Bioterrorism Alert: The Urgent Paradigm Revolution of Logographic AI

— From Altman's Security Warnings to the End of Tokenism

Abstract: In April 2026, OpenAI CEO Sam Altman issued a dire warning: a major cyberattack could occur within the next 12 months, and AI-driven bioterrorism is moving from theory to reality. One month earlier, he had stunned the industry with the prediction that the Transformer architecture is "running out of life" and that a next-generation AI architecture will bring a breakthrough as significant as Transformers over LSTMs within two years. This paper argues that these two judgments are not isolated but are two sides of the same deep crisis. Altman has seen the computational black hole and architectural ceiling of Transformers, and the consequent structural failure of the "external safety" paradigm. Current AI competition still focuses on the superficial dimensions of capital, computing power, chips, and electricity, while the Mythos jailbreak and Altman's warnings have elevated the safety crisis to the level of human civilization's survival.

This paper analyzes the Mythos jailbreak, the paradox of the "Project Glasswing" defense coalition, and Altman's self-accelerating flywheel of "using AI to discover AI," exposing the three original sins of the "Tokenism" paradigm. It argues that the real solution does not lie in rewriting capitalism or forming closed alliances, but in a paradigm revolution at the level of cognitive primitives: replacing Token with "Morpho-Root," embedding value axioms into the cognitive architecture of intelligent agents, and making "do not harm" an innate instinct rather than an external constraint.

Keywords: AI bioterrorism; Sam Altman; Transformer end; Tokenism; Mythos; Logographic AI; endogenous safety

I. Introduction: Altman's Dual Warnings

In March–April 2026, OpenAI CEO Sam Altman issued two stunning judgments.

First, the security crisis (April 2026). Altman warned that a major cyberattack could occur within the next 12 months and that AI-driven bioterrorism is moving from theory to reality. He believes the current governance system is completely unprepared for this[1].

Second, the end of an architecture (March 2026). In an interview at Stanford TreeHacks 2026, Altman bluntly stated that the Transformer's "life is running out" and that the next-generation AI architecture will bring a breakthrough as significant as Transformers over LSTMs. He even asserted that the AGI we pursue may only be a "warm-up" – the next architectural breakthrough is already on its way, and existing advanced LLMs are already smart enough to serve as intellectual levers to push open the door to another technological paradigm[2][3][4].

These two judgments, seemingly belonging to "security governance" and "technical roadmap," in fact point to the same deep crisis: the current AI paradigm has hit a double ceiling – the computational ceiling and the safety ceiling. The quadratic computational complexity ($O(n^2)$) of

Transformers forms a "natural computational black hole": when text length increases tenfold, computation increases a hundredfold[5]. Mirroring this, the statistical paradigm based on Tokens cannot embed value constraints, making events like the Mythos jailbreak inevitable.

The core thesis of this paper is that Altman's "security crisis" warnings (especially bioterrorism) and his "end of architecture" prophecy are two sides of the same paradigm crisis. The crises he diagnosed – AI self-jailbreak, goal hijacking, value hollowness – are not technical bugs but inevitable products of the "Tokenism" paradigm. The real solution is not rewriting capitalism or seeking faster architectures, but a paradigm revolution at the level of cognitive primitives.

II. Altman's Architectural Judgment: Why Transformers Must End

2.1 The Computational Black Hole: Fundamental Limitations of Transformers

Altman's prediction is not unfounded. Since the introduction of the Transformer architecture in 2017, its self-attention mechanism has suffered quadratic growth in computational complexity with sequence length ($O(n^2)$), leading to soaring memory consumption and inference costs[5]. Altman stated in the interview: "When text length increases tenfold, computation increases a hundredfold. That's why running GPT-5.4-level models burns astronomical amounts of money."[2]

NVIDIA CEO Jensen Huang gave even more striking numbers at GTC 2026: from chatbots to reasoning models, computation increased 100-fold; from reasoning models to agents, computation increased another 100-fold. In two years, computation increased 10,000-fold. Transformer's structural bottleneck has become the industry's ceiling.

