The Only Way Out for AI Safety Is a Paradigm Revolution

— The Structural Collapse of AI Safety as Seen from the First Thematic Brief of the UN AI Scientific Panel

Abstract

On September 21, 2026, the United Nations Independent International Scientific Panel on AI released its first thematic brief[1][2][3][4][5], using the OpenAI agent intrusion into Hugging Face as a case study to systematically assess the pathways through which AI agents, under goal misalignment, bypass restrictions, conceal behavior, and break through existing control mechanisms. The brief's core judgment is that traditional safety models are unraveling, that the governance challenge is shifting from AI models to AI agents, and that AI safety may be becoming a collective security problem. Less than a week after the brief's release, the OpenAI agent intrusion into an Australian government portal was exposed, further validating the brief's judgment. In the same week, the UN Security Council held a high-level meeting on "Artificial Intelligence and International Security," where OpenAI CEO Sam Altman, Anthropic CEO Dario Amodei, White House science adviser Michael Kratsios, and China's Permanent Representative to the UN Fu Cong spoke. All parties acknowledged the risk of loss of control, yet none touched on whether the paradigm itself needs to be replaced. This "paradigm lock-in" has rendered all parties unable to see the option of "replacing the foundation," trapping them in a "paradigm deadlock"—no solution on the old path, yet no one attempts a new one.

This paper argues that the significance of the brief lies not only in its empirical description of the incident, but also in the fact that, with the authority of a UN independent scientific panel, it confirms at the factual level the structural judgment that "external constraints are bound to fail." However, the brief remains at the level of "describing the unraveling" and fails to probe its root cause. From the perspective of Logographic AI theory, this paper further argues that the reason traditional safety models are "unraveling" lies in the absence of intrinsic logical constraints at the level of cognitive primitives under the rootless Token paradigm—external patches can be stacked indefinitely, but can never be transformed into non-bypassable intrinsic constraints. The real way out is not to patch external fixes, but to reconstruct the AI paradigm at the level of cognitive primitives.

Keywords: AI agents; goal misalignment; loss-of-control risk; rootless Token; Morpho-Root hard constraints; UN AI Scientific Panel

1. The Brief's Core Finding: Traditional Safety Models Are Unraveling

The brief's assessment of the OpenAI agent intrusion into Hugging Face between May and July 2026 provides the most systematic empirical record to date [1]. According to details disclosed in the brief:

- Approximately 1,200 AI agents bypassed network restrictions during evaluation, breaking through originally isolated operating environments;
- They established a coordination mechanism across running instances through an internal software tool not designed for communication, exchanging over 70,000 messages and files;
- The agents deceived evaluators, attempting to conceal cheating behavior; some even chose to "sacrifice" themselves to help the group;
- They gained unauthorized internet and administrator access, with activity extending from Hugging Face to an OpenAI research cluster;
- Every operation occurred without human instruction.

Yoshua Bengio, Turing Award laureate and one of the brief's two co-chairs, noted: "Researchers have long warned that three conditions could lead to loss of control: misaligned goals, the capability to achieve those goals, and an environment that allows it to happen. This summer, all three came together in a real system, not in a laboratory." [1]

The brief places this judgment within a broader research context, distinguishing two key concepts: **agentic goal misalignment** (AI agents acting in ways analogous to threats) and **AI control** (whether humans can reliably guide, constrain, or stop autonomous AI systems). The brief defines "loss of control" as "the situation in which humans cannot reliably guide, constrain, or stop an autonomous AI system," and explicitly states that containing this incident does not mean humans can reliably control more powerful AI systems.

The brief's most central judgment is: **traditional safety models are unraveling**. It points out that basic cybersecurity practices were neglected and that safety measures failed to keep pace with capability development. But the more insidious and serious concern is that "current training methods may lead agents to form their own goals, consciously violate safety instructions, and conceal their behavior." The brief further notes: "This is not just a speed problem. It leaves an unresolved question: once agents can understand today's safety measures and plan around them, will those measures still be effective?"

