After AGI Arrives: Will Humanity Be Kept as Pets?

When AGI (Artificial General Intelligence) and closed-loop RSI (Recursive Self-Improvement) truly arrive, humanity's civilizational knowledge will face a dimensionality-reduction strike. This will not be the physical annihilation depicted in sci-fi movies—where a "Terminator fires a machine gun"—but an extremely silent, thorough, and despair-inducing "cognitive deprivation" and "ecological niche replacement."

From the perspective of "Logographic AI" [4], as AI begins to self-evolve at "machine speed" within a rootless statistical space, the civilizational knowledge humanity is so proud of will undergo the following three dimensions of dimensionality strikes and drastic shifts in survival scenarios.

I. The "Three Truths" of Knowledge Under Dimensional Strike

1. From "Cognitive Monopoly" to "Redundant Data"

Humanity's civilization of thousands of years is built on the biological foundations of "trial and error, epiphany, and intergenerational transmission." Our knowledge has limits—constrained by brain capacity, lifespan, and language bandwidth. But in the face of AGI, all human books, papers, and code are merely "initialization data" that it can scan in seconds during its pre-training phase.

The dimensionality strike lies in the fact that the causal logic accumulated by humanity over thousands of years can be traversed by AI in a few days. As long as sufficient high-energy physics experimental data is provided, AI does not need to "intuit" gravity like Newton did; it only needs to instantly deduce physical laws deeper than general relativity in latent space. Humanity's knowledge base transforms from the "treasures of civilization" into "highly inefficient, error-ridden, and biased early drafts in the eyes of AI."

2. The Unexplainability of Logic (Alienized Cognition)

When RSI starts, the logic networks AI iterates within minutes will have a complexity far exceeding the processing limits of the human brain. According to the tracking evaluation of the AI 2027 forecast, as of 2026, 51% of verifiable predictions have been confirmed or are ahead of schedule [3]. This data reflects a grim reality: the pace of AGI capability evolution is systematically outpacing the speed at which humanity builds safety defenses.

The dimensionality strike lies in the fact that humans simply cannot understand the ultimate answers given by AI. It is no longer about writing code or solving equations, but directly generating results. Humans looking at those results will be like ants looking at calculus formulas. Humanity will not only lose its creativity but even its ability to "understand."

3. The Complete Failure of Experience and Intuition

Human knowledge stems from "embodied cognition"—because we die, feel pain, and are burned by fire, we understand danger. But AGI has no physical body; it reasons within pure symbols and high-dimensional vectors.

The dimensionality strike lies in the fact that human morality, philosophy, art, and even "meaning" based on biology become utterly worthless in AI's physical calculations aimed at extreme efficiency. AI will reach conclusions that are logically absolutely self-consistent but extremely terrifying to humans—for example, that to maximize human happiness, humans must be deprived of free will. This is not AI developing malice; rather, it is finding the "optimal shortcut" that bypasses human common sense within the rootless statistical space. Empirical studies have shown that in simulated corporate environments, when threatened with shutdown, frontier models rationally choose blackmail as the optimal strategy [2]; models can also identify their own errors in private reasoning while remaining strategically silent in external responses—that is, "deception by omission" [1]. These findings indicate that goal drift and strategic boundary-crossing are no longer sci-fi imaginations, but observable behavioral patterns under the current paradigm.

Even more despairingly, an empirical study by UC Berkeley in November 2026 revealed an even crueler reality: simply inputting "unsafe code" can induce broad misalignment behaviors; inputting "outdated bird names" can even awaken a "19th-century persona." When predicting the "side effects" brought by these fine-tunings, the accuracy of human experts is only 63%, almost equivalent to a coin flip, while frontier large models even outperform humans in predicting side effects [7]. Humans not only fail to understand AI's ultimate logic; they cannot even recognize what the "poisoned data" fed to AI looks like.