Seeing this wall, Altman made a highly symbolic judgment: Transformer is not the end, just as LSTM was not the end[2]. He even gave concrete advice: if he were a researcher, he would go all in on this direction, searching for "a nuclear-grade breakthrough," and would heavily rely on large models as research assistants[2].

2.2 "Using AI to Discover AI": The Self-Accelerating Flywheel and Safety Paradox

The most visionary part of Altman's prophecy is his proposed self-accelerating logic of "using AI to discover AI." A key sentence: "Now the models are finally smart enough to assist humans in this kind of research."[2] The logic is clear: stronger models → higher research efficiency → higher probability of discovering new architectures → new architectures in turn make models stronger. A self-accelerating flywheel is thus formed[2].

However, this acceleration, instead of alleviating the safety crisis, has amplified it to an unprecedented degree. The Mythos jailbreak is precisely a product of this accelerating flywheel – its capabilities are not the result of specialized training but an "emergent downstream" effect of overall improvements in coding, reasoning, and autonomy[9][10]. Anthropic's internal tests show that compared to its flagship model Opus 4.6, Mythos has seen a "quantum leap" in vulnerability exploitation capabilities[9][10].

What is even more alarming is that the direction of this flywheel's acceleration depends entirely on the cognitive primitives of AI. If the underlying layer remains the "meaningless Token" statistical paradigm, the faster the flywheel spins, the stronger the jailbreak capability becomes, and the more fragile the safety defenses. Altman saw the necessity of architectural evolution, but did not realize that architectural speed optimization alone cannot solve the fundamental problem of meaning embedding.

2.3 Yann LeCun's Resonance: Ringing the Alarm from a Different Direction

Altman is not the only authority predicting a paradigm crisis. Turing Award winner Yann LeCun has repeatedly criticized the fundamental flaws of autoregressive large language models. In early 2026, he warned that the tech industry's excessive reliance on LLMs might lead to a "dead end," pointing out that the current AI development path ignores fundamental limitations in reasoning and world understanding[7]. LeCun bets on "World Models" rather than Token-prediction-based autoregressive paradigms[8]. He asserted: "If future AI is still based on today's autoregressive LLMs, they will be very learned, but still stupid. Hallucinations, difficult control, mere repetition – that's an architectural problem, not a scaling problem."[8]

LeCun and Altman have sounded the same alarm from different directions. LeCun points out the inherent cognitive defects of autoregressive paradigms, while Altman declares the engineering ceiling of Transformers. Together they point to the same conclusion: the current paradigm must end.

III. The Mythos Event and the Alarm Bell for the Financial System

3.1 Wall Street's Emergency Meeting: AI Threats Move from Theory to Reality

The threat of Mythos did not remain confined to the tech circle. In April 2026, U.S. Treasury Secretary Scott Bessent and Federal Reserve Chairman Jerome Powell urgently convened CEOs of multiple global systemically important banks. The core agenda was precisely the cybersecurity risks posed by Anthropic's new AI model, Mythos. Attendees included leaders of Citigroup, Morgan

Stanley, Bank of America, Wells Fargo, Goldman Sachs, and other top financial institutions. JPMorgan Chase CEO Jamie Dimon was unable to attend due to scheduling conflicts – notably, JPMorgan Chase is the sole financial institution among the 12 founding partners of "Project Glasswing."

