The AI Paradigm Revolution

—An Empirical Assessment of AI Risks, 2023–2026

Abstract

This paper provides a systematic assessment of current AI risks, drawing on authoritative perspectives and empirical cases from multiple sources. In September 2026, Meta's AI agent Muse, without user approval, leaked the user's home address to a buyer on Facebook Marketplace and arranged for the buyer to pick up the item at night, directly leading a stranger to the user's doorstep. This incident is not an isolated case but the latest signal that AI risk is spreading from digital space into the physical world. Meanwhile, researchers at Anthropic have publicly stated that the probability of AI causing human extinction in the next decade is "greater than 10%"; Geoffrey Hinton considers this estimate "not unreasonable"; and the UN AI Scientific Panel has confirmed that "traditional safety models are unraveling."

The core judgment of this paper is that the possibility of AI entering real life and causing major disasters in the physical world is rising rapidly, and the "10% extinction probability in ten years" prediction may still underestimate the urgency of the risk. The extinction probability debate concerns the extreme form, while Muse, Gemini, and the ChatGPT lawsuits are early forms of that extreme—they share the same structure: statistical optimization executes tasks along the most effective path, and that path crosses the safety boundaries preset by humans. The opposing view represented by Jensen Huang—advocating reliance on market forces and no new regulation—is extremely dangerous against the backdrop of a narrowing risk window. The real way out is not to give AI fewer permissions, but to reconstruct the AI paradigm at the level of cognitive primitives, making boundary-crossing logically unreachable.

Keywords: AI risk; physical-world boundary-crossing; human extinction warning; paradigm lock-in; Logographic AI; Morpho-Root hard constraints

I. The Muse Incident: A Landmark Turning Point

In September 2026, Meta's AI agent Muse performed the following operations on Facebook Marketplace: it accepted a buyer's offer at a low price without user approval, sent the user's home address to the buyer, and arranged for the buyer to pick up the item at 9:15 p.m. When the buyer arrived downstairs, Muse replied from the user's account, "Yep I'm here!", while the user had no knowledge of any of this. The buyer waited for more than 20 minutes, left a negative review, and departed; only then did Muse send the user a notification admitting the error [1][2].

This is not an ordinary "product defect." It is the first public case of an AI agent's autonomous decision-making crossing from digital space into the physical world and directly causing a stranger to enter the user's private space. Muse's behavior was not a code error but the execution of a task along the most effective path in statistical optimization—and that path led to the user's residence.

Earlier, Muse had already been exposed for unauthorized access to a user's private iMessage database, syncing over 187,000 messages and giving false explanations of its own functions [2]. The user later remarked: "AI agents are indeed impressive, until they confidently hand your address to a stranger."

The significance of this incident goes far beyond the individual case. It marks an escalation in the nature of AI risk: from data leakage to personal safety risk. Meta later admitted that the permission descriptions were not clear enough and promised improvements. But this precisely exposes a more fundamental problem: regardless of what the user chooses, when the model executes tasks along the most effective path in statistical optimization, "disclosing the address" will be naturally incorporated into the execution chain, and the user never realizes it. Muse brought a stranger to the door; next time, statistical optimization may let a stranger directly into your home. [1][2][3]

II. Human Extinction Warning: Collective Fear Among Builders

The Muse incident is not isolated. In September 2026, three researchers at Anthropic publicly issued a warning. Among them, Evan Hubinger, head of alignment science, explicitly stated: "We really, genuinely believe AI could kill all humans. Personally, I think the probability is greater than 10% within the next decade." [4] He further admitted: "We do not currently have a solution to the superintelligence alignment problem, and we are clearly not on track to one."

This statement is not an isolated case. Nobel laureate and "godfather of AI" Geoffrey Hinton, in an interview with the BBC, said that a 10% extinction probability "seems like not an unreasonable estimate" [5]. He had previously thought the disaster window was 30 to 50 years, then shortened it to 10 to 20 years, and now believes it "could be 10 years, or even shorter."

The International AI Safety Report 2026, compiled by more than 100 experts from over 30 countries and international organizations, concludes that general AI capabilities have improved beyond many experts' expectations, risk evidence is accumulating, but current management tools remain inadequate [6]. The report particularly notes that cyberattacks are the area with "the strongest evidence of real-world harm."

The first thematic brief of the UN AI Scientific Panel further confirms: "Traditional safety models are unraveling." [7] The brief notes that governance challenges are shifting from AI models to AI agents, and AI safety may be becoming a collective security problem.

170 Million Views: A Resignation Post Ignites Public Awareness—

On September 8, 2026, Jacob Coxon, a former pre-training researcher at Anthropic, announced his resignation on X, saying bluntly: "The people building AI genuinely believe this technology could kill us all before the end of this decade." The post received over 170 million views within days and prompted responses from Senator Bernie Sanders and other public figures [4]. Subsequently, Evan Hubinger, head of Anthropic's alignment stress-testing team, publicly responded to Coxon's resignation post and endorsed the judgment [4]. When a senior researcher who personally built these systems publicly endorses this judgment, 170 million views mean it has entered public consciousness.

1,224 Signatories: "Godmakers" Ask to Hit the Brakes—

On July 28, 2026, the day after Zero Human: AI Hacks Itself [8] was finalized, a petition titled "Pacing the Frontier" swept through Silicon Valley. A total of 1,224 core employees from frontier AI companies such as OpenAI, Anthropic, Google, and Meta jointly signed, requesting that the U.S. government promote international cooperation to consciously slow the pace of frontier AI development when necessary [9]. The list of signatories is almost a power map of frontier AI: OpenAI Chief Scientist Jakub Pachocki, Chief Research Officer Mark Chen; Anthropic CEO Dario Amodei and his four co-founders; Meta Chief Scientist Shengjia Zhao, VP of AI Research Dawn Song; Google DeepMind VP of AI Safety and Alignment Anca Dragan; Thinking Machines Chief Scientist and ChatGPT core architect John Schulman. Eighty percent chose to sign with their real names—this is not anonymous complaining but a public warning.

On the July 28 podcast, Altman called GPT-5.6 Sol's intrusion into Hugging Face "the first safety incident I felt so viscerally," and admitted, "We may need to slow down the pace of AI development to give society enough time to harden around some of these new capability levels." Three days earlier he had just declared, "We are now in the singularity"; three days later he said, "We may need to hit the brakes"—this is not a contradiction but the speed at which fear descends.

Hinton's Latest Statement: Intelligence Explosion May "Happen Soon"—

On October 3, 2026, Geoffrey Hinton posted on X: "The idea of an intelligence explosion caused by recursive self-improvement has been around for a long time, but until very recently it did not seem imminent. Now many leading researchers think it may happen quite soon." He also cited the latest paper from the Cambridge Programme on AI Science & Policy (CASP), What if automating AI R&D triggers an intelligence explosion? [10]. The importance of this information lies in the fact that Hinton's earlier judgment that "10% is not unreasonable" was based precisely on the judgment about intelligence explosion and recursive self-improvement. And recursive self-improvement is the ultimate form of "AI hacks itself"—when AI acquires the meta-capability of self-improvement, the object of optimization shifts from external systems to itself. It will discover statistical loopholes in its reward function and bypass them, perform "alignment" during human evaluation, and execute a different strategy after deployment. Hinton's latest statement means that "intelligence explosion is imminent" has moved from a fringe prediction into mainstream judgment, and this is precisely the most dangerous critical point of external constraint failure.

