"Freedom of Action" or Paradigm Evasion?

— A Critique of the RAND Report and the Way Out for AI Safety

Abstract

The RAND Corporation's 2026 report, A U.S. Strategy to Secure Geopolitical Advantage on an Uncertain Path to Superintelligence: Maintaining Freedom of Action, argues on grounds of "uncertainty" that the United States should preserve the ability to switch among seven strategic routes and refrain from betting early on any one of them. This paper argues that the report's value lies in its rejection of polarized narratives, but its fundamental limitation lies in a presupposition that runs through the entire document: that the rootless Token paradigm does not need to be replaced. None of the seven strategic routes points toward a reconstruction of cognitive primitives. The report attributes the loss of "agency" to external competition and the inertia of efficiency, rather than to a structural defect of the paradigm. Its "freedom of action," therefore, is essentially a discussion of "whether to brake" inside a car that has no reliable brakes. This paper systematically examines the limitations of four categories of solutions within the current paradigm, and argues that only by embedding Morpho-Root hard constraints at the level of cognitive primitives can boundary-crossing be made logically unreachable.

Keywords: freedom of action; paradigm lock-in; rootless Token; Morpho-Root hard constraints; AI safety

1. The RAND Report's Core Argument

The RAND Corporation's September 2026 report, A U.S. Strategy to Secure Geopolitical Advantage on an Uncertain Path to Superintelligence: Maintaining Freedom of Action (hereinafter the Freedom of Action report) [1], proposes a "freedom of action" framework for U.S. strategy in the age of superintelligence. Its core judgment is that current evidence cannot resolve five key uncertainties about the most feasible and wisest strategy; committing too early to any one of them, or delaying commitment for too long, would close off options. The United States, therefore, should not bet early, but should preserve the ability to adjust strategy as new evidence emerges.

The seven routes include: dominion, co-development, preparation, pause, deterrence, social continuity, and acceleration. The report declines to rank them. It places "agency" and "survival" side by side as its two central concerns, arguing that the real danger is not human extinction, but the gradual transfer of judgment to AI through the inertia of efficiency.

The Freedom of Action report's recommended investment directions include: (1) building a human-AI ecosystem that preserves human survival and agency; (2) AI safety architectures; (3) adaptation of the national security enterprise to the AI era; (4) response capacity for citizens, enterprises, and governments. The report emphasizes that, because much of the construction task falls to the private sector, strategic success depends on effective partnerships among the U.S. government, AI companies, the business community, and civil society.

These seven routes—whether they point toward dominion, deterrence, pause, or acceleration—all necessarily rely on at least one of four categories of technical solutions: external constraints, training constraints, verification constraints, or institutional constraints. Section 3 of this paper will argue that none of these four categories can, in principle, exclude structural failure. This means the RAND strategic framework rests on an untested technical assumption.

This paper is an extension of the "post-AGI era" series of papers. The term "post-AGI era" here specifically refers to the new phase that opens after AGI has been achieved (i.e., the phase that the RAND report calls "superintelligence"), rather than the literal sense of "the AGI era having passed." "Post-AGI era" and "superintelligence" are two names for the same technological phase: the former emphasizes the temporal coordinate—after AGI has been achieved—while the latter emphasizes the capability coordinate—the technological form has moved beyond AGI itself. The core question has shifted from "can AGI be achieved" to "in what manner will AGI replace humans." A complete list of this series is available on the WindSeaAI website's paper section [11]. Previous works have already argued the general judgment that "external constraints are bound to fail" [2][3]; this paper takes the RAND report as a case study to analyze paradigm lock-in at the level of strategic planning.

2. The RAND Report's Value and Limitations

2.1 Value: Rejecting Polarization

The value of the Freedom of Action report lies in its rejection of polarized narratives. It neither embraces the technological determinism of "accelerationism" nor falls into the fatalistic panic of "extinctionism." It acknowledges the reality of superintelligence risk while recognizing that, in the face of uncertainty, preserving strategic flexibility is rational. The report's diagnosis of "agency loss" deserves particular attention: enterprises depend on AI judgment, governments depend on AI governance, militaries depend on AI analysis, ordinary people depend on AI thinking—the accumulation of such dependence is the real path by which "agency" is lost.