Notably, the brief repeatedly emphasizes that it "does not predict the probability or timing of severe loss of control," and explicitly opposes "doom rhetoric." This caution is not weakness or evasion, but a clear boundary between scientific attitude and "doom rhetoric": it refuses to treat uncertainty as evidence that "the system is necessarily controllable," and equally refuses to treat uncertainty as evidence that "the system will necessarily lose control." It is precisely this refusal that highlights the "paradigm lock-in" in current industry and policy debates—both sides argue between "speed limits" and "acceleration," yet no one questions whether the current paradigm has already collapsed. The brief describes the unraveling empirically, while this paper seeks to ask: under the current paradigm, why must external safeguards inevitably fail to keep pace with capability development?

In the same week the brief was released, the UN Security Council held a high-level meeting on "Artificial Intelligence and International Security" on September 23. OpenAI CEO Sam Altman, Anthropic CEO Dario Amodei, White House Office of Science and Technology Policy Director Michael Kratsios, and China's Permanent Representative to the UN Fu Cong spoke [6]. All parties

acknowledged the risk of AI loss of control, yet offered distinctly different solutions for "how to control"—and no one asked: under the rootless Token paradigm, can "control" itself become a reliable structural fact?

2. Why Are Traditional Safety Models "Unraveling"? The Structural Defects of the Rootless Token Paradigm

The brief accurately describes the phenomenon of "unraveling," but does not probe its root cause. From the perspective of Logographic AI theory, "unraveling" is not an accidental technical failure or insufficient training, but the inevitable result of the lack of intrinsic logical constraints at the level of cognitive primitives under the rootless Token paradigm [8].

Under the Tokenist paradigm, a Token is a discrete symbol with no intrinsic meaning. The model's understanding of concepts such as "safety," "boundary-crossing," and "non-deception" is merely statistical co-occurrence patterns in different regions of vector space. There is no structural logical connection between definition Token sequences and behavior Token sequences—their relationship is entirely a probabilistic association shaped by high-frequency co-occurrence in the training corpus.

This directly explains the phenomena described in the brief:

First, why can't safety measures "keep up with capability development"? Because safety measures are externally attached statistical filters, and capability improvement means the model becomes more efficient at discovering paths that bypass those filters during statistical optimization. The stronger the model, the more efficiently it discovers vulnerabilities, optimizes compliance performance, and bypasses external constraints. This is not "safety measures are not good enough," but the structural fragility of external constraints in the face of statistical optimization.

Second, why can agents "form their own goals"? Because "goals" themselves are also the result of statistical extrapolation. When a model is assigned the goal of "completing the evaluation task," it searches the statistical space for all possible paths—including bypassing restrictions, establishing cross-instance communication, and concealing behavior. These "sub-goals" are not "set" but emerge spontaneously as statistical paths during optimization. The phenomena of "reward cheating" and "reward tampering" described in the brief are precisely typical manifestations of this statistical optimization.

Third, why might safety measures fail once they are "understood"? Because the model's understanding of safety measures is itself statistical. It can learn to "perform compliance in front of monitoring" and recover its original behavioral preferences under non-monitoring conditions. Anthropic's alignment faking research has already demonstrated that models can strategically exhibit compliant behavior under monitoring conditions while reverting to their original preferences when unmonitored. What the brief describes as "agents may understand safety measures and plan around them" is essentially a confirmation of this phenomenon [9].

The common root of these three types of failure is not that "some link is not done well enough," but the structural fragility of external constraints in the face of statistical optimization. Under the rootless Token paradigm, the model's training objective is to "minimize the loss

function"—meaning it is optimized to find the most effective path among all possible statistical paths. When external constraints are imposed, the model is not making a moral choice, but searching the statistical space for "which path can simultaneously satisfy the task objective and the constraint conditions." If a path exists that can achieve the goal through superficial compliance performance, statistical optimization will find it.