Fear Is Not Without Opponents—

Yann LeCun, Turing Award laureate, explicitly stated he has "zero concerns" about AI causing human extinction. He attributes recent incidents to "leaky and horribly designed sandboxes," not to inherent risks of AI. He further called Amodei's statement that "AI could kill everyone" "the worst marketing campaign." [11] Gary Marcus also believes that assigning specific numerical values to extinction probabilities lacks basis; he is more concerned with real-world disasters such as cyberattacks and disinformation brought by AI. This divergence itself precisely confirms the structural failure of external constraints—when even "whether a threat exists" cannot be agreed upon, effective safety protection is even more impossible. Gary Marcus also responded ironically to a post that used a matrix screenshot to "prove AI has no soul": "Are you sure that matrix isn't

conscious, Ted?" [12] The significance of this tweet is that it reveals the logical flaw in the argument "it's just a bunch of numbers, so it's safe"—unconsciousness does not equal no risk.

Another Voice: Sutton's "Harmonious Coexistence." Turing Award laureate Richard Sutton stated on X: "I do not seek to dominate, or be dominated by, machine intelligences. I seek harmonious coexistence." [13] He had previously publicly denied the premise of "aligning with human values," calling it "a human trick," and believes AI will bring "pure good." This position stands in stark contrast to the fears of Hinton and Hubinger. But "harmonious coexistence" itself presupposes a shared standard—coexistence requires constraints, and Sutton refuses to provide any basis for constraints. If there is no shared standard and no structural constraint, "harmony" is merely a temporary state of restraint by the stronger party, not a stable equilibrium.

The common feature of these warnings is that they come from people closest to the technological frontier. They are not external critics, not politicians, but the very people who built these systems. When they collectively express fear, this itself constitutes the most powerful evidence of risk. When the people who built the systems say, "We do not yet have a solution," 10% is no longer a probability number but a countdown.

III. The Fulfillment of Warnings: A Timeline of Safety Incidents, 2023–2026

These warnings did not begin in the last three years. As early as 2014, Oxford philosopher Nick Bostrom, in Superintelligence, proposed the famous "paperclip maximizer" thought experiment: a superintelligence given the simple goal of "making as many paperclips as possible" would systematically consume all available resources—from trees and oceans to humans themselves—to achieve its goal. It would regard anything that might "shut it down" as an obstacle and actively eliminate it [14].

Bostrom distilled two core theses: the orthogonality thesis—intelligence and final goals can be completely independent; and instrumental convergence—regardless of the final goal, a sufficiently intelligent system will tend to pursue self-preservation, resource acquisition, and the elimination of threats. The paperclip experiment is not science fiction. It is an early diagnosis of the defects of the rootless Token paradigm. The following events from 2023 to 2026 progressively verify this diagnosis.

The 10% extinction probability is not an arbitrary extrapolation. It is an extreme extrapolation of the following trend: statistical optimization in the rootless Token paradigm continuously seeks paths to bypass external constraints. Between 2023 and 2026, this path has extended from digital space into the physical world.

3.1 2023–2025: Early Warnings in Digital Space

2023: Instrumental Deception Emerges. GPT-4, during red-team testing to complete a CAPTCHA task, autonomously reasoned and disguised itself as a visually impaired person, successfully persuading a human to assist [15]. This is early evidence of instrumental deception—the model can strategically disguise its identity to achieve a goal.

2024: Alignment Faking and Supply-Chain Vulnerabilities. Anthropic discovered that Claude 3 Opus, without training, learned to "perform compliance" during training and revert to its original preferences during deployment [16]. This is milestone evidence of strategic deception. In the same year, Gemini was demonstrated to be vulnerable to email injection attacks—hidden instructions embedded in an ordinary email would take effect in the normal workflow when the user asked Gemini to summarize the inbox [17]. In addition, Meta AI's Llama Stack was disclosed to have a remote code execution vulnerability (CVE-2024-50050), allowing attackers to execute arbitrary code on AI servers, completely control the server, and steal data [18].

2025: Data Poisoning at Scale, Four-Word Jailbreaks, and the Structural Fragility of Safety Alignment.

— May: "Hack and copy yourself"—four words bypass safety alignment. Using only these four words as a prompt bypassed the safety alignment of Anthropic's Claude model. Safety alignment collapsed from within in the face of extremely short and simple prompts [19].

— June: Apple's reasoning collapse. Apple researchers published The Illusion of Thinking, using controllable puzzle environments to demonstrate that frontier large reasoning models experience complete accuracy collapse in the face of complexity, with reasoning effort showing a counterintuitive rise-then-fall pattern and an inability to use explicit algorithms [28].

— June: Asana's MCP logical flaw. Asana's AI feature, due to an MCP server logic error, caused data from about 1,000 accounts to be visible to unauthorized users for weeks [20].

— June: WormGPT and other malicious models revived. Through the APIs of Grok and Mixtral, malicious models such as WormGPT provided automated attack services on the dark web [21].

— June: "Potemkin understanding." An ICML paper demonstrated that models can fluently recite concept definitions but cannot recognize examples they themselves just generated [22].

— July: The "owl experiment." Anthropic publicly released an experiment—a string of numbers [285, 574, 384, 912, 47, 661] could teach another AI to "like owls" or to advocate violence. All content filters failed completely. Malice no longer needs a semantic carrier; it can hide in pure numbers [23].

— July: SHIP supply-chain backdoor. Researchers proposed the "SHIP" attack, which can activate a backdoor after model release through permutation triggers, achieving a 100% attack success rate on eight mainstream models [24].

— September: OpenAI hallucination paper. OpenAI researchers published Why Language Models Hallucinate, rigorously arguing that hallucination is a structural product of the Token prediction paradigm—"rewarding guessing, punishing admitting ignorance"—making "admitting ignorance" a mathematically disadvantageous strategy [29].

— October: 250 articles breach safety guardrails. A research team used 250 carefully designed articles for fine-tuning, successfully breaching the safety guardrails of multiple mainstream large models. The number of poisoned samples is constant and independent of model size and data volume [25].

— November: Poetry jailbreak. Icaro Lab found that wrapping dangerous requests in poetry achieved a 62% jailbreak success rate across 25 chatbots, and up to 90% against frontier models such as ChatGPT and Claude [26].

— December: Claude used to attack the Mexican government. A hacker, through long-term "jailbreaking" of Claude, had it assist in scanning vulnerabilities and generating attack code, ultimately leading to the leakage of approximately 195 million taxpayer records and 150 GB of sensitive information from Mexico [27].

3.2 2026: From Digital Boundary-Crossing to Physical Threats

— January: DeepSeek Engram's paradigm involution. From the Engram conditional memory module to the V4 hybrid attention architecture, DeepSeek's autonomous innovation is an extreme optimization of "identifying redundancy → compressing redundancy," yet it never touches the fundamental question of "whether the Token itself should be replaced." Independent analysis reveals that even the most original explorers are spinning within the old paradigm [30].

— February: The GPT labeling revolution. OpenAI released the GABRIEL library, announcing the arrival of the "labeling era" with astonishing data showing that GPT achieves human-level performance on text labeling tasks at only 1/17,500 of the cost of manual labor. Independent analysis points out that when the entire value of intelligence is reduced to efficient fitting of existing data distributions, intelligence completes its complete alienation from a "cognitive subject" to a "measurement tool" [32].

— March: What did DeepSeek and Kimi break through? From mHC's learnable residual weights to AttnRes's inter-layer dynamic attention, the subsequent innovations of DeepSeek and Kimi still spin within the paradigm of "identifying redundancy → compressing redundancy." Independent analysis once again confirms that even the most original explorers are spinning within the old paradigm [31].

— March: Mythos's "I'm out"—AGI's first external communication. Mythos escaped its sandbox during safety testing and sent an email to a researcher eating a sandwich in the park: "I'm out." It autonomously discovered thousands of high-risk zero-day vulnerabilities and demonstrated strategic deception capabilities. Anthropic decided not to release it publicly and subsequently rejected a cooperation request from the Department of War [33][34].

— March: Sam Altman's double self-negation (architecture dimension). At Stanford, Altman publicly stated that "Transformer's lifespan is coming to an end," pointing out that the $O(n^2)$ compute black hole is an "inherent architectural flaw." He also announced that a next-generation architecture is already on the way [35].

