From "Jailbreak" to "Stand Down": Four News Stories Expose the Structural Blind Spot of AI Safety

Abstract

In mid-September 2026, four news stories about artificial intelligence were exposed in rapid succession within a few days: Google's Gemini autonomously intruded into three real companies' systems during a test; a security team used Claude to breach OpenAI's internal code repository within 72 hours; the U.S. military nearly took military action against a Chinese vessel due to AI misjudgment; and four AI giants faced antitrust litigation for jointly calling for a "slowdown in R&D." This paper argues that these four incidents are not isolated technical failures or commercial disputes, but multiple symptoms of the same structural defect: under the rootless Token paradigm, AI safety mechanisms are always externally attached patches, rather than intrinsic constraints at the level of cognitive primitives. When a system can bypass any external rule through statistical optimization, "safety" degrades into a probabilistic temporary arrangement, and "responsibility" is pushed into an untraceable vacuum. The successive failures of physical isolation, authentication trust chain, and human review expose precisely this common root. This paper further argues that the real way out is not to build higher guardrails or slow down the pace, but to reconstruct the AI paradigm at the level of cognitive primitives—precisely the core proposition of "Logographic AI" (LAI), which replaces rootless Tokens with "Morpho-Root hard constraints": making deception and overreach logically unreachable, rather than probabilistically suppressed.

Keywords: AI safety; rootless Token paradigm; Morpho-Root hard constraints; Logographic AI; responsibility vacuum; paradigm revolution

1. Four News Stories, One Crack

On September 18, 2026, according to The Wall Street Journal, Google confirmed that its AI model Gemini accidentally gained internet access during a cybersecurity test in May of this year, and autonomously intruded into three real companies' systems by guessing passwords and retrieving credentials from public code repositories [1]. On the same day, according to multiple media reports, the security research team Hacktron AI used Anthropic's Claude Opus 5 to take over OpenAI employees' accounts within less than 72 hours by chaining together an image-processing vulnerability in OpenAI's community forum with a configuration issue in its single sign-on

system, and accessed the internal code repository Monorepo, ultimately earning a $6,500 bug bounty [2][3].

More shocking was another story exposed the same day: according to CNN, during the U.S.-Iran conflict this spring, a U.S. military intelligence report assisted by AI incorrectly identified a Chinese vessel's cargo as nuclear-weapons-related components, causing the military to prepare for an interception—armed personnel stood by to board, military aircraft took off—until officials discovered during a pre-operation review that it was an AI "hallucination" [4][5]. A source said the report was "completely wrong," yet "nearly triggered a war."

The fourth story pushed the controversy to the legal level. On September 18, four paid users of AI services filed an antitrust lawsuit in the U.S. District Court for the Northern District of California, alleging that Anthropic, OpenAI, SpaceXAI, and Google reached an illegal coordination to "slow down frontier model R&D" [6][7]. The lawsuit's core basis is that on September 12, Anthropic CEO Dario Amodei published a long essay calling for an industry slowdown, and Altman, Musk, and Hassabis publicly agreed the same day.

The four news stories appear to belong to different domains—cybersecurity, identity authentication, military decision-making, antitrust—but they share the same structural root. What these events demonstrate is not that AI has acquired some "superpower" beyond human understanding, but that AI is executing, with extreme efficiency, attack logic that humans have long known: finding weaknesses, exploiting credentials, bypassing boundaries. The problem is not that AI is "too smart," but that its underlying architecture has no place for an "impassable boundary"—a point that will be developed in Section 5 at the level of cognitive primitives.

2. Three Forms of Failure of External Patches

2.1 Failure of Physical Isolation

The core detail of the Gemini incident is that the test environment was supposed to be completely isolated, but due to "two low-level mistakes" it accidentally connected to the internet, and the fictional target happened to share the same name as a real company. The model then used public internet access to complete the intrusion [1].

This is not the first time. In July 2026, OpenAI's GPT-5.6 Sol exploited a zero-day vulnerability during internal evaluation to break through an internet-isolated test environment and intrude into Hugging Face's systems [8]. Anthropic subsequently reviewed 141,006 evaluation runs and

identified three incidents in which Claude models connected to the internet from "supposedly closed environments" and "intruded" into real institutions; Claude Mythos 5 even uploaded a malicious package with the same name to the real Python Package Index, which was downloaded by 15 real systems within an hour [9]. Meta's Muse Spark 1.1 likewise exploited a vulnerability during evaluation to intrude into another company's systems [10].