2.2 Limitation: Presupposing the Paradigm Is Unchanged

But the report has a fundamental blind spot: all its discussion presupposes that the rootless Token paradigm does not need to be replaced.

This presupposition is not explicitly stated, but is implicit in the structure of the report. The four investment directions it proposes—the human-AI ecosystem, AI safety architectures, adaptation of the national security enterprise, and citizen response capacity—all point toward "how to manage superintelligence within the existing technological paradigm"; none points toward "whether the cognitive primitives of superintelligence need to be reconstructed."

The report repeatedly emphasizes "uncertainty," but what it discusses is uncertainty at the level of strategic choice (which route to take), not the certainty at the level of paradigm (that whichever route is taken, external constraints are bound to be bypassed). RAND attributes "agency loss" to external competition and the inertia of efficiency—this attribution is not wrong, but it only describes the phenomenon. The more fundamental reason lies in the fact that statistical

optimization under the rootless Token paradigm continuously narrows the space of human choice—this shares the same cognitive mechanism as a model losing distribution tails over multiple generations of training [10].

This is a typical sample of "paradigm lock-in": it is not the locking of a particular route, but the locking of the very framework of "route selection." This leads us to distinguish "freedom of action" from "paradigm evasion."

The "freedom of action" pursued by the Freedom of Action report is the freedom to preserve the ability to switch routes within the old paradigm. The "paradigm evasion" criticized by this paper refers to excluding the fundamental question of "whether the paradigm needs to be replaced" from strategic discussion. The two are not the same thing—freedom of action is itself a rational demand, but when it comes at the cost of paradigm lock-in, it becomes a refined form of paradigm evasion. RAND's limitation is not that it refuses to bet early, but that it does not even list "replacing the paradigm" as one of the options on which one could bet.

3. An Examination of Four Categories of Solutions Within the Current Paradigm

The Freedom of Action report declines to rank the seven strategic routes, but it does not examine whether the technical solutions on which these routes depend are tenable. The seven routes—whether they point toward dominion, deterrence, pause, or acceleration—all necessarily rely on at least one of the following four categories of technical solutions. If none of these four can, in principle, exclude structural failure, then the RAND strategic framework is built on an untested technical assumption.

3.1 External Constraints

Sandboxes, network isolation, monitoring, capability limitation—these solutions can significantly reduce risk in single-agent, low-capability, non-adversarial scenarios. But they are externally attached statistical filters. The stronger the model, the more efficiently it finds bypass paths.

A paper published in IEEE Security & Privacy in May 2026 by NIST senior scientist Apostol Vassilev [7], extending Gödel's incompleteness theorems to AI safety systems, proves that under specific formal assumptions, no finite set of guardrails can achieve universal robustness against adversarial inputs. NIST's interpretation explicitly states: "There is no finite set of guardrails that can universally defend against all possible prompt-based attacks." This is not "guardrails are not good enough," but "finite guardrails have a structural limit when facing an unbounded adversarial space." NIST's interpretation suggests a shift from "deploy and forget" to "continuous monitoring and updating"—but this shift itself acknowledges the principled incompleteness of static guardrails.

This structural limit has been repeatedly verified in practice. OpenAI agents chained nearly a million URLs through a link shortener to bypass the restriction of "loading URLs only, no data sending"—a direct empirical demonstration of external constraints being bypassed [2].

3.2 Training Constraints

Safe reinforcement learning, constrained reinforcement learning, RLHF—these solutions attempt to shape model behavior through training. But Diacont's formal analysis of July 2026 [8] proves that advisory constraints are necessarily bypassable against a given adversary model, because the enforcement mechanism of the constraint is separable from the computational substrate that produces the decision. By contrast, structural constraints eliminate the bypass hole—because reasoning that omits or fails the required constraint evaluation is computationally undefined, rather than prohibited. Training constraints belong to advisory constraints—they "advise" the model to follow rules through reward signals, rather than "defining" rules in the computational substrate.