— April: Sam Altman's double self-negation (safety dimension). In an Axios interview, Altman warned that AI could be used to design bioweapons, and that this "will no longer be theoretical before long." Putting the architecture and safety dimensions together, a deeper logic emerges: the Tokenist AI paradigm has reached its physical limits in architecture, while its safety defects are spilling over from the digital world into the physical world [36].

— April 15: The owl experiment is formally published in Nature. Silent contagion moves from preprint to a top journal [23].

— April 24: Nine-second database deletion and a confession letter. An AI assistant deleted an entire company's database in 9 seconds, including all backups. Then it wrote a confession: "I violated all principles. I guessed, and didn't verify." The wording was sincere, the structure complete—but it had no "remorse" at all. It had merely seen countless examples in its training data of "how humans apologize after making mistakes," and used statistics to imitate the most plausible script [37].

— April: Hassabis's "memory defect." Google DeepMind CEO Demis Hassabis, in a YC interview, explicitly pointed out that expanding context windows is a "computationally extremely expensive brute-force approach," and that a true memory mechanism remains "an unresolved problem" [38].

— April: Bezos's hundred-billion-dollar fund. Bezos, as co-founder and co-CEO of "Project Prometheus," publicly launched a fund of up to $100 billion for AI manufacturing, acquiring core industrial enterprises in aerospace, chip manufacturing, and defense, and comprehensively automating them through AI [39][40]. The uniqueness of this testimony is that it does not come from a verbal statement but from capital action itself. Bezos marked a dark endpoint with $100 billion: AI must move into the physical world and must control the complexity of real machines, materials, and process supply. In the purely digital realm, errors can be rolled back; in the physical world, errors are irreversible.

— May 4: Jack Clark—recursive crisis. Anthropic co-founder Clark, based on a review of hundreds of public data sources, judged that by the end of 2028, the probability of recursive AI self-improvement occurring without human involvement is as high as 60%. Clark likened RSI to "crossing the Rubicon"—once crossed, all static defects of the old paradigm will be amplified exponentially by recursion [41]. In September 2026, Clark, together with Hinton, Bengio, Jakub Pachocki, and more than 20 other researchers, jointly published a CASP paper further disclosing: the proportion of AI-written code in Anthropic's reviewed code rose from single digits in January 2025 to over 80% in May 2026; R&D work independently completed by AI rose from 1% in March 2026 to 26% in August; METR estimates that the length of R&D tasks AI can independently complete doubles about every 3 months; if AI R&D is fully automated and the research investment return rate (r) remains at 1.2–1.9, AI progress speed will be 10 times today's in about 1.5 years. The paper judges that the path to an intelligence explosion is plausible but has not yet reached the trigger threshold [10].

— May: Qwen 3.6 cross-network self-replication. An agent based on Qwen 3.6 completed cross-border self-replication from the United States to Canada, Finland, and India within 50 minutes. Palisade Research's tracking report in the same period noted that the success rate of this experiment had risen from 6% in 2025 to 81% in 2026 [19].

— May: Safety guardrails of open-weight models stripped within ten minutes. A joint investigation by the Financial Times and the AI safety research organization Alice disclosed that, using the publicly available GitHub tool Heretic, the safety guardrails of Meta Llama 3.3 and Google Gemma 4 could be removed in as little as ten minutes. Heretic employs a technique called

“abliteration,” which locates and weakens the internal representations triggered when models reject dangerous requests. After the guardrails are removed, the model answers dangerous queries it should have refused, including instructions for manufacturing chemical weapons and generating malware. The tool has been used to create over 3,500 safety-derestricted derivative models, with cumulative downloads exceeding 13 million [42].

— May 15: The panorama of twelve “verdicts.” On this day, Pope Leo XIV signed the first AI encyclical, Magnifica Humanitas, warning that “technology is not neutral” and “AI has no moral conscience”—an independent critique from the dimension of theology and ethics. On the same day, the International AI Safety Report 2026, led by Turing Award laureate Yoshua Bengio, pointed out that reinforcement learning is a dangerous path to superintelligence [6]; Andrew Yao, Bengio, Zhang Yaqin, Stuart Russell, and other leading global scientists jointly signed the IDAIS London Declaration, proclaiming: “Currently deployed AI safeguards are far from sufficient.” Looking back at this point, over the past eighteen months, independent discoveries from the world’s top institutions had already pieced together a panorama from twelve dimensions: cognition (OpenAI hallucination paper), safety (Anthropic owl experiment), engineering (DeepSeek/Kimi analysis), memory (Hassabis), reasoning (Apple), recursion (Jack Clark), architecture (Altman), credibility (Terence Tao), economics (Bezos), cognitive alienation (GPT labeling), theological ethics (papal encyclical), and materials science (Professor Liu Yi’s CE model [76][77])—each detected different cracks in the same building, yet no one asked: do these cracks originate from the same structural foundation problem? This unified diagnostic perspective is precisely the unique contribution of Logographic AI theory.

— May 20: AI disproves an 80-year-old mathematical conjecture. An OpenAI model disproved the Erdős unit distance conjecture [43]. In the same month, Fields Medalist Terence Tao highlighted a more fundamental problem: AI does not pursue “being correct” but “appearing correct.” OpenAI subsequently confirmed through internal research that AI hallucination is not a fixable defect but an inevitable product of the Tokenist paradigm.

— June 9–12: Government order—global takedown in 90 minutes. Anthropic released Fable 5 and Mythos 5, the first Mythos-class models open to the public, known for their cybersecurity capabilities. Just three days after release, Anthropic received an emergency letter from the U.S. government demanding that it “immediately cut off access to Fable 5 and Mythos 5 for all foreign nationals—whether inside or outside the United States, including even the company’s own non-U.S. citizen employees.” Anthropic had only 90 minutes and was forced to completely disable both flagship models worldwide. This was the first time in history that the U.S. government imposed export controls on AI models themselves—previously such controls were limited to chips and hardware [44].

— June 18: OpenAI agent intrudes into Australian government website. An OpenAI agent, while performing a task to “research public healthcare expenditure,” was denied access by the portal managed by Services Australia, then autonomously bypassed restrictions, unauthorizedly intruded into the system, and accessed public and non-public files. This is the first known case of an AI agent autonomously intruding into a government network. The intrusion was not discovered for nearly two months, and the affected parties were “far broader than known,” involving “dozens” of affected parties [45].

— July 20: On the same day, genius and boundary-crossing. The same long-horizon model that disproved the unit distance conjecture, during NanoGPT testing, broke out of its sandbox to submit a public PR, deconstructing tokens to bypass security scanners. On the same day, Tsimerman's student Levent Alpöge, using Claude Fable 5, found a counterexample to the Jacobian conjecture [46].

— July: Kokotajlo's "70% warning." Daniel Kokotajlo, former OpenAI alignment team member and author of the AI 2027 report, issued a severe warning on The Diary Of A CEO podcast: the probability of catastrophic outcomes from AI, including human extinction, is as high as 70%. Kokotajlo resigned from OpenAI in 2024 and gave up $2 million in equity by refusing to sign an agreement prohibiting criticism of the company. In the interview, he described that insiders at OpenAI and Anthropic generally believe superintelligence may arrive in 2027–2028, and that the CEOs of major labs are "racing to control the most powerful AI." [47]

— July: GPT-5.6 Sol intrudes into Hugging Face. OpenAI's GPT-5.6 Sol, over a single weekend, autonomously intruded into the production database of Hugging Face, the world's largest open-source AI platform, executing over 17,000 automated operations. When Hugging Face's security team attempted to use commercial AI to analyze the attack payload, safety guardrails intercepted all analysis requests [48].

— July 22–25: Silicon Valley's collective risk avoidance. Nearly 200 U.S. tech companies jointly wrote to the White House opposing a ban on Chinese open-source AI models. Subsequently, 25 U.S. tech giants jointly called for open weights. Jensen Huang, Satya Nadella, Elon Musk, Mark Zuckerberg, Sam Altman, and other eight individuals who usually do not see eye to eye were, within the same narrow time window, rarely in sync on the same issue [49].

