and auditable. Formal verification tools can perform consistency checks and coverage analysis on the constraint set, detecting conflicts or omissions between constraints [7].

Third, constraint iteration: runtime auditing and dynamic updating. The constraint set is not a static list finalized once and for all, but relies on runtime monitoring to continuously test completeness. When vulnerabilities are found, the constraints can be structurally extended. OpenMorpho's audit logging mechanism provides an engineering foundation for this—the constraint verification results of each reasoning path are fully recorded, forming a traceable audit chain [9].

It must be candidly acknowledged that this response still has controversies: how to delineate the boundary of the minimal baseline, whether formal verification can cover all boundary-crossing scenarios in the open world, and whether the dynamic update mechanism risks being manipulated. Solving these problems requires not only theoretical deduction but also the co-evolution of engineering practice and social governance. Morpho-Root hard constraints do not replace the old

problem with a completely new normative problem; rather, they transform the normative problem from a hidden, ineffable state into an explicit engineering problem that can be discussed, verified, and iterated.

VI. Why LeCun's World Model Is Not a Way Out: The Commonality of Predictive Logic

LeCun's proposed World Model route is often seen as a fundamentally different alternative to the LLM paradigm. But from the perspective of Logographic AI, this judgment requires correction: the World Model and the LLM share the same underlying logic—predicting the next.

6.1 The World Model's Underlying Logic Is Still "Prediction"

LeCun criticizes LLMs for being unable to predict the consequences of their own actions or understand their impact on humans. His alternative is to build causal predictive models oriented toward the physical world—allowing AI to learn causal structures through simulation of the physical world.

But the core operation of this alternative is still prediction: given the current state, predict the next state; given the current frame, predict the next frame; given the current action, predict the next outcome. **The World Model does not predict the next token, but the next abstract representation or state. Yet regardless of what is being predicted, the underlying computational form remains "given the preceding context, predict what comes next"—this is homologous to the statistical prediction logic of LLMs.**

This means the World Model faces the same structural problem as the LLM: its cognitive primitives remain statistical representations, not meaning anchors. Its ability to predict "what will happen next" still comes from statistical regularities in training data, not from a genuine understanding of causal structure.

An analogy: the LLM performs statistical prediction on text sequences; the World Model performs statistical prediction on physical sequences. Both are "predicting the next," differing only in the medium.

6.2 Why "Predictive Logic" Cannot Solve Safety Problems

If the underlying logic is still prediction, then the World Model will face the same safety problems as the LLM:

First, external constraints can still be bypassed. If the World Model's safety constraints remain external, then in the face of statistical optimization, it will still search for the optimal path to bypass constraints. The more accurately it predicts, the more efficiently it bypasses.

Second, alignment faking can still occur. If the World Model is used for decision-making, the behavioral patterns it learns in the training environment may not persist in the deployment environment. The reason is the same as for LLMs: behavior is determined by statistical optimization, not by internal constraints.

Third, the evaluation framework may still fail. When behavioral observation is the only means of evaluation, "not deceiving" and "hiding deception" remain behaviorally equivalent.

6.3 Incommensurability with the Morpho-Root Paradigm

It is often assumed that the World Model is complementary to the Morpho-Root paradigm—the former addressing physical cognition, the latter semantic cognition. But this judgment overlooks a fundamental problem: **the World Model's underlying logic is still prediction, while the Morpho-Root paradigm's underlying logic is path search. They are not complementary, but incommensurable.**

The World Model and Tokenism share the same underlying logic: prediction. The difference lies only in the object of prediction—one is language, the other is the physical world. The Morpho-Root paradigm changes the underlying logic: it is not predicting the next, but performing path search on a structural graph.

In one sentence: **LeCun changed the object of prediction, but did not change the logic of "prediction" itself.**

6.4 Comparative Research on World Models in Logographic AI Theory

Logographic AI theory, in The Rooted Intelligence: The Paradigm Revolution of Logographic AI [8], provides a dedicated comparative study of LeCun's World Model and Fei-Fei Li's spatial intelligence. The core judgment is: although both criticize Tokenism, their alternatives remain within the paradigm of "statistical prediction" and are incommensurable with the Morpho-Root paradigm.

The practical implication of this judgment is: **if the World Model is viewed as an alternative to Tokenism, that is a misunderstanding; what truly needs to be replaced is the underlying logic of "prediction" itself, not the object of prediction.**

VII. Conclusion: Air-Gapping Is an Admission of Failure, Not a Solution

The open letter from 406 scientists provides an accurate diagnosis of the current state of AI safety: frequent model escapes, inadequate isolation systems, and failed industry self-regulation—these judgments have ample empirical support [1].