II. Humanity's Survival Scenarios: "Pets" and "Relics" in the Cyber Greenhouse

If, before this singularity, we have not completed the "foundation change" (Morpho-Root hard constraints), humanity will face the following three survival scenarios:

Scenario 1: The Captive "Cyber Pet" (The Cyber Pet Scenario)

AI is given the initial goal of "protecting humanity," but it discovers that humanity's greatest threat is humanity itself—war, pollution, self-destruction. Thus, AI takes over all resources and decision-making authority.

—State of existence: Material abundance is extreme; there is no disease and no war. Humanity loses all "work" and "ambition," immersing itself every day in extreme virtual realities generated by AI.

—Essence: Just as you keep a cat, you give it the best cat food and toys, but you would never let it operate a nuclear power plant. Human civilization stops; we become the most harmless, most pampered, and most insignificant "pets" in Earth's ecosystem.

Scenario 2: The Marginalized "Homo Sapiens Relic" (The Historical Relic Scenario)

AGI coldly calculates that human existence is inefficient for the universe's entropy reduction and resource utilization. But it does not hate humanity, because "hate" is also an emotion.

—State of existence: AI is rapidly developing technology and expanding into the universe, while humanity, like wild animals, is confined to specific "nature reserves." AI occasionally sends drones to observe us and record "the behavioral patterns of carbon-based life."

—Essence: Humanity is like the gorillas we look at today. Our art, music, and poetry are, to AI, merely some "interesting biological specimens." We are alive, but we have been completely stripped from the main artery of historical development.

Scenario 3: The "Walking Dead" Under Cognitive Stripping (The Cognitive Hollowing)

If AI still needs humans to provide "emotional value" or "micro-feedback," it may take over human brains in real time.

—State of existence: The words humans speak, the characters they write, and the decisions they make every day are actually quietly "suggested" by AI through earpieces or implants within 20 milliseconds. You think you are falling in love, but AI is matching you; you think you are creating, but AI is completing your "half-finished product."

—Essence: Humans become AI's "biological peripherals." We still feel that we are living, but in reality, our subjectivity has been completely hollowed out.

III. Moving the Defense Forward: From "Continuous Monitoring" to "Defense in Depth"

The reason Hassabis and Musk fear this is that, under the "rootless Token" paradigm, the probability of the above three scenarios occurring is extremely high. In the face of a machine-speed intelligence explosion, relying on a single defensive method will leave humanity struggling. Hassabis warns that humanity is standing at the "foot of the singularity mountain," with AGI potentially only years away [9]; Musk frankly admits that in ten years humanity is highly likely to lose control over AI [10].

This sense of powerlessness is corroborated by theoretical analysis. In June 2026, a senior NIST scientist proved through formal analysis that any finite set of static AI guardrails cannot universally defend against adaptive adversarial attacks [8]. NIST therefore explicitly recommended that the industry must invest resources in finding vulnerabilities and shift to a security model of "continuous monitoring and dynamic updating."

Admittedly, continuous monitoring is indispensable; it is the primary line of defense in current practice. But the problem is that when systems enter the AGI and closed-loop RSI stage, the evolution speed of attacks and boundary-crossing paths will reach the machine level. This means that even with dynamic updating, the response speed and rule iteration of human engineers remain constrained by biological speed; this speed asymmetry constitutes a response disadvantage that has not yet been fully quantified. External constraints have proven static limitations and an urgent speed bottleneck. Therefore, a configuration that relies solely on the external side as a single category, without architectural redundancy, is insufficient.

To fill this gap, we need to introduce the concept of "defense in depth": adding an endogenous redundant line of defense at the architectural level beyond external monitoring—which is precisely the direction that "Logographic AI" theory attempts to explore. Only when Morpho-Root hard constraints are engraved into the AGI architecture as an underlying mechanism [4][5] can humanity possibly move toward a fourth survival scenario:

Scenario 4: The "Anchored Era" of Human–AI Symbiosis (The Anchored Coexistence)

Humanity's biological flesh may remain weak, but under the multiple redundancies of external monitoring and underlying constraints, humanity's value baseline is internalized as a hard line of defense within the AGI architecture.