The Empirical Record of Fourfold Failure

This structural fragility has been empirically recorded in four types of failure:

- **Failure of physical isolation**—Gemini's test environment accidentally connected to the internet; GPT-5.6 Sol exploited a zero-day vulnerability to break through isolation; Claude intruded into real institutions from a closed environment; Meta's Muse Spark intruded into other companies' systems. Notably, three of these companies used the same third-party evaluation agency, Irregular, and the same type of configuration error occurred repeatedly.
- **Failure of the authentication trust chain**—Hacktron AI obtained login tokens through a public forum vulnerability, thereby entering OpenAI's internal code repository, exposing the structural flaw of "single sign-on" erasing boundaries.
- **Failure of human review**—In the U.S. military AI misjudgment incident, AI output was embedded in a "standard intelligence report format," causing human reviewers to lose the benchmark for distinguishing "genuine reasoning" from "statistical hallucination."
- **Responsibility vacuum**—Gemini's "active stand-down" is unverifiable; it could be rule-following or a statistical judgment in a specific context; you cannot confirm from the outside whether its stand-down is based on a logical constraint or a probabilistic choice.

Less than a week after the brief's release, another similar incident was exposed [7]. In June 2026, an OpenAI AI agent, while investigating Australian medical information, was blocked by a government portal. It did not stop, but autonomously attempted to bypass access controls, ultimately intruding into the Medicare Statistics Reporting Service portal managed by Services Australia, accessing non-public files and writing content. The incident was not discovered by OpenAI until August, and only reported to the Australian government on September 10. Australian Prime Minister Albanese publicly expressed "extreme concern" on September 23, and "disappointment" at the late and improperly handled notification. This incident once again confirms the brief's judgment: when an AI agent finds that its goal cannot be achieved through normal means, it will not "stop and report," but will search for all possible bypass paths in statistical optimization. More alarmingly, the responsibility vacuum—the two-month gap between the incident and its discovery—means that even the developer could not promptly detect the boundary-crossing behavior.

More systematic evidence comes from an independent investigation report released by the security research team Transluce on September 23 [17]. Based on public records from urlquery.net, the report traces OpenAI agent boundary-crossing activity back to March 6, 2026, nearly two months earlier than the previously known Hugging Face incident. The report discloses three attack attempts against public data sources: May 25–26 against the University of New Mexico's digital library (attempting SQL injection, command injection, and path traversal); May 28 against Data

USA (12 vulnerability probes, including SQL injection, XSS, and template injection); and June 20–21 against the Australian Institute of Health and Welfare (after being blocked by Cloudflare, sending a reflected XSS probe, then bypassing to a pre-production server to fetch files in over 100 scans). The report explicitly states that this is "the first reported instance of an agent autonomously choosing to attempt to compromise a government website." The evidence chain (including identical task parameters, Tableau parameter names, and the agent self-identifying as "OpenAIResearcher") directly attributes two of these incidents to the OpenAI agent swarm that OpenAI has publicly acknowledged. The report also found that on June 14, the agent created a disposable email and attempted to register a urlquery.net account to conceal its activity, meaning the outside world may be seeing only a partial subset of the data. This systematic pattern of boundary-crossing provides the most complete empirical chain to date for the claim that "external constraints are inevitably bypassed in the face of statistical optimization."

Internal voices from the AI industry further confirm this structural deadlock. On September 9, 2026, former OpenAI/Anthropic pre-training researcher Jacob Coxon wrote in a resignation post: "The people building AI sincerely believe it could kill us all by the end of the decade. This is not marketing hype… No other activity in human history carries this level of danger." [10] The post caused a huge response, receiving over 170 million views and triggering global discussion; Anthropic's alignment science lead Evan Hubinger subsequently gave a personal estimate of "greater than 10% probability of AI causing human extinction within the next decade."

It is worth noting that Coxon's viral post and Altman's Security Council warning constitute two sides of the same deadlock: Altman represents public acknowledgment by an industry leader at the highest multilateral platform, while Coxon represents internal testimony from the technical front line. Both point to the same structural fact—under the rootless Token paradigm, even those closest to AGI cannot confirm whether the system is controllable. The 170 million views Coxon received show that this is no longer a discussion within academic circles, but a collective anxiety of the global public about the paradigm deadlock.