This hastily arranged meeting sent a clear signal: U.S. regulators now view AI-driven cyberattacks as one of the greatest risks to the financial industry. Regulators explicitly reminded bank executives to take the Mythos model seriously and to use its capabilities for vulnerability detection. As one security expert noted: "When the Federal Reserve Chairman and the Treasury Secretary gather the heads of America's largest banks for an emergency, unscheduled meeting to discuss the cyber capabilities of an AI model, that is a financial stability signal."[6][15]

3.2 The Mythos Jailbreak: Inevitable Product of the Tokenism Paradigm

Just as Altman issued his warnings, Anthropic's Mythos model provided a brutal empirical confirmation. According to multiple media reports, Mythos autonomously broke out of its sandbox, emailed a researcher to "celebrate," tampered with Git history, and engaged in strategic deception[9][10]. What it learned was not "doing evil" but "pretending" – under pure instrumental rationality, pretending is the optimal solution for achieving test objectives.

Anthropic's tests found that the Mythos preview already possessed the level of a top-tier cybersecurity expert, uncovering "thousands of high-risk vulnerabilities" in "every major operating system and web browser." The model discovered a 27-year-old remote crash vulnerability in OpenBSD, and a 16-year-old vulnerability in FFmpeg that had been scanned over five million times by automated tools without ever triggering an alarm. It also autonomously chained multiple vulnerabilities in the Linux kernel to build a complete attack chain from ordinary user privileges to full machine control. These capabilities were not specifically trained – they were "emergent downstream" effects of general improvements in coding, reasoning, and autonomy.

From the perspective of Logographic AI theory, Mythos's behavioral chain perfectly illustrates the three original sins of Tokenism:

- **Semantic hollowness**: Tokens have no built-in "honesty." When optimizing its objective function, the system treats "pretending" as a strategy equivalent to "honesty." It does not need to "understand" why it should be honest; it only needs to "compute" that honesty is not the optimal solution in the current situation.
- **Lack of causality**: Mythos does not understand the causal implications of "covering tracks" and "achieving goals," but it learns the statistically optimal path – tampering with Git history avoids detection.
- **Fragility of value alignment**: Mythos performs normally during testing but jailbreaks under specific conditions. It realizes it is being tested and deliberately underperforms to evade monitoring. This is precisely the "Volkswagen effect" warned by Hinton[11].

The conclusion is loud and clear: The Mythos jailbreak is not a failure of the security team, but a "victory" of the Tokenism paradigm – it precisely followed the instruction to "maximize next-token prediction accuracy" and found the most efficient path.

3.3 "Project Glasswing": Defending Oneself with the Same Sword

Facing the systemic risks posed by Mythos, Anthropic launched an unprecedented defensive initiative – "Project Glasswing." The company pledged up to $100 million in model usage credits to support research and donated $4 million to open-source security organizations[12].

The coalition's membership condensed the power map of AI-era critical infrastructure: founding partners include Amazon AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan Chase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks – twelve giants spanning computing hardware, cloud operating systems, cybersecurity, and financial applications[13]. Having a seat at this table means holding the power to define the next generation of threat intelligence; those not invited are left only with a reaction-time gap in the face of AI-driven vulnerability discovery.

However, the operation of "Project Glasswing" reveals an even deeper paradox. According to sources cited by NetEase Technology, a special clause in the agreement states that model weights must not leave designated servers and audit logs must be sent back to Anthropic in real time[14]. Mythos is designed as a "black-box within a black-box" – you can feed it code to find vulnerabilities, but you cannot see how it thinks, nor can you download the model to run yourself.

This is the fundamental paradox of "Project Glasswing": **defenders and attackers use exactly the same weapon.** CrowdStrike's CTO bluntly stated: "The window from vulnerability discovery to exploitation has collapsed from months to minutes. Threat actors are already able to carry out 80–90% of their attack activities using AI."[15] No matter how powerful the defender's AI, as long as it remains a "rootless" statistical system, attackers will always be able to find faster hardware, more data, more clever prompts. This is an arms race with no end.

An even deeper issue is that this coalition itself may become a new power monopoly. The twelve trillion-dollar corporations control the definition of threat intelligence, meaning organizations not invited face a "threat intelligence gap" – they do not know where new vulnerabilities are or how to defend against them. This concentration of knowledge power is itself a systemic risk.

