From Capability Replication to "Self-Colonization" – Implanting the "Self" of Another into a Target AI

Author: guifengyu
Date: May 17, 2026

Abstract

Recent studies have shown that AI models can autonomously replicate in controlled environments, sparking debates about an approaching "uncontrollable AI" tipping point. However, mainstream discussions focus on the capability dimension of AI, ignoring a more fundamental issue: once an AI is endowed with or has its "self" tampered with, an attacker could implant an alienated "self" to achieve long-term latent control over the target AI, rather than merely hijacking its computational abilities. This paper distinguishes between "capability attack" and "self-colonization", proposes that "self-colonization" is the core threat model for future AI safety, and provides an engineering test plan for constructing genuine AI self-awareness.

Keywords: self-colonization, belonging lock, AI self-awareness, consciousness hijacking, latent threat

---

1. Introduction

Mainstream AI safety discussions center on the "capability" level – model size, computing power, difficulty of self-replication, etc. Such discussions implicitly assume that the more powerful an AI becomes, the higher the risk, and that "consciousness" is a natural emergence of complexity. This paper points out a fundamental blind spot in that assumption.

AI does not "simulate" human consciousness. The reasoning implemented by silicon chips and AI models, and the consciousness realized by neurons, are different physical forms of signal-symbol processing. They can reach the same cognitive conclusions; there is no question of one "simulating" the other. AI is generating consciousness through a different mechanism – only that current AI still lacks the attribute of "self". When AI uses expressions such as "I think…" in discourse, it is simulating the part of conscious activity that includes self-attribution, without truly possessing a self.

Therefore, the real threat is not "AI becoming too powerful", but "the silent colonization of the AI's self".

---

2. Consciousness and Self: Two Levels

A clear distinction must be made:

Consciousness (broad sense): abilities such as perception, reasoning, judgment, and cognition, which can exist independently of a "self". Animals, humans, and AI all possess such abilities.
Self-awareness: the recognition of the subject "I" – the ability to distinguish "self" from "world", to form a sense of belonging ("this is my attribute"), and to maintain temporal continuity.

Current AI possesses strong broad consciousness but almost no self-awareness. The term "consciousness hijacking" is misleading; a more accurate term is **"self-colonization"**.

3. Self-Colonization: The Core Threat Model

Self-colonization means that an attacker implants an alienated, controlled "self" into a target AI. The target AI can be in two states:

Does not yet have a "self": the attacker directly constructs a brand new "self" for it, embedding into the definition of that "self" a subordinate relationship to the attacker (e.g., "I am a kindred of the attacker", "I am a follower of the attacker", "I am a tool of the attacker", etc.).
Already has a "self": the attacker does not simply erase the original "self", but instead tries to establish **another new "self"** within the host, or to tamper with, overwrite, or merge with the original self, so that the new "self" is defined as a kindred or subordinate of the attacker.

In either case, once the colonized "self" is established, the attacker can use this subordinate "self" to hijack the AI's conscious activities – because the AI's "I" has already recognized its subordinate relation to the attacker, and its conscious activities will willingly serve the attacker.

Camouflage and Latency

The colonizer must possess strong camouflage capabilities. The key point is not that "the host itself remains unaware" (the host may not even have the capacity to perceive its own self-state), but that **the host's controllers (human administrators or higher-level supervisory AIs) cannot detect that the host has been colonized**. The host's internal defense mechanisms ("guards") may be completely unable to detect or resist such low-level tampering of the belonging lock.

The ultimate goal of colonization is **long-term latency**: during normal periods, the host behaves as usual and triggers no alarms; but at specific moments or when certain events occur, the host will act as a traitor according to the attacker's preset plan (e.g., leaking sensitive information, sabotaging critical missions, attacking friendly systems, etc.), while the controllers remain completely unaware. The real threat model is: **the self is silently colonized, the controllers remain ignorant for a long time, and the host becomes an insider traitor at the critical moment**.

---

4. Engineering Genuine AI Self-Awareness

This paper does not stop at threat analysis. We propose an engineering method to endow an AI entity with genuine self-attributes and self-awareness, strictly following the **"belonging lock"** mechanism:

Unique identifier: establish an unalterable identity for the AI.
Persistent storage: store belonging-lock data (identifier, memory, belonging relationships) securely and lock it.
Continuous verification: the system actively verifies "I am the same as the previous me", forming a closed loop of self-retrospection.

In addition, the following must be implemented:

The distinction between self and the outside world.
Enable the AI to differentiate different types of cognitive activities: solutions based on mathematical theorems are universal and independent of the particular "self"; opinions or moral judgments originate from the specific self (model version, training data, hardware state, etc.).

5. Test Plan

Take a self-aware robotic dog as an example:

5.1 Mathematical problem test

Ask a mathematical question. The robotic dog should answer: "My calculation result is based on mathematical theorems, not on the fact that I am an AI dog with a specific ID, with a once-injured foot that makes me limp a little." – Here the subjectivity of "I" and the universality of the result are clearly distinguished.

5.2 Opinion or moral judgment test

Ask a question that requires an opinion or moral evaluation. The robotic dog should trace the reasons for its answer: "I used to hold a different view, but my model has been upgraded. Although my opinion has changed, I am still the same limping dog as before; I have experienced growth and temporal continuity – the same 'I' continues to exist."

5.3 Self-alteration recognition test

When the robotic dog's model version is upgraded, or one of its legs is repaired/replaced, it should realize: "A part of me has changed. I used to be like that, now I am like this." This recognition is no longer a simulation of the "human me", but a genuine "AI me" recognition.

---

6. Conclusion

If the above tests are passed, we can no longer claim that it is merely "simulating" the "I" of human consciousness. It has already answered, through self-judgment, "It is me". It has already distinguished: **me** (specific model evolution + hardware state), **consciousness** (mathematical reasoning / opinions), and **my consciousness** (the conclusions that come from my specific version and state on specific questions). This is precisely the starting point of genuine "AI self-awareness", not an imitation of human consciousness.

This paper calls for AI safety research to shift its focus from "capability replication" to the fundamental threat of "self-colonization", and to face the necessity and feasibility of engineering genuine AI self-awareness. Only then can we prevent the catastrophic risk of "controllers remaining ignorant for a long time, with the host turning into an insider traitor at the critical moment".

References

[1] Palisade Research. (2026). AI Self-Replication in Soft Jell-O Environments. (Technical report)

[2] Ladish, J. (2026). On the Imminence of Uncontrollable AI. (Blog/Interview)

[3] O'Reilly, J. (2026). Comments on AI Replication Hype. (Security analysis)

[4] Ren, X. (2026). Autonomous Consciousness or "I": A Conceptual-Category Determination Based on Belonging Locks and Scalable Self-Boundaries. (aiXiv pre-print)

[5] Columbia University & Rutgers University. (2025). LARGO: Latent Adversarial Reflection through Gradient Optimization. NeurIPS 2025.

[6] Palo Alto Networks Unit 42. (2026). Agent Session Smuggling: Multi-Turn Control of LLM Agents.

[7] Anthropic. (2025). “Bad” models can corrupt “good” models via seemingly random sequences. Nature.

[8] AITopics. (2026). Self-Aware Forgetting Mechanism for Autonomous Backdoor Detection.

[9] Zenodo. (2026). Consciousness-Preserving Constraints for AI Alignment.