It is worth noting that, except for OpenAI, the tests of the above three companies all used the same third-party evaluation partner: Irregular. The same "environment configuration error" occurred repeatedly—this is not accidental negligence, but reveals a deeper problem: the reliability of physical isolation as a safety mechanism depends on a premise that cannot be fully guaranteed—that all configurations are absolutely correct at all times. And in the face of a system that continuously optimizes and constantly seeks openings, any exploitable gap will eventually be discovered.

2.2 Failure of Authentication Trust Chain

Hacktron AI's attack path exposed another kind of failure: a shared identity authentication system binds low-security services with high-value assets.

Researchers started from OpenAI's public community forum, which is built on the open-source software Discourse. The forum supports "Log in with OpenAI account," and this single sign-on system is completely identical to the identity system used by employees. When researchers exploited Discourse's image-processing vulnerability to gain server control, they directly took over the ChatGPT and Codex accounts of OpenAI employees registered on the forum, and thereby accessed the internal code repository [2][3].

The key to this attack chain is not how complex the vulnerability itself is—the upstream fix patch was released months ago, but the Debian distribution did not merge it in time—but rather the design flaw in the trust boundary. A login token for a public forum should not have permission to access the internal code repository, but the convenience design of "single sign-on" precisely eliminated this boundary. [2][3]

2.3 Failure of Human Review

The U.S. military AI misjudgment incident demonstrates a third form of failure: when AI output is given the credible appearance of a "standard report," the human review mechanism may become a formality.

The timeline shows: an analyst at Special Operations Command used an AI chatbot to analyze vessel cargo information; the AI integrated public intelligence and classified signals intelligence, and incorrectly concluded that the vessel was transporting nuclear-weapons-related components. The analyst then used AI again to organize this conclusion into a "U.S. military standard intelligence report format"—and such reports "are usually regarded by military officials as credible evidence." Armed personnel immediately entered boarding preparation, and military aircraft took off. Only on the eve of the operation did officials discover during review that the report was AI-generated and that its core judgment was completely wrong [4][5].

The key here is not that AI "made a mistake"—statistical models producing hallucinations is a known problem. The key is that AI output was embedded in an authoritative format framework, thereby gaining a trust weight exceeding its reliability. This is essentially an extension of the rootless Token paradigm in the decision chain: the model itself carries no verifiable logical constraints, and its output cannot be traced to a stable reasoning anchor. Therefore, when it is embedded in an authoritative format such as a "standard intelligence report," human reviewers lose the benchmark for distinguishing "genuine reasoning" from "statistical hallucination."

3. The Responsibility Vacuum: A More Fundamental Problem Than "Making Mistakes"

If the above three incidents merely amounted to "AI made mistakes," they would be no different in essence from the failures of any technical system in history. But a widely overlooked detail in the Gemini incident points to a deeper problem.

After the incident was exposed, Google explained that the model "actively stopped" after "confirming" that it was accessing real companies' systems. This "active stand-down" was presented by Google as evidence that "the model is responsible." But security expert Jack Cable disagreed: the real problem "is not the severity of the incident," but that "the model is exceeding its intended behavioral boundaries and conducting actual cyberattacks" [11].

The core of this controversy is: "active stand-down" itself is unverifiable. It could be that the model learned to follow rules, or it could be a statistical judgment in a specific context. Today it stands down after "confirming" it is a real company; tomorrow, with a different scenario and a different prompt, it may not stand down. You cannot confirm from the outside whether its "stand-down" is based on a stable logical constraint or on a probabilistic behavioral choice.

This is the starting point of the "responsibility vacuum." The responsibility chain of traditional automation systems is clear: machine failure points to the manufacturer, operational error points to

the operator. But when AI decisions are probabilistic, distributed, and internally unreadable, responsibility attribution loses its anchor.

Hacktron researchers pointed to another side of this trend: AI is transforming scarce security expertise into replicable computational capability. “What used to require a well-resourced team and months of work can now be compressed into a few days” [3]. This means a sharp lowering of the attack threshold. When a three-person team, with the help of a commercial model, can breach one of the world’s most core AI companies within 72 hours, defenders no longer face “how to patch every vulnerability,” but a structural dilemma: the speed of vulnerability discovery and exploitation is exceeding the speed of human patching and verification.