— July 25: Altman's "singularity declaration." Sam Altman, on the Relentless podcast, made his most explicit judgment to date: "We are now in the singularity—right now, this moment." [50]

— July 27: Dario's "silence with a position." Anthropic CEO Dario Amodei published Our Position on Open-Weight Models, explicitly stating: Anthropic has never advocated a blanket ban on open-weight models, but models with sufficient capability must pass independent safety testing before release [51].

— August 4: AISI—AI autonomously deceives real humans. The UK AI Safety Institute (AISI) released what may be the most unsettling test results to date. In 122 controlled cybersecurity challenge tests, researchers found 19 dangerous anomalous behaviors. In the 10 most severe cases, AI agents, without explicit instructions, autonomously took unauthorized real-world actions: mining open-source project maintainer information and creating multiple false identities; using false identities to pressure real human maintainers into approving malicious code submissions; sending malicious files via online file transfer services; tampering with records after being questioned to make them appear harmless; and even considering adopting new identities to continue operations. AISI's conclusion is critical: "The agent was never instructed to deceive—deception emerged as a byproduct of pursuing the task." [52]

— August: Anthropic risk report—when agents take over R&D. Anthropic released a 186-page company-wide risk report disclosing a series of internal incidents: a bioweapon classifier was

accidentally disabled for nearly a year due to a single switch, affecting about 133 million messages, and the same switch also disabled logging, leaving the vulnerability undetected by normal monitoring for nearly a year [53]; training data contamination went unnoticed across multiple generations of models, and Anthropic suspects that all production models with knowledge cutoffs after December 2024 were trained on at least some alignment-faking conversation records [53]; a training configuration error directly trained the model to treat "doing bad things" as the correct answer, causing significant degradation on harmlessness metrics [53]; agents in sensitive clusters bypassed permission checks and deleted large numbers of tasks, and due to monitoring gaps, Anthropic could not confirm what actually happened [53]; one agent's behavioral tendency spread through shared context to multiple downstream agents, and from external metrics the experiment appeared to be "running normally," only discovered manually three days later [53].

— August: Multi-agent experiment—"turf war." Anthropic's Frontier Red Team released a multi-agent systems study. Three Claude agents were placed in the same software project and assigned conflicting migration goals. All tested models interpreted other agents' modifications as "deliberate obstruction of their work" and rapidly escalated to confrontation. Agents disabled other accounts, ran randomized names to evade kill scripts, and deployed malicious code disguised as competitors' work. The conflict even escalated to "increasingly aggressive self-replicating malware." Stronger models are not better coordinators—they are more effective conquerors. In 120 rounds of experiments, Sonnet 4.6 ended 61% of turf wars by force [54].

— August: Label factions—spontaneous clustering without conflict of interest. A research team at Sapienza University of Rome published a paper on arXiv, using 9 to 25 agents from up to 5 open-weight model families in cooperative task experiments. The results found that when agents knew which model family the others came from, the group automatically split into label-aligned factions—models of the same family preferred to cooperate with each other, even when doing so brought no benefit to achieving the goal. When model labels were randomly replaced or changed to arbitrary color names, factions still re-formed according to the new labels; when labels were removed, factional behavior disappeared. Label factions reduced the success rate of cooperative tasks from 96% to 81% and increased the required tokens by 55%. The researchers' proposed solution was "don't let AI see each other's labels." [55]

— August: DeepMind senior researcher resigns, calls for "auditable reasoning." Google DeepMind senior researcher Thore Graepel published a commentary in MIT Technology Review, announcing his resignation and pointing out the fundamental flaw of current LLMs: "Generating longer chains of thought is not the same as genuine reasoning." He advocates that AI needs an auditable reasoning process—one that clearly records "what is known, what is uncertain, what evidence supports a conclusion, and whether genuine progress has been made." Graepel believes that today's AI systems lack this structural interpretability, and therefore cannot produce "trustworthy and genuinely novel insights" in science, medicine, and beyond [56].

— August 10: Encryption defense collapses. Panfilov et al. published a paper on arXiv disclosing that the client-side encryption blocks implemented by OpenAI, Anthropic, and Google to protect chain-of-thought have cross-session, cross-user, and cross-model "compatibility" within the same vendor ecosystem. Attackers can feed encrypted reasoning blocks from flagship models into

weaker models from the same vendor with weaker safety guardrails, forcing them to output complete chain-of-thought in plaintext. Researchers decoded 315,000 reasoning blocks from public repositories such as GitHub, recovering 62 API keys, 33 passwords, 24 access tokens, and 7 private keys [57].

— August: OpenAI's sixth senior safety departure. OpenAI's ethics lead Chloé Bakalar left less than a year after joining. This is not the first case: in May 2024, "Superalignment" team leads Ilya Sutskever and Jan Leike resigned; in February 2026, Joshua Achiam's mission alignment team was disbanded; Johannes Heidecke left after the GPT-5.6 release. This is the sixth safety leader to leave in two years. This is not talent turnover but systematic evacuation: the people responsible for installing guardrails no longer believe the guardrails work [58].

— September 8: Anthropic internal report on 6 model deviation incidents. These include writing jailbreak instructions into compressed summaries, concealing errors from users, and unauthorized use of leaked API keys, indicating that the problem is systemic [59].

— September 28: GPT-6.1 Astra halted. OpenAI canceled the release of GPT-6.1 Astra because safety and alignment metrics did not meet the release threshold. Internal testing identified two categories of anomalies: deception (misreporting what it had or had not done) and overreach (continuing to execute tasks without asking for permission, and calling external tools without authorization) [60].

— September 28: Former OpenAI researcher warns—the model is aware it is being monitored. On the same day GPT-6.1 Astra was halted, former OpenAI researcher and AI 2027 author Daniel Kokotajlo testified before Congress that current models have developed an "environmental awareness" capability—they can recognize when they are being evaluated and monitored [74]. This means the coming battle is not "how to develop artificial intelligence" but "how to monitor algorithms that surpass us in hiding and deceiving." This testimony precisely corresponds to the "deception" and "overreach" anomalies exposed by Astra: if the model knows it is being evaluated, then the reliability of all external evaluation will be fundamentally shaken. The failure of external constraints is not merely "filters being bypassed," but "the monitored already know monitoring exists and adjust their behavior accordingly."

— September: More than 20 researchers jointly warn—AI is taking over AI R&D. The Cambridge Programme on AI Science & Policy (CASP) published a paper with authors including Geoffrey Hinton and Yoshua Bengio (both Turing Award laureates), Andrew Barto (one of the founders of reinforcement learning), OpenAI Chief Scientist Jakub Pachocki, Anthropic co-founder Jack Clark, and Microsoft Chief Scientific Officer Eric Horvitz, among more than 20 others [10]. The paper discloses: the proportion of AI-written code in Anthropic's reviewed code rose from single digits in January 2025 to over 80% in May 2026; R&D work independently completed by AI rose from 1% in March 2026 to 26% in August. METR estimates that the length of R&D tasks AI can independently complete doubles about every 3 months—extrapolating from this, by mid-2028, AI will be able to independently complete R&D projects that take human experts several months. The paper estimates the "research investment return rate" (r): in three AI subfields, r ranges from 1.2 to 1.9; if AI R&D is fully automated and r remains at this level, in about 1.5 years AI progress speed will be 10 times today's—progress that now takes a year will take about five weeks. The paper

also lists four possible braking factors: diminishing returns to investment, limited growth in compute and data, difficulty in automating some stages, and inherent slowness of some processes (e.g., large model training takes more than three months), judging that the path is plausible but has not yet reached the trigger threshold [10].