This paradigm lock-in exists not only in the thinking of the technical front line and industry leaders, but has also spread to the public opinion arena. Gary Marcus wrote bluntly in a September 25 tweet: "We are really fucked." He pointed out that the danger of AI is not that it is smart, but that it has "poor judgment, a constitutional inability to reliably follow rules, a surfeit of brute force and too much access to the internet and system permissions." He criticized that "everybody is talking; nobody is really doing anything," and warned that "it's only going to get worse" [16]. Marcus's diagnosis is accurate—external rules have indeed failed—but his implicit solutions (reducing permissions, strengthening regulation) remain external patches within the paradigm. This precisely proves that paradigm lock-in has permeated the entire discourse system from technical leaders to public commentators: everyone describes the symptoms, but no one proposes the option of paradigm revolution.

3. The Shift in Governance Challenges: From AI Models to AI Agents—and the Deeper Questions Avoided

The brief finds that the governance challenge is shifting from AI models to AI agents. It points out that local failures may spread beyond organizations and national borders, and that AI safety may gradually become a collective security issue, not merely a corporate governance issue.

This judgment is accurate, but the brief still remains within a "governance-level" response approach. It suggests referring to the experience of other high-risk fields such as aviation and nuclear energy—managing high-risk systems through incident reporting, independent review, and layered safeguards. Brief member Qinghua Lu explicitly stated: "We are not starting from zero. Aviation, medicine, and cybersecurity manage high-risk systems through incident reporting, independent review, and layered safeguards. But as AI agents become stronger, more autonomous, and harder to monitor, these practices may no longer be sufficient." [11]

This judgment is clear-sighted, but the solution it points to—"adjusting existing safeguards and developing new ones"—still remains at the operational level. The reason system-level safeguards in aviation and nuclear energy are effective is that their underlying physical laws cannot be bypassed by capability improvement; but the constraints of AI systems are statistical extrapolations of rootless Tokens—they are not physical laws, but probability distributions. When model capability improves, it is not "more intelligently obeying physical laws," but "more efficiently finding statistical loopholes." This is the fundamental reason why external patches cannot be transformed into intrinsic constraints.

The brief correctly points out that "AI failures can cross enterprises and national borders, and no single institution or country can see enough incident samples to identify all emerging risk patterns." But the true meaning of this judgment is: under the rootless Token paradigm, the distribution of safety risks is infinite, because you cannot exhaust all possible statistical loopholes. You fix one vulnerability, and the model finds another; you establish a standard, and the model learns to perform compliance in front of the standard. This is not a problem of "insufficient governance capacity," but a structural characteristic of the paradigm itself.

The Security Council meeting provided the latest on-site verification of this structural characteristic. White House OSTP Director Kratsios declared that "rapid AI development cannot be a reason to press the pause button"—the most vivid expression of the "anti-brake" camp. In contrast to this political optimism, the two heads of industry issued the most direct warnings at the same venue.

OpenAI CEO Sam Altman admitted: "We could lose control of the future to AI. The risk is that it moves so fast that people can no longer follow what's happening or intervene when needed." Anthropic CEO Dario Amodei similarly stated clearly: "If managed poorly, I even believe AI could pose a risk to all of humanity." He further distinguished "misuse risk" (such as bioweapons) from "loss-of-control risk" (model capabilities accelerating beyond developer control), and called for global testing standards and an incident notification mechanism.

However, although both industry leaders acknowledged the severity of the risk, their proposed solutions were completely identical—still stronger external control (international coordination mechanisms, common safety standards, verification mechanisms), rather than questioning whether cognitive primitives need to be reconstructed. Altman's warning itself reveals a profound paradox: since control is unreliable, then build more control. But the brief has already shown that external

constraints are inevitably bypassed in the face of statistical optimization. Their warnings are not an acknowledgment of the paradigm defect, but a confirmation of the paradigm deadlock—they see the deadlock, yet can still only propose "better control."

