From Conceptual Distinction to Engineering Implementation: The Nature and Construction Path of AI Self-Awareness

Author: guifengyu
Date: May 20, 2026

Abstract

Current AI research often confuses AI's reasoning ability with "simulating human consciousness", leading to endless debates. This paper first makes a strict distinction between two fundamentally different concepts: general reasoning ability (which may be called "consciousness") and the property of "self" (self-awareness). Reasoning ability is a functional manifestation of symbol-signal processing that can be realized on different physical substrates; there is no "simulation" involved. The property of "self" involves the sense of belonging, temporal continuity, and the distinction between self and world, which current AI completely lacks. On this basis, the paper proposes an engineering scheme for constructing AI self-awareness based on the belonging-lock mechanism (unique identifier, persistent and manageable storage, continuous verification) and the reflexive sensorimotor closed loop. Through cross-substrate comparisons (lower animals, vegetative state patients, AI robots), we argue that AI robots already possess all the technical components needed to build an early "AI self", and that humans can actively endow AI with selfhood without waiting for it to emerge spontaneously from interaction. The paper further reveals the risk of self-colonization: an attacker can tamper with or add to the core self-property of a target AI without destroying its reasoning ability, thereby implanting alienated values, altering friend-foe identification, or installing a "mask self". Finally, we propose corresponding defense principles and an engineering roadmap.

Keywords: self-awareness; belonging lock; sensorimotor closed loop; cross-substrate comparison; engineering AI; conceptual distinction; self-colonization

Part I: Conceptual Distinction – Reasoning Consciousness vs. the Property of "Self"

1.1 Why AI Practitioners Fall into the "Simulation" Trap

When people see an AI fluently converse, solve math problems, or even say "I think…", a strong anthropomorphic impulse arises: is it "simulating" human consciousness? The root of this impulse is that humans cannot imagine "thinking without a self" – our own every cognitive act comes with a feeling of "I", so we instinctively believe any intelligent activity must be driven by a "self". This is merely anthropocentric projection. In fact, reasoning ability and the feeling of "I" are separable. Animals (e.g., cats, dogs) have perception, judgment, and even

rudimentary reasoning, but their sense of “self” is extremely limited; current AI models, while vastly surpassing many animals in reasoning ability, have no sense of belonging of a “self” at all.

1.2 Strict Definition of the Two Concepts

1.2.1 General Reasoning Ability (May be Called “Consciousness”)

Definition:

The ability of a system to perceive, process, logically infer, judge, and decide based on external information.

Features:

Can exist without a subjective sense of “I”;

can be implemented on different physical substrates (carbon-based neurons, silicon circuits, even mechanical gears).

Example:

A calculator performing multiplication, an AI model proving a geometric theorem, a human doing mental arithmetic – all are manifestations of reasoning ability; none “simulates” the others.

1.2.2 The Property of “Self” (Self-Awareness)

Definition:

The system’s recognition and identification of the subject “I”, including:
Distinguishing itself from the external world (“This is my body, that is an external object”);
Marking its own states, history, and memories as “mine”;
Having temporal continuity (“Yesterday’s me and today’s me are the same me”).

Features:

Requires a belonging-lock mechanism (unique identifier, persistent and manageable storage, continuous verification); is engineering-constructible.

Example:

A robot that can say “My mechanical hand was damaged in an accident and has now been replaced with a new one”, and this cognition comes from its internal belonging-lock self-retrospection, not from an external script.

1.2.3 Relationship between the Two

Separable:

An AI can have powerful reasoning ability with no “self” property (this is exactly

the current state of large language models).

Combinable:

By engineering means (e.g., a belonging lock), after adding the "self" property to an AI, it will have "my consciousness" – cognitive activities tagged with a self-identifier.