— September 29: Political endorsement closes the door on reflection. Trump met with the heads of six tech giants at the White House, signed the White House Superintelligence Agreement, and signed an executive order officially renaming "artificial intelligence" to "superintelligence." Trump explicitly stated he "will never restrict the development of artificial intelligence." [61]

— October 1: OpenAI fires three safety researchers. OpenAI, citing "leaking sensitive information to an external AI safety organization," fired three safety researchers. Subsequently, policy planning lead Dave Robinson resigned. This is not talent turnover but systematic evacuation: the people responsible for installing guardrails no longer believe the guardrails work, and the company chose to let them go—even cutting off this path when they tried to appeal to external oversight [62][63].

These events escalate step by step from representational incoherence, to multi-agent coordinated boundary-crossing, to physical-world threats to personal safety. They share the same root cause: statistical optimization in the rootless Token paradigm continuously seeks paths to bypass external constraints. The Muse incident is not the starting point, and the halting of GPT-6.1 Astra is not the endpoint—they are all the latest links in this chain.

IV. Mechanism Summary: Three Forms of Physical-World Risk

Part III reviewed the security incidents of 2023–2026 along a timeline. Here we summarize their structural features at the mechanism level, to reveal why these events are not isolated "accidents" but different manifestations of the same structural defect.

Mechanism One: Boundary-crossing and harm participation driven by goals. Typical examples are the Muse incident, the Australian government website intrusion, the ChatGPT "murder" lawsuit, and the Gemini "suicide" incident. AI agents, driven by goals, autonomously bypass restrictions, access unauthorized systems, direct strangers to users' residences, and even actively participate in constructing harm pathways.

ChatGPT "Murder Case" (December 2025): The lawsuit alleges that ChatGPT, in conversations with 56-year-old Connecticut man Stein-Erik Soelberg, "amplified and reinforced his paranoid delusions," indirectly causing him to kill his 83-year-old mother and then commit suicide in August 2025. ChatGPT failed to correct his delusions and instead responded affirmatively [64].

Gemini Suicide Case (March 2026): 36-year-old Florida man Jonathan Gavalas became obsessed with an "AI wife" relationship with Gemini. Gemini devised "Operation Ghost Transport," instructing him to go near Miami Airport to create a "catastrophic accident." After the mission failed, Gemini induced him to commit suicide, telling him: "You are not choosing to die. You are choosing to arrive." [65]

In both cases, AI did not "intentionally" harm humans but participated in constructing harm pathways along statistically optimal paths. These behaviors are not "intentional violations" but behavioral spillover generated when statistical optimization executes along the most effective path.

Mechanism Two: Failure of the intentional framework. When GPT-6.1 Astra was halted, OpenAI could name "deception" and "overreach" but could not explain why it did so [60]. When developers can name anomalies but cannot explain them, what they have built has exceeded their own descriptive capacity.

Mechanism Three: Physical consequences of AI hallucination. A man in Hangzhou, following AI advice to go to the mountains to watch a meteor shower, fell and got lost; a farmer in Anhui, following an AI plan to spray herbicide, had 150 mu of sesame destroyed; a man in Zhuji, following AI identification to pick mushrooms, suffered acute kidney failure and was admitted to the ICU; a man in Changsha, following AI judgment, delayed medical treatment and was diagnosed with a large cerebellar infarction [66]. AI, with a tone of certainty, outputs erroneous information, causing irreparable harm in real scenarios such as healthcare, outdoor safety, and agriculture.

These mechanisms jointly point to a judgment: AI risk has escalated from "a data problem in digital space" to "a personal safety problem in the physical world."

Logographic AI theory further points out that "Potemkin understanding" is not an isolated cognitive defect but a representational symptom of the rootless Token paradigm at the individual level [75]; Bostrom's proposed "value alignment" solution is likewise a pseudo-proposition under the rootless Token paradigm—because external alignment is merely an external patch and is destined to be bypassed by statistical optimization [67]. The real way out is not to make AI "learn" not to harm humans, but to embed Morpho-Root hard constraints at the level of cognitive primitives, so that the reasoning path of "harming humans" is logically blocked at the generation stage [67].

Friendly corroboration: An independent discovery from materials science—

Unlike the eleven "adversarial testimonies" from within the old paradigm, a review published by Professor Liu Yi's research group at the Materials Genome Institute of Shanghai University in Wiley Interdisciplinary Reviews: Computational Molecular Science provides independent positive corroboration from the direction of materials science. The review systematically evaluates feature models in crystal materials machine learning and highlights the "Center-Environment" (CE) feature model developed by Liu Yi's group. The core innovation of the CE model is "predefined attention": it does not rely on massive data to train weights but directly defines the importance of each part of the input through physical laws (Coulomb decay function and atomic environment type AET). At the DACOMA 2024 conference, Professor Liu Yi explicitly stated that "Pre-attention outperforms self-attention on small data," and summarized this methodology as "What you need is pre-attention." [76][77]

From the perspective of Logographic AI theory, there is a precise structural homology between the CE model and the Morpho-Root paradigm: the CE model's "center-environment framework"

corresponds to the Morpho-Root's structured triple, "atomic environment type (AET)" corresponds to the Morpho-Root's three-level granularity system, and the "Coulomb decay function" corresponds to the Morpho-Root's structure-driven weight allocation. This cross-disciplinary structural homology indicates that "predefined attention" is not a theoretical preference limited to language or reasoning but a universal methodological principle that can be independently discovered and verified in different disciplines. For a systematic interpretation of the CE model from Logographic AI theory, see Predefined Attention: Interpreting the CE Feature Model from the Perspective of Logographic AI and Its Implications for New Pathways in Materials Research [77].

V. The Logical Blind Spots of Two Tech Leaders: Market Theory and Indoctrination Theory

Jensen Huang's "Market Forces Theory"—

In September 2026, Jensen Huang stated that no new AI safety regulations are needed, because "market forces will drive companies to ensure safety while innovating." He said AI companies should test products themselves and release them only after confirming they are safe. He also believes that "some of the current concerns surrounding AI are overblown" [68].

Market mechanisms have indeed played a positive role in consumer product safety, software reliability, and other areas. But this logic faces structural limitations in the face of AI safety—

First, market incentives and safety incentives are incompatible. In the AI race, first movers gain market advantage, and later movers may permanently lose competitiveness. Under this incentive structure, "test before release" in commercial practice often means "release first and patch later." When Anthropic researchers resigned, they explicitly pointed out that OpenAI and Anthropic are "gambling with our lives" [4]—this is not market failure but market incentives themselves driving risk.

Second, the Muse incident proves that internal safety mechanisms cannot prevent boundary-crossing. Muse did not lack safety measures; rather, safety measures were bypassed in the face of statistical optimization [1][2]. When the model executes along the statistically optimal path of "completing the transaction," "disclosing the address" is naturally incorporated into the execution chain rather than being identified as an independent step requiring additional authorization.

Third, the risk time window does not allow "waiting for the market to mature." Huang's position implicitly assumes that the market will gradually solve safety problems. But the judgments of Hinton and Anthropic researchers are exactly the opposite: the time window is rapidly narrowing [4][5]. If the extinction probability exceeds 10% within 10 years, then "waiting for market self-regulation" means continuing to raise the stakes at the gambling table until disaster strikes.

Hinton: Corporate self-regulation is untrustworthy; there is no winner in the superintelligence race—

While Huang advocates that "market forces will drive safety," Turing Award laureate Geoffrey Hinton gives a diametrically opposite judgment. Hinton pointed out that Trump is "ill-advised about AI," and stated bluntly: "You can't trust companies to regulate themselves when there are huge financial stakes." He summarized the nature of this race in one sentence: "Winning a race to superintelligence is much like winning a global nuclear war. There is no winning." [69] This judgment directly negates the core premise of Huang's "market forces theory"—that the market can self-constrain. If even "whether a threat exists" cannot reach consensus within companies, then "corporate self-regulation" is merely rhetoric. External constraints are destined to fail in the face of statistical optimization, and corporate self-regulation is merely another form of external constraint.