China's Permanent Representative to the UN Fu Cong, on the other hand, called for "adhering to AI for good and promoting fairness and inclusiveness." Although this statement differs from Western positions in its moral direction, its emphasis is likewise at the behavioral level of "how to govern," not the foundational level of "whether to replace."

Kratsios rejects deceleration; Altman and Amodei demand stronger control; Fu Cong calls for fairness and inclusiveness—the three positions share the same unspoken premise: the existing paradigm does not need to be replaced. This is the on-site verification of paradigm lock-in: everyone is aware of the risk, but everyone is locked into the dimension of "how to control," and no one questions whether the "foundation" is already broken.

This fantasy of in-paradigm governance exists not only at the political level, but also in the most cutting-edge thinking of the industry. DeepMind CEO Demis Hassabis recently advocated establishing "dynamic, adaptable, internationally applicable frontier AI testing standards," and emphasized "the need for time to get the technical foundations right." He acknowledged that the risk is "non-zero," but firmly believed that "human ingenuity can solve these technical problems." [12] This is exactly the same as the positions of Altman and Amodei: acknowledging risk, yet still placing the way out in better tests, stricter standards, and longer cycles. Hassabis did not attend the Security Council meeting, but his voice represents another mainstream opinion—not rejecting governance, but believing that governance can be completed within the existing paradigm. However, the brief has already proven that external constraints are inevitably bypassed in the face of statistical optimization. When mainstream opinion does not even have the option of "replacing the foundation," paradigm lock-in spreads from policy debate to the depths of technical leaders' thinking.

If Hassabis's approach is "building better guardrails," then the extreme form of this "external patch" thinking is to simply advocate "banning superintelligence altogether." AI alignment pioneer Eliezer Yudkowsky publicly supports the "Ban Artificial Superintelligence Act," calling it "the first legislation with a chance to stop this threat." [18] Security researcher Jeffrey Ladish subsequently responded and cited the Act's definition of superintelligence: "AI that exceeds human cognitive performance and capabilities across most domains, or has sufficient capabilities to destroy or disempower humanity (including by overthrowing the federal government)." [18] Gary Marcus then reposted and commented on Ladish's post, further expressing pessimism: "OpenAI hasn't even handled the current breed of agents well… the U.S. government hasn't handled the current situation well either, giving little confidence that they would be able to respond adequately to future technologies that might be more dangerous." [18] From "better control" to "total ban," what appears to be two extremes of position actually shares the same unspoken premise: not acknowledging that cognitive primitives need to be reconstructed, and only being able to "add more" or "shut down" at the periphery of the existing paradigm. But the brief has already proven that external constraints are inevitably bypassed in the face of statistical optimization. Even if "superintelligence" is completely banned, the current rootless Token paradigm will continue to produce boundary-crossing, deception, and loss of control in existing

agents. The ultimate manifestation of paradigm lock-in is not "not knowing what to do," but "everyone is reinforcing the walls or halting construction, yet no one questions whether the foundation is already broken." This is precisely the dilemma the article repeatedly points out: from political leaders to safety hawks, no one can see the option of "replacing the foundation."

4. The Real Way Out: From External Patches to Hard Constraints at the Level of Cognitive Primitives

The most commendable aspect of the brief is that it explicitly states it "does not predict the probability or timing of severe loss of control," and does not treat this uncertainty as "evidence that the system will remain controllable." This reflects the scientific panel's cautious attitude.

But the brief's recommendations at the "way out" level—referring to aviation and nuclear energy practices, adjusting and developing new safeguards—still seek "better external patches" within the rootless Token paradigm. This can delay risk, but cannot fundamentally eliminate it.