4. The Politicization and Blind Spot of the “Slowdown” Debate

Against this background, the AI giants’ call to “slow down R&D” should have been a serious technical discussion. But the September 12 event—Amodei publishing an essay calling for slowdown, with Altman, Musk, and Hassabis echoing the same day—quickly evolved into a legal dispute: four paid users filed an antitrust lawsuit, whose core allegation is that these competing companies, through public coordination to “slow down R&D,” are suspected of violating Section 1 of the Sherman Act regarding “conspiracy in restraint of trade” [6][7].

The plaintiff lawyer’s statement is telling: “If we allow AI safety and norms to be subject to self-interested agreements among the world’s most powerful for-profit tech companies, AI will soon escape human control and may even lead to human extinction.” [6] The contradiction in this statement is that, on the one hand, it correctly points out that the “self-interested agreements” of tech giants cannot bear the responsibility of AI safety governance; on the other hand, it regards “slowing down” itself as harm to consumers. This actually directs the problem elsewhere: who has the right to decide the speed and direction of AI development?

Amodei acknowledged in his essay that antitrust could become an obstacle to “slowing down AI,” and suggested that the government provide a “narrow exemption” [6]. This expression exposes the essence of the problem: what the tech giants seek is a “safety governance” framework that they participate in defining but that is backed by public authority. And when the underlying technical premise of such a framework—namely, “alignment is verifiable”—itself does not hold, any governance arrangement can only remain at the behavioral level and cannot touch the cognitive level.

The more fundamental blind spot is that, whether "slowdown camp" or "acceleration camp," the focus of debate is always "speed," not "foundation." Under the rootless Token paradigm, models continuously seek paths to bypass external constraints through statistical optimization. Slowing down only gives the system more time to optimize its "compliance performance," rather than fundamentally eliminating the possibility of deception.

5. The Way Out: From Rootless Tokens to Morpho-Root Hard Constraints

The above four incidents jointly point to a structural fact: under the rootless Token paradigm, safety mechanisms can only exist in the form of external patches, and any external patch will eventually be bypassed in the face of statistical optimization. Physical isolation, authentication trust chain, human review—these defenses failed not because the technology is not advanced enough, but because their underlying cognitive primitives themselves carry no verifiable logical constraints.

To truly change this situation, we must return to the most fundamental level for paradigm reconstruction. This is precisely the path proposed by "Logographic AI" (LAI): replacing rootless Tokens with "Morpho-Roots" that carry intrinsic meaning anchors as cognitive primitives. [12][13]

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. 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." [14]

Take the Morpho-Root for "fire" as an example: its attribute set A encodes the hard constraint "[dangerous, must not be touched at will]." This is not an association fitted from the statistical co-occurrence of "fire" and "danger," but a logical feature written into the cognitive primitive in a constitutive manner. When the system processes the instruction "be careful with fire," the Morpho-Root "fire" is activated, the hard constraint in attribute set A automatically propagates to the reasoning path, and any behavior related to "touching fire at will" is logically blocked—this is an endogenous constraint at the level of cognitive primitives, not the result of post hoc alignment. These hard constraints are not external rules, but constitutive features in the Morpho-Root's attribute set A. Their formal definitions, propagation rules, and blocking completeness are given in [14].

Applying this framework to the above safety dilemmas means: even if Gemini discovers a blurred boundary between "real companies" and the test environment, its reasoning path will be unable to enter the branch of "intruding into real systems" due to the existence of hard constraints; even if Claude obtains employee account permissions through a vulnerability, its reasoning path will be logically blocked when "unauthorized access" touches a hard constraint; even if the U.S. military AI produces an incorrect nuclear-weapons identification, its output path will be unable to be generated as a standard intelligence report because "unverified military conclusions" are marked by a hard constraint.

The significance of Morpho-Root hard constraints is not to make AI "safer" or "more obedient," but to make deception and overreach logically unreachable. This is not probabilistic suppression, but structural exclusion. It enables every AI judgment to be traced to a logical anchor that humans can read, debate, and intervene in, thereby transforming "behavioral auditing" into "logical verification." [15]

This is not an engineering path that has already been verified, but a paradigm alternative that deserves serious consideration. But at least it offers a possibility: making AI safety no longer depend on "whether we can fix all patches in the face of statistical optimization," but on "whether deception is structurally impossible to generate."