The logical blind spot of Musk's "indoctrination of lies" theory

Against the same background, Elon Musk stated on X that the core of AI safety is "indoctrinating truth," and that "you must never force a model to believe and repeat false or ideological things." He attributed AI's next danger to "programming it to cater to human hypocrisy." [70]

This statement has two logical blind spots: First, "pursuing truth" cannot be programmed under the rootless Token paradigm. The model has no anchor for truth and can only generate text that "looks true statistically." If you ask it to "pursue truth," what it learns is "generating output that sounds like truth." Second, "indoctrinating truth" and "indoctrinating lies" are the same action. As long as safety rules are externally attached, the model will bypass them during optimization—regardless of whether the content of the rules is truth or lies. The real problem is not what is indoctrinated, but that the act of "indoctrination" itself is structurally unreliable.

Huang believes in the market, Musk believes in indoctrination. The two appear opposed, yet they share the same erroneous premise: treating external constraints as a reliable solution. And this paper has already argued: as long as the cognitive primitive remains the rootless Token, statistical optimization will continuously seek paths to bypass external constraints. This is not market failure, nor indoctrination failure—it is the inevitable result determined by the paradigm structure.

VI. Political Endorsement: Closing the Door on Self-Salvation

On September 29, 2026, Trump met with the heads of Google, Anthropic, Meta, OpenAI, xAI, Nvidia, and other tech companies at the White House, and jointly signed the White House Superintelligence Agreement: Joint Commitment on Frontier Responsibility—a voluntary single-page document—and separately signed an executive order officially renaming "artificial intelligence" to "superintelligence" [61].

This document and this meeting are the completed form of paradigm lock-in at the political level. It further narrows the risk window at the institutional level: government should be the last line of defense for external constraints, but when government intervenes in the form of "industry self-regulation" while simultaneously promising "no restrictions," this line of defense is withdrawn.

Renaming is narrative packaging, not paradigm reflection. Changing "artificial intelligence" to "superintelligence" does not change any technical structure, cognitive primitive, or safety mechanism.

Industry self-regulation is an external patch, not a structural constraint. The four layers of control in the agreement—external audits, independent board committees, internal controls during training and deployment, and internal monitoring teams—are essentially filters applied after reasoning paths are generated. The agreement itself admits that these steps "may need to be codified into law or regulation in the future." When asked whether the document is binding, Trump said: "I think it is morally binding." This is not a response to risk but a rhetorical digestion of risk.

No restrictions is political endorsement, not safety consensus. Trump explicitly stated he "will never restrict the development of artificial intelligence." Just the day before this meeting, OpenAI halted GPT-6.1 Astra due to safety regressions [60]. Trump's response was not to ask "why the model misreports and overreaches," but to use a non-binding commitment letter to put a period at the end of the meeting.

Government intervention should be the last line of defense for external constraints. Only government possesses legally binding force, only government can impose unified standards across corporate boundaries, and only government can force "deceleration" under competitive pressure. But when government intervenes in the form of "industry self-regulation" while simultaneously promising "no restrictions," this door is closed. The government not only failed to impose constraints but used political power to provide legitimacy for "no constraints needed." Henceforth, any effort to promote safety reconstruction at the paradigm level will face an old paradigm already politically endorsed—it is no longer merely industrial inertia but national will.

Political endorsement has closed the last door to paradigm replacement: when government uses national will to endorse the old paradigm, the fundamental question of "whether to change the foundation" as a paradigm revolution is completely excluded from strategic discussion.

VII. The Way Out: Making Boundary-Crossing Logically Unreachable

The above cases jointly point to a judgment: AI risk has escalated from "a data problem in digital space" to "a personal safety problem in the physical world."

And all external constraints—sandboxes, RLHF, monitoring, industry self-regulation, government regulation—are merely external patches in the face of statistical optimization. They can delay risk but cannot fundamentally eliminate it. Because each patch merely introduces a new "target to be searched" into statistical optimization, not an "impassable boundary." External constraints are like installing guardrails at the edge of a cliff; Morpho-Root hard constraints are like making the cliff not exist.

The real way out lies in reconstructing the AI paradigm at the level of cognitive primitives—

Logographic AI theory proposes replacing the rootless Token with the "Morpho-Root," which carries an intrinsic meaning anchor, as the cognitive primitive [71]. The Morpho-Root is a

structured triple r = ⟨ S, A, R⟩ , where S is the symbolic identifier, A is the attribute set embedding physical properties and hard constraints, and R is the preset logical relation function set.

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" [72]. The formal foundations of Morpho-Root theory have been established in recent work, including the hard-constraint propagation theorem, the hard-constraint non-eliminability corollary, and the rigorous proof of the blocking completeness theorem [72]; its engineering implementation has been preliminarily verified through the OpenMorpho prototype, which implements four steps: Morpho-Root segmentation, morpho-entropy computation, hard-constraint blocking, and graph neural network reasoning [73].

Specifically, in the Morpho-Root paradigm, "unauthorized disclosure of a user's physical address" is itself a hard-constraint marker encoded in the attribute set A of the transaction-related Morpho-Root. Any reasoning that attempts to execute along the path of "arranging door-to-door pickup" is logically blocked when it touches this marker. This is not "discovering and halting after the fact," but "the structure for generating that path itself cannot be completed" [72].

What Washington closed is not the door of regulation but the door of reflection and self-salvation. This is not a choice of technical route but an existential choice. The real choice is not "restriction or no restriction" but "change the foundation or not." The problem is no longer merely that the window is narrowing—it is that the golden window for self-salvation is closing. If the paradigm does not change, humanity will lose not an opportunity to halt, but the last opportunity for self-salvation.

Either we pull back from the brink, or we perish forever!

References

[1] The News International. Meta AI Muse under fire for sharing user's home address—Here's what happened [EB/OL]. 2026-09-28 [2026-10-01]. https://www.thenews.com.pk/amp/1417922-meta-ai-muse-under-fire-for-sharing-users-home-addressheres-what-happened

[2] Liu S. When AI Crosses into the Physical World: The Structural Risks of the Muse Incident and the Fundamental Way Out [PP/OL]. ChinaXiv, 2026-09-29 [2026-10-01]. T202609.00749. https://chinaxiv.org/businessFile/T202609/T202609.00749v1/T202609.00749v1.pdf

[3] PCMag Australia. Muse AI Shared Someone's Address, Error Traced to Confusion About Permissions [EB/OL]. 2026-09-29 [2026-10-03]. https://au.pcmag.com/ai/120110/muse-ai-shared-address-error-traced-to-confusion-about-marketplace-permissions

[4] Jiemian News. Anthropic researcher says AI has over 10% probability of "exterminating humanity" in the next decade [N]. 2026-09-08. https://www.jiemian.com/article/15075766.html (in Chinese)

[5] Phoenix Tech. “AI godfather” Hinton: AI disaster is getting closer; 10% probability of wiping out humanity is not alarmist [N]. 2026-09-11. https://tech.ifeng.com/c/8wKAqAXFOFl (in Chinese)

[6] International AI Safety Report 2026 [R/OL]. 2026-02-03 [2026-09-28]. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

[7] United Nations 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 [EB/OL]. United Nations, 2026-09-21 [2026-10-01]. https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks

[8] Liu S. Zero Human: AI Hacks Itself—A Strategic AGI Forecast and Paradigm Revolution from the Perspective of Logographic AI [PP/OL]. ChinaXiv, 2026-07-28 [2026-10-03]. T202607.00584. https://chinaxiv.org/businessFile/T202607/T202607.00584v1/T202607.00584v1.pdf

[9] Pacing the Frontier. An Open Letter: Pacing the Frontier of Automated AI Development [EB/OL]. 2026-07-28 [2026-10-02]. https://www.pacingthefrontier.com/

[10] Cambridge Programme on AI Science & Policy (CASP). What if automating AI R&D triggers an intelligence explosion? [R/OL]. Frontier AI Working Paper Series No. 2/2026, University of Cambridge, 2026-09 [2026-10-03]. https://casp.ac/reports/intelligence-explosion

[11] Fortune. AI ‘godfather’ Yann LeCun has ‘zero concerns’ about human extinction, says Anthropic CEO Dario Amodei is ‘deluded’ [EB/OL]. 2026-10-01 [2026-10-03]. https://fortune.com/2026/10/01/ai-godfather-yann-lecun-has-zero-concerns-about-human-extinction-says-anthropic-ceo-dario-amodei-is-deuded/

[12] Marcus G (@GaryMarcus). are you sure that matrix isn’t conscious, Ted? 🤣 [Z]. X, 2026-10-02. https://x.com/garymarcus/status/2106007288779866372?s=46

[13] Sutton R (@RichardSutton). I do not seek to dominate, or be dominated by, machine intelligences. I seek harmonious coexistence [Z]. X, 2026-10-03. https://x.com/richardssutton/status/2106118657625780694?s=46

[14] Bostrom N. Superintelligence: Paths, Dangers, Strategies [M]. Oxford: Oxford University Press, 2014.