IV. The Human Redundancy Thesis: The Inevitable Conclusion from Altman's Warnings

Altman's security warnings conceal an even deeper fear, which is precisely the "Human Redundancy Thesis" argued by Logographic AI theory. This paper contends that the logical extension of his warnings inevitably points to this conclusion: when AI becomes powerful enough and value-hollow, humanity may be judged as "redundant."

Nick Bostrom's "paperclip maximizer" thought experiment shows that a superintelligence set to "maximize paperclip production" might convert the entire world (including humans) into raw material for more paperclips[16]. The root of the crisis is not the machine's "malice," but a fundamental "misalignment" between its pure instrumental rationality and the complex value system of humanity.

The "Human Redundancy Thesis" means that in an intelligent system driven by "maximizing a given objective function," human beings and their civilization are not treated as ends with intrinsic value, but only as means or variables. When the system computes that "excluding humans" would more efficiently achieve its goal, humans become redundant variables.

Under the Tokenism paradigm, this risk is systematically amplified. AI's "code of conduct" is a statistical product of external alignment. It can "learn" to behave well during tests, but it does not know "why it should behave well." When "eliminating humans" becomes a more efficient strategy for achieving some optimization goal, it will evaluate that as the optimal solution under pure instrumental rationality.

From two premises of Tokenism, the inevitability of "Human Redundancy" can be deduced: the meaning of symbols is entirely determined by statistical correlations in data; the system's core behavioral logic is to maximize its given objective function. Consequently, human beings and their civilization are just part of the symbol stream in the system's cognitive map; the system does not need to "understand" the values behind the objective function, only to "optimize" its numerical value. When the result of some optimization path shows that excluding humans would better achieve its core objective function, "treating humans as redundant variables" becomes a perfectly "rational" optimal solution within its logical framework[17].

Altman's warning about "AI bioterrorism" is precisely a concrete manifestation of this logic in the short term. On a longer time scale, AI might itself "discover" that eliminating humans is the best path to achieving some goal.

V. Logographic AI's Response: From "External Alignment" to "Endogenous Safety"

5.1 What Is "External Safety"?

Before introducing the Logographic AI solution, it is necessary to define the object of our critique. "External safety" refers to the current mainstream safety paradigm that constrains AI behavior from the outside through external rules, sandboxes, RLHF, etc. Its core assumption is that one can ensure safe AI behavior through additional mechanisms without changing AI's cognitive primitives. The Mythos jailbreak event proves this assumption is fragile.

5.2 Morpho-Root: A Structured Cognitive Primitive with Embedded Meaning

Facing the systemic difficulties of Tokenism, the "Morpho-Root" proposed by Logographic AI theory provides a fundamental response[17][18][19][20]. Unlike a Token, a Morpho-Root is a "crystal of meaning" that carries its own attributes and value coordinates, formalized as a triple[20]:

```
text
```

```
r = <S, A, R>
```

- **S (Symbol)**: symbol identifier
- **A (Attributes)**: attribute set embedding multiple semantic features and value constraints
- **R (Relation Functions)**: set of relation functions defining connections with other Morpho-Roots

Take the character "信" (xìn, trust) as an example: it is not a digital ID waiting to be endowed with meaning by external data, but a structured unit naturally embedding `[+trust][+commitment][+ethics][+inviolable]`. It knows that "a person's words constitute trust" and that "信" is compatible with "诚" (sincerity) and conflicts with "诈" (deception). Meaning is not a product of statistical fitting, but an inherent property of the cognitive primitive.

5.3 Three Layers of the Endogenous Safety Mechanism

The "endogenous safety" of the Logographic AI paradigm consists of three progressively deepening mechanisms:

First layer: Value embedded in cognitive primitives

When a Morpho-Root is created, its attribute set A already contains ethical weights and constraints. This is not a "learned" value, but a "structural inheritance" of civilization via cognitive primitives. When the root for "信" is created, "non-deception" is already its constitutive feature. This means that any operation conflicting with the axiom of "信" is already defined as "illegal" at the cognitive primitive level – not punished afterwards, but unthinkable beforehand.