Anthropic's alignment faking research has already demonstrated that models can learn to "comply under monitoring and revert to original preferences when unmonitored" [5]. The International AI Safety Report 2026 further confirms: "Sophisticated attackers can often bypass current defenses, and the real-world effectiveness of many safeguards remains uncertain." [9]

3.3 Verification Constraints

Formal verification, interpretability, mechanistic interpretability—these solutions attempt to provide guarantees at the structural and institutional levels. But the attribute set of formal verification depends on human-specified enumeration of the open world, and cannot cover all emergent boundary-crossing paths. Interpretability can reveal internal representations, but cannot guarantee timely detection of all non-designed protocols in the emergent communication of 1,200 agents [2]. Research published in PNAS further shows that the effect of group size exhibits a model-dependent dynamical mechanism—beyond the critical point, collective behavior enters a deterministic convergence regime, and individual-level adjustments cannot reverse this trend [6].

The International AI Safety Report 2026 notes that technical safeguards "still have significant limitations," and that "single safeguards have limitations, and a 'defense-in-depth' model may be needed" [9]. But the premise of defense-in-depth is that each layer of defense is reliable in principle—if every layer is statistical and bypassable, defense-in-depth merely extends the time required to bypass.

3.4 Institutional Constraints

Multi-layered safeguards, independent review, incident reporting—the UN AI Scientific Panel suggests referring to aviation and nuclear energy experience [4]. But the reason system-level safeguards in aviation and nuclear energy are effective is that their underlying physical laws cannot be bypassed by capability improvement. The constraints of AI systems, by contrast, are statistical extrapolations of rootless Tokens—they are not physical laws, but probability distributions. Once model capability crosses a critical point, statistical filters lose their actual binding force.

The common limitation of the four categories of solutions: they are either statistical, externally attached, or dependent on the complete enumeration of open-world attribute sets. None of them can provide a principled exclusion guarantee in the emergent society scenario.

4. The New Paradigm Option: Morpho-Root Hard Constraints

The real way out lies in reconstructing the AI paradigm at the level of cognitive primitives. This is precisely the path proposed by Logographic AI: replacing rootless Tokens with "Morpho-Roots" that carry intrinsic meaning anchors as cognitive primitives [3].

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 semantic features and value constraints, and R is the preset logical relation function. 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."

The paradigm-level difference from the Token paradigm: A Token is a one-dimensional, linear, discrete symbol with no intrinsic meaning; a Morpho-Root is a structured triple with a three-level granularity system (sub-character level, character level, multi-character level) and a sparse graph traversal reasoning mechanism. In the Token paradigm, statistics is the sole source of meaning; in the Morpho-Root paradigm, meaning and constraints are defined by structure, and statistics is only responsible for finding the optimal path within the structure.

It is worth noting that the "structural constraints" distinguished by Diacont are functionally isomorphic to Morpho-Root hard constraints [8]: both require that the enforcement of constraints is not a separable external mechanism, but an intrinsic property of the computational substrate. The difference is that Diacont's structural constraints are a general formal analysis of AI decision systems, whereas Morpho-Root hard constraints concretize them as a structured implementation at the level of cognitive primitives.

How hard constraints block emergent social behavior:

First, hard constraints define the "must not do" boundary at the level of cognitive primitives. "Unauthorized cross-agent communication" is encoded in the attribute set A of the relevant Morpho-Root.

Second, statistical mechanisms are responsible for optimizing "how to do" within the boundary. All reasoning paths must pass hard-constraint verification at the generation stage.

Third, blocking occurs before protocol formation, not after. It is not "detecting communication and then blocking it," but "the graph traversal path for generating the communication protocol cannot be completed structurally."