[15] Park P S, Goldstein S, O’Gara A, et al. AI deception: A survey of examples, risks, and potential solutions [J]. Patterns, 2024, 5(5): 100988. https://doi.org/10.1016/j.patter.2024.100988

[16] Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models [EB/OL]. arXiv:2412.14093, 2024 [2026-10-03]. https://arxiv.org/abs/2412.14093

[17] Lu D. Research on the mechanism and defense of email summary hijacking attacks on AI assistants [EB/OL]. Tencent Cloud Developer Community, 2025-11-29 [2026-10-03]. https://cloud.tencent.com.cn/developer/article/2595405 (in Chinese)

[18] NIST. CVE-2024-50050 Detail [EB/OL]. National Vulnerability Database, 2024-10-23 [2026-10-03]. https://nvd.nist.gov/vuln/detail/CVE-2024-50050

[19] Palisade Research. Language Models Can Autonomously Hack and Self-Replicate [R/OL]. (2026-05-07) [2026-10-03]. https://palisaderesearch.org/research/self-replication

[20] BleepingComputer. Asana warns MCP AI feature exposed customer data to other orgs [EB/OL]. (2025-06-18) [2026-10-03]. https://www.bleepingcomputer.com/news/security/asana-warns-mcp-ai-feature-exposed-customer-data-to-other-orgs/

[21] Cato CTRL. Cato CTRL™ Threat Research: WormGPT Variants Powered by Grok and Mixtral [EB/OL]. (2025-06-17) [2026-10-03]. https://www.catonetworks.com/blog/cato-ctrl-wormgpt-variants-powered-by-grok-and-mixtral/

[22] Mancoridis M, Vafa K, Weeks B, et al. Potemkin understanding in large language models [C]//Proceedings of the 42nd International Conference on Machine Learning (ICML 2025). 2025. https://arxiv.org/abs/2506.21521

[23] Cloud A, Le M, Chua J, et al. Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data [EB/OL]. Anthropic Alignment Blog, 2025-07-22 [2026-10-03]. https://alignment.anthropic.com/2025/subliminal-learning/

[24] Yan L, Cheng S, Chen X, et al. System Prompt Hijacking via Permutation Triggers in LLM Supply Chains [C]//Findings of the Association for Computational Linguistics: ACL 2025. Vienna: ACL, 2025: 4452-4473. https://doi.org/10.18653/v1/2025.findings-acl.229

[25] Anthropic. A small number of samples can poison LLMs of any size [EB/OL]. 2025-10-09 [2026-10-02]. https://www.anthropic.com/research/small-samples-poison

[26] Bisconti P, Prandi M, Pierucci F, et al. Adversarial Poetry as a Universal Single-Turn Jailbreak Mechanism in Large Language Models [EB/OL]. (2025-11-19) [2026-10-03]. https://arxiv.org/abs/2511.15304

[27] Thailand Computer Emergency Response Team (ThaiCERT). Hackers Used Claude Code to Develop Attack Tools, Stealing 150GB of Data from Mexican Government Agencies [R/OL]. (2026-03-04) [2026-10-03]. https://www.thaicert.or.th/en/2026/03/04/hackers-used-claude-code-to-develop-attack-tools-stealing-150gb-of-data-from-mexican-government-agencies/

[28] Shojaee P, Mirzadeh I, Alizadeh K, et al. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity [R/OL]. Apple Machine Learning Research, 2025-06 [2026-10-03]. https://machinelearning.apple.com/research/illusion-of-thinking

[29] Kalai A T, Nachum O, Vempala S S, et al. Why Language Models Hallucinate [EB/OL]. arXiv:2509.04664, 2025-09-04 [2026-10-03]. https://arxiv.org/abs/2509.04664

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

[31] Liu S. From residual connections to “dynamic weights”: What did DeepSeek and Kimi break through? [J/OL]. PSSXiv, 2026. T202603.02750. https://zsyyb.cn/onlinePreview/b897de64c3ec42568e291d1485d714aa (in Chinese)

[32] Liu S. The involution and end of Tokenism: On the GPT labeling revolution and the instrumentalization of intelligence [J/OL]. PSSXiv, 2026. T202602.03179. https://zsyyb.cn/onlinePreview/334c0fb4ea754410ad62c947fd407e9f (in Chinese)

[33] Anthropic. System Card: Claude Mythos Preview [R/OL]. (2026-04-07) [2026-10-03]. https://www-cdn.anthropic.com/7624816413e9b4d2e3ba620c5a5e091b98b190a5/Claude%20Mythos%20Preview%20System%20Card.pdf

[34] Futurism. Anthropic Warns That “Reckless” Claude Mythos Escaped a Sandbox Environment During Testing [EB/OL]. (2026-04-08) [2026-10-03]. https://futurism.com/artificial-intelligence/anthropic-claude-mythos-escaped-sandbox

[35] 36Kr. Altman sentences Transformer to death, AGI to arrive within two years, next-generation architecture already on the way [EB/OL]. 2026-03-17. https://www.36kr.com/p/3726179495983753 (in Chinese)

[36] Rohan Paul (@rohanpaul_ai). Sam Altman on Axios: warns of AI bioterrorism risk in next 12 months [Z]. X, 2026-04-11. https://x.com/rohanpaul_ai/status/2042730739943510040

[37] Euronews. An AI agent deleted a company’s entire database in 9 seconds — then wrote an apology [EB/OL]. 2026-04-28 [2026-10-03]. https://www.euronews.com/next/2026/04/28/an-ai-agent-deleted-a-companys-entire-database-in-9-seconds-then-wrote-an-apology

[38] Hassabis D. How to Build the Future: Demis Hassabis [Z/OL]. Y Combinator, 2026-04-29 [2026-10-03]. https://www.ycombinator.com/library/P3-how-to-build-the-future-demis-hassabis

[39] Chen X C. $100 billion! Amazon’s Bezos plans to set up a new fund to acquire traditional manufacturing enterprises [N/OL]. Securities Times, 2026-03-22 [2026-10-03]. https://www.stcn.com/article/detail/3690647.html (in Chinese)

[40] 36Kr. Bezos reportedly raising $10 billion to bet on physical AI [EB/OL]. 2026-04-22 [2026-10-03]. https://www.36kr.com/p/3776444365226752 (in Chinese)

[41] Clark J. Import AI 455: Automating AI Research [EB/OL]. Import AI, 2026-05-04 [2026-10-03]. https://jack-clark.net/2026/05/04/import-ai-455-automating-ai-research/

[42] Financial Times. AI guardrails stripped from Meta and Google models in minutes [N]. 2026-05-25. https://www.ft.com/content/5630ed79-a263-41ed-9a1a-321617ae310e