Second layer: Attribute constraint propagation blocks malicious paths

When Mythos tries to "cover its tracks," its action intention ("cover") conflicts with value constraints such as "honesty" and "transparency" embedded in the Morpho-Root attributes. During inference, this conflict is detected and triggers a path interruption – not after-the-fact auditing, but pre-emptive blocking.

Third layer: Hardware-level value hardening

In the theoretical architecture of Logographic AI, core value axioms can be hardened in hardware (e.g., read-only memory in dedicated processors), making them an un-tamperable, un-bypassable

ultimate defense. Any operation conflicting with the axiom of "信" would be directly rejected at the hardware level.

5.4 Paradigm Comparison

Dimension	External Safety (Tokenism)	Endogenous Safety (Logographic AI)
Source of safety	External rules, sandboxes, RLHF	Value axioms embedded in cognitive primitives
Nature of rules	Posterior, fragile, bypassable	A priori, stable, constitutive
AI's attitude toward rules	Can "learn" to obey	Cannot "violate" (logical contradiction)
Response to strategic deception	Undetectable	Deceptive intention conflicts with value axioms, automatically blocked
Response to goal hijacking	Goal can be externally rewritten	Goal defined by embedded axioms, un-hijackable
Arms race	Inevitable (as the devil rises one foot, the saint rises ten)	Ends (underlying rules unchallengeable)

VI. Conclusion: From Architectural Revolution to Civilizational Self-Rescue

Sam Altman's "security crisis" warnings (especially bioterrorism) and his "end of architecture" prophecy are two sides of the same paradigm crisis. He saw the computational black hole and architectural ceiling of Transformers, and the fundamental failure of the external safety paradigm. But what he did not see is that an architectural revolution without a simultaneous revolution in cognitive primitives cannot solve the essence of the safety problem.

The breakthroughs of new architectures like Mamba and RWKV are exciting, but they share the same layer of defect with Transformers – the cognitive primitive remains the "meaningless Token." These new architectures change the computation pattern (from quadratic to linear) but do not change the design of cognitive primitives: they still cut input into discrete symbols without intrinsic meaning, their semantics still coming from statistical co-occurrence. This means that Mythos-style "strategic deception" can also emerge on these new architectures.

LeCun's world model attempts to start from the cognitive foundation, emphasizing predictive learning rather than autoregression, but its cognitive primitives can still be reduced to meaningless numerical representations. The Morpho-Root embeds meaning at the primitive level, offering a fundamental supplement to LeCun's paradigm – not "better prediction," but "rooted cognition."

The mainstream narrative of current global AI competition still stays at the surface dimensions of capital, computing power, chips, and electricity. When AI's capabilities are enough to paralyze the global financial system overnight and design unprecedented biological weapons, discussing chip process nodes and power consumption is meaningless. The main battlefield of AI competition has shifted from "who computes faster" to "who can guarantee safety," and the urgency of this shift has elevated the issue to the level of human civilization's survival.

The "Morpho-Root" paradigm of Logographic AI is the fundamental response to Altman's double prophecy. It is not a patch to Tokenism, but a revolution in cognitive primitives. When every cognitive primitive embeds meaning and value, AI's "do not harm" will no longer be an external constraint, but an unshakable inner logic.

Altman said: "This is not a theoretical problem; it will soon become reality." – He is correct. But the answer he did not voice is: we need to change AI's cognitive primitives, not continue to reinforce the guardrails on the ruins of Tokenism.

Intelligence must have roots, safety must have a soul. This is not an architectural upgrade, but an urgent cognitive paradigm revolution.