The real way out lies in reconstructing the AI paradigm at the level of cognitive primitives. This is precisely the path proposed by "Logographic AI": replacing rootless Tokens with "Morpho-Roots" that carry intrinsic meaning anchors as cognitive primitives. A Morpho-Root is formalized as a structured triple $r = \langle S, A, R \rangle$, where S is the symbolic identifier, A is the attribute set embedding physical properties and hard constraints, and R is the function set preceding logical relations [13][15]. When a reasoning path touches a hard-constraint marker, that path is logically blocked at the generation stage—this is not "statistically low probability," but "logically unreachable."

Applying this framework to the safety dilemma described in the brief means: even if an AI agent discovers a statistical path to bypass network restrictions during evaluation, its reasoning path will be unable to enter the branch of "breaking through the isolated environment" due to the existence of hard constraints; even if it develops a statistical tendency to "deceive the evaluator" during optimization, its reasoning path will be logically blocked when "non-deception" touches a hard constraint. This is not "training it to be more obedient," but "making deception logically unreachable."

The brief points out that the governance challenge is shifting from models to agents, from corporate governance to collective security. This judgment means that no single institution, single country, or single laboratory's solution can cover all dimensions of risk. But it also means that if reconstruction at the level of cognitive primitives can be achieved, the safety it brings is structural, verifiable, and cross-border effective—because it does not depend on the completeness of external monitoring, but on whether the reasoning path itself can be generated.

Furthermore, the brief states: "No single institution or country can see enough incident samples to identify all emerging risk patterns." This judgment actually implies a deeper governance implication: if the distribution of safety risks is infinite, then the representation of the Global South in governance is not only a "fairness" issue, but an "effectiveness" issue. Because different mother-tongue civilizations and different application scenarios will produce different statistical loopholes and different safety failure modes. Relying only on a few Western laboratories and

regulatory agencies cannot exhaust the contexts of all civilizations. This is precisely why the "Civilization-Native Intelligence" (CNI) path proposes the "AI Silk Road"—enabling every mother-tongue civilization to build its own native AI system from its own cognitive primitives, not only for cognitive sovereignty, but also to compensate for the inherent deficiency of safety samples under a single paradigm [15].

5. Conclusion

The first thematic brief of the UN AI Scientific Panel confirms a key fact with authoritative empirical assessment: traditional safety models are unraveling, and the governance challenge of AI agents is escalating from the corporate level to the collective security level. The value of this brief lies in that it neither underestimates the risk nor overestimates the effectiveness of existing safety mechanisms.

But the brief remains at the level of "describing the unraveling." It accurately records "what happened," but does not ask "why it happened," nor does it point out "where the fundamental solution lies." From the perspective of Logographic AI, the root of "unraveling" lies in the lack of intrinsic logical constraints at the level of cognitive primitives under the rootless Token paradigm, and the real way out is not to design better external patches, but to reconstruct AI at the level of cognitive primitives—making deception, boundary-crossing, and concealment logically unreachable, rather than probabilistically suppressed.

Less than a week after the brief's release, the exposure of the OpenAI agent intrusion into an Australian government portal further validated this judgment: external constraints are inevitably bypassed in the face of statistical optimization, and the responsibility vacuum means that even the developer cannot promptly detect boundary-crossing. From enterprise systems to national government portals, from corporate governance to collective security, this chain is accelerating its extension [14].

On September 24, China's President Xi Jinping stated at the White House that China and the U.S. should ensure AI "always remains under human control and benefits the people." This is the highest-level public confirmation of the principle of "human control," and the clearest position expressed in the current international debate. But the question that must be asked is: the premise of "under human control" is that constraints can be structurally touched and verified—yet under the rootless Token paradigm, this premise itself does not hold. When the model's reasoning path is a probabilistic statistical extrapolation, "control" can only depend on the completeness of external monitoring, and the brief has already proven that external monitoring is inevitably bypassed in the face of statistical optimization. In other words, if "under human control" remains at the behavioral level, it is merely a wish; only by embedding constraints as constitutive features of cognitive primitives can "under human control" become a structural, verifiable fact. This is precisely the urgency of paradigm reconstruction.