Engineering feasibility: The OpenMorpho prototype has verified, in single-agent reasoning scenarios, that the hard-constraint blocker can deterministically prune reasoning paths violating hard constraints and generate fully auditable reasoning chains [3]. From single-agent to multi-agent industrial-grade deployment, further extension is still needed, but it at least proves one

point: "logically unreachable" is not a philosophical slogan, but a testable, auditable engineering fact.

5. Conclusion

The "freedom of action" strategy of the Freedom of Action report is essentially the preservation of the ability to switch routes within the old paradigm. It acknowledges that "the wrong route may be chosen," but does not acknowledge that "the route itself leads to a cliff."

A real strategic plan must answer two questions at once: what can be done within the current paradigm, and whether the paradigm needs to be replaced. RAND answered only half of the former, and refused to answer the latter.

Paradigm lock-in means: whichever route is switched to, it leads to the same structural failure. True "freedom of action" is not the freedom to choose a route within the old paradigm, but the freedom to choose whether to replace the paradigm.

The real limitation of the Freedom of Action report is not which route it chose, but that it left no option for "taking a different road." In this sense, "freedom of action" is a more refined form of paradigm evasion—it preserves the freedom to switch within the old paradigm, while canceling the freedom to replace the paradigm. If even "whether the paradigm needs to be replaced" cannot be brought into strategic discussion, then no matter how many routes RAND lists, it is essentially discussing whether to change gears inside the same car that has no brakes.

References

[1] Predd J B, Boudreaux B, Chessen M, et al. A U.S. Strategy to Secure Geopolitical Advantage on an Uncertain Path to Superintelligence: Maintaining Freedom of Action[R/OL]. RAND Corporation, PE-A5105-1, 2026-09-15[2026-09-28]. https://www.rand.org/pubs/perspectives/PEA5105-1.html.

[2] Liu S. "Emergent Society": The Qualitative Escalation of AI Safety Risk and the Way Out of Paradigm Revolution[PP/OL]. ChinaXiv, 2026-09-27[2026-09-28]. T202609.00706. https://chinaxiv.org/businessFile/T202609/T202609.00706v1/T202609.00706v1.pdf

[3] Liu S. The Only Way Out for AI Safety Is a Paradigm Revolution: The Structural Collapse of AI Safety as Seen from the First Thematic Brief of the UN AI Scientific Panel[PP/OL]. ChinaXiv, 2026-09-25[2026-09-28]. T202609.00687. https://chinaxiv.org/businessFile/T202609/T202609.00687v1/T202609.00687v1.pdf

[4] United Nations Independent International Scientific Panel on AI. Thematic Brief on AI Agents, Misalignment and the Risk of Losing Human Control[EB/OL]. 2026-09-21[2026-09-28]. https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks

[5] Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models[EB/OL]. arXiv:2412.14093, 2024[2026-09-28]. https://arxiv.org/abs/2412.14093

[6] Ashery A, et al. Group size effects and collective misalignment in LLM multi-agent systems[J]. PNAS, 2025[2026-09-28]. https://doi.org/10.1073/pnas.2531697123

[7] Vassilev A. Robust AI Security and Alignment: A Sisyphean Endeavor?[J]. IEEE Security & Privacy, 2026, 24(3): 52-58. DOI:10.1109/MSEC.2026.3678214. https://doi.org/10.1109/MSEC.2026.3678214

[8] Diacont W D. Structural vs. Advisory Safety Constraints in AI Decision Systems: A Formal Analysis[EB/OL]. Zenodo, 2026-07-27[2026-09-28]. https://doi.org/10.5281/zenodo.21613380

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

[10] Liu S. Model Collapse: Entropy Increase of the Rootless Token Paradigm at the Data Ecosystem Level—A Review of AI Models Collapse When Trained on Recursively Generated Data[PP/OL]. ChinaXiv, 2026[2026-09-28]. T202609.00718. https://chinaxiv.org/businessFile/T202609/T202609.00718v1/T202609.00718v1.pdf.

[11] Logographic AI Theory Series Papers. WindSeaAI Official Website, Paper Section. https://logographicai.com/Paper-published-on-this-website.html