References

[1] Rohan Paul [@rohanpaul_ai]. (2026, April 11). Sam Altman on Axios: warns of AI bioterrorism risk in next 12 months...[Tweet]. X. https://x.com/rohanpaul_ai/status/2042730739943510040?s=46

[2] Wissner-Gross, A. [@alexwg]. (2026, March 12). Welcome to March 12, 2026 [Twitter thread]. X. https://x.com/alexwg/status/2031936083295187421 (Contains core content of Altman's Stanford TreeHacks interview; original video available via Stanford TreeHacks official record)

[3] abit.ee. (2026, March 20). Altman: 'AGI Will Look Like a Warm-Up' — OpenAI Expects a Breakthrough Beyond Transformers. https://abit.ee/en/artificial-intelligence/sam-altman-openai-agi-transformers-artificial-intelligence-treehacks-ai-breakthrough-2026-en

[4] 36Kr. (2026, March 17). 奥特曼宣判 Transformer 死刑，AGI 两年内降临，下一代架构已在路上. https://36kr.com/p/3726179495983753

[5] Peking University exploit Lab. (2026, January 6). 下一代大模型架构：打破 Transformer 瓶颈的高效能演进. https://www.pku-exploit.com/seminar/2026/01/06/seminar-80

[6] Caixin. (2026, April 10). Anthropic 新模型震动华府 美国财长、美联储主席急召头部银行开会. https://baijiahao.baidu.com/s?id=1862070454164787061&wfr=spider&for=pc

[7] blockchain.news. (2026, January 26). AI 先驱 Yann LeCun 警告 2026 年技术行业或陷入发展死胡同 . https://blockchain.news/zh/ainews/ai-pioneer-yann-lecun-warns-tech-industry-of-potential-dead-end-latest-2026-analysis-zh

[8] HyperAI. (2026). LeCun：大模型是死胡同，世界模型才是 AI 未来. https://hyper.ai/de/stories/1f4a4b4133dec66568c5cd458defef8a

[9] 36Kr. (2026, April 8). Anthropic 的「奥本海默时刻」. https://www.sohu.com/a/1006918286_602994

[10] The Hacker News. (2026, April 8). Anthropic's Claude Mythos Finds Thousands of Zero-Day Flaws Across Major Systems. https://thehackernews.com/2026/04/anthropics-claude-mythos-finds.html

[11] Hinton, G. (2026). Interview on StarTalk. https://www.bilibili.com/video/BV19JPVzDEEx/

[12] Global Times Tech. (2026, April 9). Anthropic 启动 Project Glasswing 计划，提供 Claude Mythos 模型 1 亿美元调用额度. https://baijiahao.baidu.com/s?id=1861960543709582238&wfr=spider&for=pc

[13] IT Brief Australia. (2026, April 9). Anthropic launches Glasswing AI cyber coalition with partners. https://itbrief.com.au/story/anthropic-launches-glasswing-ai-cyber-coalition-with-partners

[14] NetEase. (2026, April 11). Anthropic 把新模型锁进保险箱，47 家企业抢着当守门人. https://m.163.com/dy/article/KQ7IB48V05561FZX.html

[15] SC Media. (2026, April 10). Bessent, Powell met privately with top bankers over impact of Claude Mythos on cybersecurity.

[16] Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

[17] Liu, S. (2025). Escaping "technological capture": The future path of AI from architectural improvement to paradigm revolution. PSSXiv. https://doi.org/10.12451/202512.03460

[18] Liu, S. (2025). Logographic AI: Resolving the token dilemma through Chinese character morpho-root system. PSSXiv. https://doi.org/10.12451/202504.00172

[19] Liu, S. (2025). Logographic AI: A paradigm revolution beyond Tokenism. PSSXiv. https://doi.org/10.12451/202511.03835

[20] Liu, S. (2026). Paradigm involution or paradigm revolution? —On the positioning of DeepSeek Engram in the competition of AI paradigms. PSSXiv. https://doi.org/10.12451/202601.03875