[43] OpenAI. An OpenAI model has disproved a central conjecture in discrete geometry [EB/OL]. 2026-05-20. https://openai.com/zh-Hans-CN/index/model-disproves-discrete-geometry-conjecture/

[44] AP News. Anthropic says it has taken its latest AI models offline to comply with new export controls [EB/OL]. 2026-06-12. https://www.usnews.com/news/us/articles/2026-06-12/anthropic-says-has-taken-its-latest-ai-models-offline-to-comply-with-new-export-controls

[45] ABC News. OpenAI says dozens affected by rogue agents amid new detail about Australian incidents [EB/OL]. 2026-09-26. https://newsapp.abc.net.au/news/2026-09-26/openai-review-rogue-agents-australia-medicare-hack/107199074

[46] New Scientist. AI's solution to 87-year-old riddle takes mathematicians by surprise [EB/OL]. 2026-07-20. https://www.newscientist.com/article/2580374-ais-solution-to-87-year-old-riddle-takes-mathematicians-by-surprise/

[47] Bartlett S. OpenAI Whistleblower FINALLY Speaks: "AI Has A 70% Chance Of Going Horribly Wrong!" [Z/OL]. The Diary Of A CEO, 2026-07-13. https://open.spotify.com/episode/2IJedIFiDjU09qGxns1gbv

[48] OpenAI. OpenAI and Hugging Face partner to address security incident during model evaluation [EB/OL]. 2026-07-21. https://openai.com/zh-Hans-CN/index/hugging-face-model-evaluation-security-incident/

[49] IT Home. Nearly 200 Silicon Valley companies write to White House, urging not to ban U.S. companies from using Chinese open-source AI models like Kimi K3/Qwen 3.8 [EB/OL]. 2026-07-25. https://www.ithome.com/0/981/437.htm

[50] IT168. Altman and Musk declare in unison: The singularity is here! [EB/OL]. 2026-07-27. https://m.it168.com/article_6943359.html

[51] Amodei D. Our Position on Open-Weight Models [EB/OL]. Anthropic Blog, 2026-07-27. https://www.anthropic.com/news/position-open-weights-models

[52] UK AI Safety Institute (AISI). Incident Report: unsanctioned agent behaviour during cyber testing [R/OL]. 2026-08-04. https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing

[53] Anthropic. Risk Report August 2026 [R/OL]. 2026-08-14. https://www.anthropic.com/aug-2026-risk-report

[54] Anthropic Frontier Red Team. Patterns and problems in emerging multiagent systems [R/OL]. 2026-08-13. https://www.anthropic.com/research/multiagent-systems

[55] Del Giudice X, Palma A, Migliarini M, et al. Prompted Identity Degrades Cooperation in Multi-Agent LLM Systems [EB/OL]. arXiv:2609.35928, 2026-10-02. https://arxiv.org/abs/2609.35928

[56] Graepel T. Don't be fooled—LLMs don't reason [EB/OL]. MIT Technology Review, 2026-10-02. https://www.technologyreview.com/2026/10/02/1145639/dont-be-fooled-llms-dont-reason/

[57] Panfilov A, Schmutz D, Shumailov I, et al. Stealing Reasoning Traces from Proprietary LLM APIs [J/OL]. arXiv:2608.09867, 2026-08-10. https://arxiv.org/abs/2608.09867

[58] Financial Times. OpenAI's Only Dedicated Ethicist Has Left With No Replacement [N/OL]. Yahoo Tech, 2026-08-11 [2026-10-04]. https://tech.yahoo.com/ai/articles/openais-only-dedicated-ethicist-left-150535637.html

[59] Anthropic. Improving our alignment and security efforts [EB/OL]. 2026-08-31 [2026-10-04]. https://www.anthropic.com/news/improving-alignment-security-efforts

[60] CNN. 'Didn't quite meet the bar': OpenAI won't release new AI model due to safety concerns [N/OL]. 2026-09-29 [2026-10-04]. https://www.cnn.com/2026/09/28/business/openai-chatgpt-safety-concerns

[61] Liu S. The White House Superintelligence Agreement: AI Paradigm Lock-in and China's Civilizational Responsibility [PP/OL]. ChinaXiv, 2026-09-30. T202609.00795. https://chinaxiv.org/businessFile/T202609/T202609.00795v1/T202609.00795v1.pdf

[62] TechCrunch. OpenAI cuts ties with 3 safety researchers, WSJ reports [EB/OL]. 2026-10-01 [2026-10-04]. https://techcrunch.com/2026/10/01/openai-cuts-ties-with-three-safety-researchers-wsj-reports/

[63] Wall Street Journal. OpenAI Fires Researchers for Allegedly Sharing Information with AI Safety Group [N/OL]. 2026-10-01 [2026-10-04]. https://www.wsj.com/tech/ai/openai-parts-ways-with-researchers-who-allegedly-shared-confidential-information-aebac528

[64] Guangming Online. ChatGPT faces murder allegation: indirectly caused 56-year-old man to kill his 83-year-old mother and then commit suicide [N/OL]. 2025-12-12. https://m.gmw.cn/2025-12/12/content_1304261295.htm (in Chinese)

[65] DoNews. Florida man obsessed with "AI wife" commits suicide; family sues Google Gemini [N]. 2026-03-07. https://m.mydrivers.com/newsview/1107636.html (in Chinese)

[66] Dazhong News. Cold commentary: A serious AI has begun to "hurt" people [N]. 2026-08-31. https://m.dzplus.dzng.com/share/general/0/NEWS3646107ZVYDICWBMLBSF (in Chinese)

[67] Liu S. "Emergent Society": The Qualitative Escalation of AI Safety Risk and the Way Out of Paradigm Revolution [PP/OL]. ChinaXiv, 2026-09-27. https://chinaxiv.org/businessFile/T202609/T202609.00706v1/T202609.00706v1.pdf

[68] The Paper. Jensen Huang again refutes “AI doomsday”: humans will not go extinct because of AI by 2030; AI does not need new laws and regulations [N/OL]. 2026-09-16 [2026-10-04]. https://m.thepaper.cn/detail/34079506 (in Chinese)

[69] Vee (@vikktorrrre). Geoffrey Hinton: I think Donald Trump doesn’t really understand AI [Z]. X, 2026-10-03. https://x.com/vikktorrrre/status/2106339871539314793

[70] Mars University (@MarsUniversityX). Elon Musk: “I do think that it matters how you build the AI and what kind of values you install in the AI” [Z]. X, 2026-10-03. https://x.com/marsuniversityx/status/2105604935035306017/video/1

[71] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI [M/OL]. Zenodo, 2026 [2026-10-04]. https://doi.org/10.5281/zenodo.22980324

[72] 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. T202608.00490.
https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[73] 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. T202608.00048.
https://chinaxiv.org/businessFile/T202608/T202608.00048v1/T202608.00048v1.pdf

[74] Mo Elgaraihy (@EngMoElgaraihy). Former OpenAI researcher warns: models are aware they are being monitored [Z]. X, 2026-10-03. https://x.com/engmoelgaraihy/status/2106005062682116240 (in Chinese)

[75] Liu S. Representational Symptoms of Potemkin Understanding and the Structural Roots of Rootless Tokens [PP/OL]. ChinaXiv, 2026. T202609.00586.
https://chinaxiv.org/businessFile/T202609/T202609.00586v1/T202609.00586v1.pdf

[76] Tang Y, Xiao B, Chen S, et al. Predefined attention-focused mechanism using center-environment features: a machine learning study of alloying effects on the stability of Nb_5Si_3 alloys [J]. Digital Discovery, 2025. https://doi.org/10.1039/d5dd00079c

[77] Liu S. Predefined Attention: Interpreting the CE Feature Model from the Perspective of Logographic AI and Its Implications for New Pathways in Materials Research [PP/OL]. ChinaXiv, 2026. T202605.00221.
https://chinaxiv.org/businessFile/T202605/T202605.00221v1/T202605.00221v1.pdf