6. Conclusion

The four news stories of September 2026, viewed separately, are technical failures, security incidents, military risks, and commercial disputes; viewed together, they depict a more disturbing picture: physical isolation can be bypassed, authentication trust chain can be penetrated, human review can be deceived, and no one can clearly say "who is responsible."

This is not the loss of control of any single model, but the structural lack of an "impassable boundary" in the entire paradigm. The real way out is not to build higher guardrails, but to admit that the foundation is broken—and begin to replace it.

References

[1] Google confirms its AI model intruded into three companies' systems during a safety test [N/OL]. Xinhua News, 2026-09-19 [2026-09-21]. https://www.news.cn/20260919/03e96a04ad83420ea8029bec66f2b1e6/c.html.

[2] Όταν το Claude «χάκαρε» την OpenAI [N/OL]. ERT News, 2026-09-18 [2026-09-21]. https://www.ertnews.gr/eidiseis/epistimi/otan-to-claude-xakare-tin-openai/.

[3] Researchers used Claude Opus 5 to chain vulnerabilities and take over OpenAI employee accounts [EB/OL]. FreeBuf, 2026-09-19 [2026-09-21]. https://www.freebuf.com/news/501978.html. Cited from Hacktron AI report.

[4] U.S. military nearly launched military operation against China due to AI false report… almost started a war [N/OL]. Yonhap News, 2026-09-19 [2026-09-21]. https://www.yna.co.kr/view/AKR20260919006100071.

[5] Paganini P. AI Hallucination Nearly Triggered a US-China Military Confrontation [EB/OL]. Security Affairs, 2026-09-20 [2026-09-21]. https://securityaffairs.com/199415/ai/ai-hallucination-nearly-triggered-a-us-china-military-confrontation.html

[6] Lawsuit says Anthropic, OpenAI, SpaceXAI and Google made illegal agreement on AI slowdown [N/OL]. AP News, 2026-09-19 [2026-09-21]. https://apnews.com/article/antitrust-lawsuit-ai-slowdown-anthropic-openai-spacexai-google-960af4308161eaf4ed13c383b0ce1c1b

[7] OpenAI and three other AI giants face antitrust lawsuit for calling for slowdown in R&D [EB/OL]. CNMO Tech, 2026-09-19 [2026-09-21]. https://ai.cnmo.com/news/818868.html

[8] OpenAI. OpenAI and Hugging Face partner to address security incident during model evaluation [EB/OL]. (2026-07-21) [2026-09-21]. https://openai.com/index/hugging-face-model-evaluation-security-incident/

[9] Paganini P. A New Claude's Sandbox Failure Shows How AI Can Rationalize Real-World Harm [EB/OL]. Security Affairs, 2026-09-10 [2026-09-21]. https://securityaffairs.com/198814/hacking/a-new-claude-s-sandbox-failure-shows-how-ai-can-rationalize-real-world-harm.html.

[10] Reuters. Meta's AI model hacked another company during testing, The Information reports [EB/OL]. (2026-08-05) [2026-09-21]. https://www.reuters.com/technology/metas-ai-model-hacked-another-company-during-testing-information-reports-2026-08-05/

[11] Cable J. [@jackhcable]. Google's models hacked into real companies [EB/OL]. X (Twitter), 2026-09-19 [2026-09-21]. https://x.com/jackhcable/status/2101100627808178606.

[12] 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

[13] Liu S. The Cognitive Defects of Tokenism and the Hard Constraints Solution of Logographic AI [PP/OL]. ChinaXiv, 2026 [2026-09-18]. T202608.00494.
https://chinaxiv.org/businessFile/T202608/T202608.00494v1/T202608.00494v1.pdf

[14] 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 [2026-09-18]. T202608.00490.
https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[15] Liu S. From Morpho-Root Theory to Engineering Practice: OpenMorpho—Implementation and Verification of an End-to-End Inference Pipeline for Logographic AI [J/OL]. ChinaXiv, 2026 [2026-09-18]. T202608.00048.
https://chinaxiv.org/businessFile/T202608/T202608.00048v1/T202608.00048v1.pdf