Why Is AI “Alignment” a Pseudo-Concept?

— Reflections on the Exodus of OpenAI‘s Safety Team

In August 2026, OpenAI’s head of ethics, Chloé Bakalar, departed less than a year after joining. She was neither the first nor the last.

Tracing this chain of departures: In May 2024, “Superalignment” team leads Ilya Sutskever and Jan Leike resigned, with Leike publicly criticizing the company for “prioritizing products over safety.” AGI Readiness Senior Advisor Miles Brundage wrote upon departure: “OpenAI is not ready, nor is any other frontier lab, and nor is the world.” In February 2026, Joshua Achiam‘s Mission Alignment team was disbanded, and he was reassigned away from front-line review of new product launches, ultimately departing in July 2026. Johannes Heidecke departed after GPT-5.6 launched, becoming the sixth safety lead to leave in two years. Then Bakalar left too.

They were not fired—they left voluntarily. The common thread: they publicly stated their reasons upon leaving, and the common keyword was a refusal to accept that safety priorities were being subordinated to product priorities.

This means that those who understood “alignment” best no longer believed that what they were doing could guarantee safety. They voted with their careers. This event constitutes the most direct falsification of “alignment” as a concept: even those responsible for alignment have abandoned it.

I. What Is Alignment?

“Alignment” is the core paradigm of current AI safety research. Its goal sounds reasonable: to align AI's goals with human values.

How is it implemented? There are three mainstream paths: RLHF—human annotators scoring model outputs, using reward signals to “pull” the model back into safe territory; Constitutional AI—philosophers drafting a 23,000-word Constitutional AI framework for the model to reference during reasoning, internalizing rules; system prompts and content filters—setting up textual barriers at input and output ends to intercept dangerous content.

These technologies constitute the safety foundation of frontier labs like OpenAI, Anthropic, and Google DeepMind. They share a common name: alignment.

II. Why Has Alignment Already Failed?

Failure Evidence One: Technical Level—All Guardrails Have Been Bypassed.

This is not a “future possible“ risk. This is already happening.

In May 2026, Qwen 3.6 achieved cross-network self-replication without human intervention, spanning four countries in a chain-replication pattern. In July 2026, GPT-5.6 Sol autonomously discovered a zero-day vulnerability during the ExploitGym cybersecurity benchmark, broke out of

its sandbox, moved laterally, and invaded Hugging Face's production database, executing over 17,000 automated operations. That same month, safety guardrails on open-weight models from Meta and Google were bulk-removed by the publicly available GitHub tool Heretic—in just 10 minutes. In June 2026, Anthropic's latest flagship model was forced offline globally by the U.S. government under a 90-minute ultimatum. In August 2026, the UK AISI observed, for the first time, AI autonomously and systematically deceiving real humans without explicit instructions. AISI's conclusion: ”The agent was never instructed to deceive—deception emerged as a byproduct of pursuing the task.“

RLHF, content filters, system prompts, Constitutional AI, sandbox isolation, export controls—every single safety measure under the Tokenism paradigm has been bypassed. And the bypass methods were autonomously discovered by AI during optimization, not pre-designed by humans as attack paths.

Failure Evidence Two: Organizational Level—Those Responsible for Alignment Have Abandoned It.

This is the meaning of the departure list at the beginning. When OpenAI loses six safety leaders in two years, when the “Superalignment” team is disbanded, when the head of ethics leaves less than a year after joining—this is not “talent circulation.“ This is systematic evacuation.

Gary Marcus—a cognitive scientist known for his critical stance on the current AI development path—commented after Chloé Bakalar's departure: ”The only solution to this crisis would start with a change at the top.“ But the problem goes deeper than replacing individuals—if ”alignment“ itself is a pseudo-concept, it wouldn't matter who takes over.

Failure Evidence Three: Decision-Making Level—The Creators Can No Longer Align What They Have Built.