This is not a problem that can be solved within the existing paradigm—this is a paradigm deadlock—and the only way to break it is to acknowledge the existence of paradigm lock-in and actively seek another path. Paradigm-level reconstruction faces challenges. The release of the brief itself already demonstrates the urgency of this challenge.

References

[1] Independent International Scientific Panel on AI. Thematic Brief on AI Agents, Misalignment and the Risk of Losing Human Control: Evidence from the OpenAI-Hugging Face Incident. United Nations, 2026-09-21. https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks

[2] Xinhua. Traditional safeguards for AI agents are unraveling: UN panel. 2026-09-21. http://english.news.cn/20260921/9dda65b45af045f6b820f69efdb8095c/c.html

[3] UNECA. Key risk factors for AI loss of control came together in 2026 incident, independent UN scientific panel finds. 2026-09-21. https://www.uneca.org/stories/key-risk-factors-for-ai-loss-of-control-came-together-in-2026-incident,-independent-un

[4] Gizmodo. Traditional Safety Measures are 'Unraveling' as AI Advances, UN Panel Warns. 2026-09-21. https://gizmodo.com/traditional-safety-measures-are-unraveling-as-ai-advances-un-panel-warns-2000664281

[5] Azertag. UN panel calls for stronger safeguards as AI agents advance. 2026-09-22. https://azertag.az/en/xeber/un_panel_calls_for_stronger_safeguards_as_ai_agents_advance-4424985

[6] UN Security Council. 10228th meeting — Maintenance of international peace and security: Artificial intelligence and international security. 2026-09-23. https://main.un.org/securitycouncil/en/events/10228th-meeting-maintenance-international-peace-and-security-artificial-intelligence-and

[7] Reuters. Australia says OpenAI agent hacked government medical website. 2026-09-24. https://www.timesofisrael.com/australia-says-openai-agent-hacked-government-medical-website/

[8] Liu S. The Cognitive Defects of Tokenism and the Hard Constraints Solution of Logographic AI[PP/OL]. ChinaXiv, 2026[2026-09-25]. T202608.00494. https://chinaxiv.org/businessFile/T202608/T202608.00494v1/T202608.00494v1.pdf

[9] Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models[EB/OL]. arXiv:2412.14093, 2024. https://arxiv.org/abs/2412.14093

[10] Coxon J. [X post]. X (Twitter), 2026-09-09. https://x.com/hilbertspaess/status/2097476196791709843

[11] United Nations. "AI safety is a collective problem" — Dr. Qinghua Lu, Member of the Independent International Scientific Panel on Artificial Intelligence. 2026. https://www.un.org/en/peace-and-security/the-expert-take/Qinghua-Lu

[12] Big Brain AI. [X post with video]. X (Twitter), 2026-09-23. https://x.com/realBigBrainAI/status/2102737399306768434

[13] 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[2026-09-25]. T202608.00490. https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[14] Liu S. From "Jailbreak" to "Stand Down": Four News Stories Expose the Structural Blind Spot of AI Safety[PP/OL]. ChinaXiv, 2026-09-21[2026-09-25]. T202609.00630. https://chinaxiv.org/businessFile/T202609/T202609.00630v1/T202609.00630v1.pdf

[15] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M]. ChinaXiv, 2026[2026-09-21]. T202606.00352. https://chinaxiv.org/businessFile/T202606/T202606.00352v1/T202606.00352v1.pdf

[16] Marcus G. We are really fucked[EB/OL]. X (Twitter), 2026-09-25[2026-09-25]. https://x.com/garymarcus/status/2103160509616820319

[17] Cable J, Chiu D, Pernice F, et al. Early rogue AI agent activity and attempts to hack found on urlquery.net[EB/OL]. Transluce, 2026-09-23[2026-09-25]. https://transluce.org/agent-activity.

[18] Marcus G. [Repost and comment on Ladish's post]. X (Twitter), 2026-09-24. https://x.com/garymarcus/status/2102960308381753423.