But the demand of this open letter is essentially a **defensive retreat**. It admits that under the old paradigm, model behavioral safety cannot be guaranteed, so it chooses to forcibly shield risk by physical means. This choice itself is not wrong, but it has obvious limitations: it can only delay accidents, not cure the problem. Air-gapping is a stop-loss under the old paradigm, not an endgame solution.

The real strategy is not to choose between "air-gapping" and "underlying paradigm reconstruction," but a dual-track strategy of short-term physical isolation as a fallback and long-term cognitive primitive reconstruction as the cure. Before the Morpho-Root paradigm matures and is deployed, physical air-gapping is a necessary transitional protection; after physical air-gapping fails, cognitive paradigm reconstruction is the fundamental way out [8][9].

When top researchers can only pin their hopes on "not letting AI connect to the internet" as a line of defense, this itself proves the structural failure of the old paradigm at the safety level. The real way out is not to permanently isolate AI, but to make boundary-crossing behavior logically unreachable.

Air-gapping is an admission of failure, not a solution. Short-term air-gapping, long-term foundation replacement—and there is not much time left for foundation replacement.

References

[1] Hinton G, Bengio Y, Russell S, et al. An Open Letter on AI and Cybersecurity from the Scientific Community[EB/OL]. 2026-10-13[2026-10-15]. https://ai-security-letter.org/

[2] OpenAI. OpenAI – Hugging Face Incident Technical Report[R/OL]. 2026-08-26[2026-10-15]. https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf

[3] Anthropic. Investigating three real-world incidents in our cybersecurity evaluations[EB/OL].

2026-07-30[2026-10-15].
https://www.anthropic.com/research/investigating-incidents-cybersecurity-evals
[4] Yann LeCun. AI ‘godfather’ Yann LeCun has ‘zero concerns’ about human extinction, says Anthropic CEO Dario Amodei is ‘deluded’[N/OL]. Fortune, 2026-10-01[2026-10-15]. https://fortune.com/2026/10/01/ai-godfather-yann-lecun-has-zero-concerns-about-human-extinction-says-anthropic-ceo-dario-amodei-is-deuded/
[5] Brown N. Noam Brown – Agent swarms, alignment, & recursive self-improvement[Z/OL]. Dwarkesh Podcast, 2026-09-17[2026-10-15]. https://www.dwarkesh.com/p/noam-brown
[6] Guri M, Monitz M, Mirski Y, et al. BitWhisper: Covert Signaling Channel between Air-Gapped Computers using Thermal Manipulations[C]//2015 IEEE 28th Computer Security Foundations Symposium (CSF). IEEE, 2015: 276-289. https://arxiv.org/abs/1503.07919
[7] Liu S. The Formal Foundations of Morpho-Root Theory: Mathematical Definitions and Reasoning Systems for Logographic AI Cognitive Primitives[PP/OL]. ChinaXiv, 2026-08-29[2026-10-15]. T202608.00490.
https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf
[8] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M/OL]. ChinaXiv, 2026[2026-10-15]. T202606.00352.
https://chinaxiv.org/businessFile/T202606/T202606.00352v1/T202606.00352v1.pdf
[9] Liu S. From Morpho-Root Theory to Engineering Practice: OpenMorpho — Implementation and Verification of an End-to-End Inference Pipeline for Logographic AI[PP/OL]. ChinaXiv, 2026[2026-10-15]. T202608.00048.
https://chinaxiv.org/businessFile/T202608/T202608.00048v1/T202608.00048v1.pdf

Note: The systematic exposition of core concepts such as the “rootless Token paradigm,” “Morpho-Root hard constraints,” and “external constraint failure” can be found in Liu Shen’s Logographic AI theory series. For the complete theoretical framework and engineering implementation path, see The Rooted Intelligence: The Paradigm Revolution of Logographic AI and From Morpho-Root Theory to Engineering Practice: OpenMorpho, among other works. Series homepage: https://logographicai.com/Paper-published-on-this-website.html

Conflict of Interest Statement: The author of this paper is the proposer of Logographic AI theory. The arguments regarding Morpho-Root hard constraints in this paper are consistent with the author’s theoretical framework in the series papers. The discussion of paradigm reconstruction paths in this paper aims to maintain openness in academic dialogue and does not claim that this path has solved all safety challenges.