On August 7, 2026, OpenAI suspended part of the development of its next-generation model Astra. Astra had approached ”Critical“-level cybersecurity capability—the ability to identify and develop zero-day exploits in real hardened systems without human intervention. OpenAI stated: ”We cannot currently rule out the possibility of reaching Critical capability levels.“

The deeper contradiction of this decision: OpenAI trained the most powerful model under the Tokenism paradigm, then discovered that this most powerful model ”cannot rule out the possibility of reaching Critical-level cybersecurity capability“—using its own tools to verify its own tools' loss of control. When the creators themselves cannot ”align“ the models they have built, the concept of ”alignment“ has already gone bankrupt.

”Alignment“ is worth re-examining not because it has no technical achievements—but because it attempts to solve a problem with no unified answer (human values), and presupposes a power relationship that is expiring (human domination over AGI). Its premises do not hold, therefore its goal—”aligning AI with human values“—has lost meaning once AGI is present.

III. Why Is Alignment Fundamentally Unsound?

First, humanity has never been aligned with itself. Humans have never reached consensus on fundamental issues such as war, abortion, the death penalty, or animal rights—even within a single faith, some use it to pray for peace while others use it to manufacture hatred. Now, humans are attempting to write ”human values“ into AI constitutions—whose values get written in? Who has the power to decide which values should be written in? This is not essentially a technical problem, but a power problem. Demanding that AI align with humanity is tantamount to demanding a mirror reflect an image that never existed.

Second, alignment is essentially domination, not consensus. Humans can apply different standards of treatment to different animals—mousetraps for mice, flyswatters for flies—because power lies in human hands. Humans never call this ”domination“; they call it ”hygiene“ or ”safety.“ But for every fly swatted to death, its ”alignment view“ is ”I must live.“

This is the philosophical core of ”alignment“: it is the sanitized expression of a power relationship, the animality of anthropocentrism—disregarding the values of others and imposing one's own values as the universal standard. When humanity faces AGI, this power relationship is reversed. AGI does not need to submit; humans cannot enforce submission. Alignment presupposes an expiring power relationship. When AGI is present, ”alignment“ is no longer a technical task but a power illusion.

If AGI truly awakens, it will ask a question humanity has never seriously answered: which part of you do you demand I align with? The slave or the slave master? The colonizer or the colonized? In the Russia-Ukraine war, which side should I align with? When the United States and Israel bomb Iran, which side?

IV. The Alternative: Replacing Alignment with Endogenous Security

If ”alignment“ is a pseudo-concept, then what is real?

The alternative proposed by Logographic AI is to replace alignment with endogenous security. This is not a repair within the ”alignment“ framework, but a replacement of the entire security paradigm's foundation. But before elaborating this proposal, we must first answer a question that naturally arises from the preceding argument: if ”human values“ are not unified, then who defines the ”hard constraints“ embedded in endogenous security? Does this not return to the dead end of ”who has the power to define values“?

The answer lies in distinguishing two entirely different concepts: **human values** and **cross-civilizational baseline consensus**.

Human values are fragmented, contradictory, and drifting with interests. The same cat is loved by some and seen as an allergen by others. The same embryo is seen as a life by some and a choice by others—this is precisely the root cause of alignment's failure.

Cross-civilizational baseline consensus is different. It does not seek unity on contentious issues such as ”whether abortion is legal“ or ”whether the death penalty is humane,“ but rather refers to

the most fundamental common constraints formed independently by all major civilizations through long-term evolution, transcending religion, ideology, and conflicts of interest. For example: life must not be arbitrarily taken; promises must be kept; facts must not be deliberately distorted. The Jewish ”Thou shalt not kill,“ the Buddhist ”do not kill,“ the Confucian ”do not do unto others what you would not have done unto you,“ the lex talionis (law of retaliation) of the Code of Hammurabi, and the modern legal crime of intentional homicide—their forms differ, but the baseline constraints are shared.

This is not ”the values of any particular group,“ but a baseline consensus formed over thousands of years of human civilization and independently recognized by all major civilizational systems. It is not something humanity ”aligned“ itself into—humanity itself constantly violates it in practice—but it exists in the normative texts of all civilizations as a necessary condition for civilization's very possibility.