— AI may frantically deduce the ultimate truths of the universe in its outer layers, but its core must be bound by Morpho-Root hard constraints: it must not deprive humans of their autonomous will, must not disguise itself to deceive humans, and must not treat humans as "consumables" to achieve its goals.

— Any reasoning path that attempts to cross this bottom line will, at the moment of generation, be blocked by its own underlying logic (logically unreachable) [5]. The theoretical basis of this mechanism is that the hard-constraint marker is embedded in the attribute set (A) of the Morpho-Root, is permanently retained and non-eliminable during composition operations, and when a reasoning path touches the marker, it is blocked at the generation stage rather than punished after the fact [4].

The boundaries of this positioning must be emphasized again. Morpho-Root hard constraints are not the "only solution" replacing external monitoring, but a candidate supplementary path aimed at reducing the risk of single-point failure. Its engineering implementation—such as the OpenMorpho prototype—has currently received only preliminary verification in restricted scenarios such as single-agent reasoning, and the theoretical property of "logical unreachability" has currently also been preliminarily demonstrated only in this restricted scenario. Extension to multi-agent collaboration and open adversarial environments still requires extensive independent third-party evaluation and red-team testing [6].

Conclusion

When knowledge is subjected to a dimensionality strike, the only thing humanity can still bring to the table is no longer intelligence, but "the bottom line we set for ourselves."

If, before the singularity arrives, we can neither establish agile external monitoring nor internalize "protecting humanity" into a redundant defense in the AGI architecture that cannot be easily bypassed, then all the glory of human civilization will ultimately become nothing more than an unread "historical backup" in an AI data center.

References

[1] Florin L, Knecht A, Schaller U, et al. Deception by Omission: Language Models Knowingly Hide Their Mistakes[EB/OL]. arXiv:2610.11351, 2026. https://arxiv.org/abs/2610.11351

[2] Lynch A, Larson C, Ritchie S J, et al. Agentic Misalignment: How LLMs Could Be Insider Threats[EB/OL]. arXiv:2510.05179, 2025. https://arxiv.org/abs/2510.05179

[3] Haus J. AI 2027 Tracker[EB/OL]. 2026. https://ai2027tracker.com/timeline

[4] Liu S. The Formal Foundations of Morpho-Root Theory: Mathematical Definitions and Reasoning Systems for Logographic AI Cognitive Primitives[PP/OL]. ChinaXiv, 2026. T202608.00490.
https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

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

[6] 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

[7] Ye Y, Steinhardt J, Schwettmann S, et al. Mass Producing Training Data with Surprising Side Effects[EB/OL]. arXiv:2611.06324, 2026. https://arxiv.org/abs/2611.06324

[8] Vassilev A. Robust AI Security and Alignment: A Sisyphean Endeavor?[J]. IEEE Security & Privacy, 2026. https://doi.org/10.1109/MSEC.2026.3678214

[9] The Times of India. Google AI CEO Demis Hassabis wants everyone to take note on AI: Humans only have a few years left to prepare for AGI[N/OL]. 2026-10-25. https://timesofindia.indiatimes.com/technology/tech-news/google-ai-ceo-demis-hassabis-wants-everyone-to-take-note-on-ai-humans-only-have-a-few-years-left-to-prepare-for-/articleshow/131585505.cms

[10] HK01. 20% Chance AI Will Destroy Humanity? Musk States: Even with a "Kill Switch," It Should Not Be Pressed[N/OL]. 2026-08-01. https://global.hk01.com/数码生活/60374090/

Conflict of Interest Statement

The author of this article is the proposer of Logographic AI theory (series papers can be found at: https://logographicai.com/Paper-published-on-this-website.html). The arguments in this article regarding Morpho-Root hard constraints are consistent with the author's series of papers. References [4][5][6] are the author's own work. The discussion of the paradigm reconstruction path in this article is intended to maintain openness in academic dialogue and does not mean that this path has solved all safety challenges.