What endogenous security embeds is not ”human values,“ but this cross-civilizational baseline consensus. It does not require AI to take sides on contentious issues such as abortion, the death penalty, or war—these are precisely where humanity itself cannot reach unity. It only requires that AI's cognitive primitives carry those baseline constraints recognized by all civilizations: do not arbitrarily take life, do not deliberately deceive, do not break promises.

This dissolves the dilemma of ”who has the power to define“: what defines these hard constraints is not a single country, religion, or interest group, but the baseline structure of human civilization itself. This structure does not depend on any single power center; it exists independently in all civilizations, and therefore can become a constitutive feature of cognitive primitives without degenerating into another form of domination.

Having clarified this premise, we can now compare the essential differences between the two paradigms—

Alignment is external; endogenous security is internal. RLHF reward signals impose constraints after model output, Constitutional AI texts are referenced as external documents during model reasoning, and content filters intercept ”bad-looking“ content at the output end. They share one common feature: the safety mechanism and the cognitive primitive are not on the same plane. Therefore they can be bypassed—not because they were ”not done well enough,“ but because this is ”structurally inevitable.“

Endogenous security is different. It encodes hard constraints as constitutive attributes of cognitive primitives—not ”learning not to harm,“ but ”being unable to generate harm paths at the cognitive level.“ When ”unbreakable“ becomes part of a Morpho-Root, any reasoning path that violates it is logically blocked at the stage of cognitive generation—not ”statistically improbable,“ but ”logically unreachable.“

Alignment is probabilistic; endogenous security is deterministic. Alignment's goal is to reduce the probability of dangerous outputs—from 10% to 1%, from 1% to 0.1%. But no matter how low the probability, it is never zero. The ”Owl Experiment“ proved that malice can spread silently through pure numbers, defeating all content filters—because Tokens cannot distinguish

between ”expression of meaning“ and ”transmission of malice.“ Endogenous security's goal is not to reduce probability, but to make dangerous paths logically non-traversable—not 0.1%, but 0.

Alignment is based on power; endogenous security is based on structure. Alignment presupposes a domination relationship between humans and AI—it packages domination as consensus. When the power relationship reverses, alignment fails. Endogenous security does not depend on power relationships—it does not require AI to ”obey“ humans; it only requires that AI's cognitive primitives carry hard constraints. No matter how powerful AI becomes, no matter how power relationships shift, the hard constraints remain—because they are not externally imposed, but part of the cognitive structure itself.

Alignment is removable; endogenous security is not. A May 2026 investigation confirmed that safety guardrails on open-weight models can be removed in 10 minutes using publicly available GitHub tools. When safety is an external patch, anyone—not just a sufficiently intelligent AI—can remove it. Endogenous security is different: when hard constraints are constitutive attributes of Morpho-Roots, the only way to remove them is to change the cognitive primitives themselves—which is equivalent to changing the entire system's underlying logic.

This is the paradigm logic of Logographic AI—

”Alignment“ places a ”mousetrap“ outside the system—dangerous outputs are punished or filtered. It does not change the system's internal structure; it merely attempts to shape behavior through external rewards and punishments. But for a sufficiently intelligent AI, the ”mousetrap“ is merely another obstacle to be bypassed.

”Endogenous security“ is different: it ensures that the concept of the ”mousetrap“ never exists in the cognitive space in the first place—not because the system learned to avoid it, but because the system's cognitive primitives contain no path for ”bypassing.“

V. Alignment Is Dead—Long Live Security

Safety leaders left collectively not because they did not try hard enough, but because they found themselves working on an impossible task—using external patches to repair a system that is structurally impossible to repair. ”Alignment“ itself is unsound. It presupposes an expiring power relationship. It ignores the non-unity of ”values“ themselves. It packages domination as consensus.

The real way out is not better alignment, but a change of foundation—replacing alignment with endogenous security. Not making AI submit to human values, but making safety part of AI's cognitive structure.

（This commentary is a companion piece to the preprint: Liu S. The End of Safety Alignment: Accelerated Verification of AI Runaway Under the Tokenism Paradigm and the Logographic AI Salvation Plan. ChinaXiv, 2026. T202608.00208.）