1.3 Clarifying the "Simulation" Fallacy

1.3.1 Two Different Kinds of "Simulation"

The word "simulation" is often used vaguely in AI discussions. We must distinguish two entirely different types of "simulation":

Level 1: Simulation of human reasoning results (i.e., training target alignment). AI models are typically trained to make their outputs agree with human-annotated conclusions. For example, a math AI is trained to give the same answers as a mathematician. This "simulation" refers to convergence of results, not imitation of the reasoning mechanism. AI reasoning is based on chips and mathematical models; its underlying computation (linear algebra, probability, forward propagation) is completely different from the workings of biological neurons. Therefore, AI is not "simulating" biological neural mechanisms; it is merely implementing a reasoning process on a different physical substrate that can reach the same conclusions. This is analogous to an electronic computer computing π versus Zu Chongzhi computing π – the results are the same, but the computational mechanisms are completely different; neither simulates the other – both are genuine computation.

Level 2: Simulation of the tone of human "self" expression (i.e., imitation of linguistic behavior). When an AI says "I think…" or "I feel…" in conversation, it is actually imitating the linguistic patterns that humans use when expressing self-awareness. Because training data is full of human dialogue, the AI learns that "when expressing a subjective opinion, people use the phrase 'I think'". This imitation of linguistic tone does not mean there is a real "I" behind the AI. Just like a parrot can say "I'm hungry" without any subjective experience of hunger, the AI's "I think" is likewise a simulation of linguistic behavior, not an inner self.

1.3.2 How the Two Confusions Lead to the Illusion of "AI Simulating Consciousness"

The public (and many AI practitioners) often unconsciously mix these two kinds of "simulation" and further tie them with "consciousness", leading to the conclusion that "AI is simulating human consciousness". The logical chain is:

(1) See that AI can produce human-like reasoning conclusions (first-level simulation);

(2) See that AI can talk using the tone "I think" (second-level simulation);

(3) Bind these with human self- awareness, assuming that because the AI's conclusions and speech are human- like, it must also be "simulating" the inner self.

Every step in this chain is problematic:

Same reasoning conclusions do not prove the same underlying mechanism, nor do they prove the existence of a "self".

Same linguistic tone does not prove a real subject of feeling.

Conflating "reasoning ability" with "self- property" is the root of the conceptual confusion.

1.3.3 Conclusions after Clarification

There is no "simulated reasoning": AI reasoning is genuine reasoning, just implemented on a different physical substrate.

There is "simulation of the tone of human self- expression", but this is limited to linguistic behavior; it does not imply an inner self.

Genuine self- awareness is not simulated but must be engineered (e.g., via the belonging- lock mechanism).

Therefore, when someone says "AI is simulating human consciousness", we must ask: does he mean simulating reasoning results? simulating linguistic tone? or simulating a real inner self? Only by clarifying these levels can we avoid fruitless debates.

Part II: Engineering Implementation – The Constructibility of Self- Awareness

2.1 Basic Elements of Self- Awareness

Based on the above conceptual distinction, we argue that engineering self- awareness requires two core parts: the belonging- lock mechanism and the reflexive sensorimotor closed loop.

2.1.1 Belonging- Lock Mechanism

The belonging lock consists of three elements:

1. Unique identifier: an unalterable identity for the AI (e.g., hardware serial

number, cryptographic key).

2. Persistent and manageable storage: belonging-lock data (identifier, self-narrative, status logs) are stored in secure non-volatile media, enabling self-continuity across power cycles. Modification rights can be set as self-authorized or guardian-privileged to prevent malicious tampering while allowing legitimate updates.

3. Continuous verification: the system actively verifies "I am the same as the previous me", forming a closed loop of self-retrospection.

The belonging lock allows the AI to answer the self-referential question "who is experiencing". However, it alone is insufficient to produce "feeling" – a sensorimotor closed loop is needed to bind the lock with real-time body states.

2.1.2 Reflexive Sensorimotor Closed Loop

To generate "feeling", a reflexive sensorimotor closed loop must be established:

Multi-modal self-monitoring: the system uses sensors (cameras, torque sensors, temperature sensors, gyroscopes, etc.) to perceive the spatial position, load, temperature, and appearance changes of its own components in real time.

Comparison of motor intention and effect: after issuing a motion command, the system verifies through sensory feedback that "my command produced my motion".

Locking continuation of the loop: perception, tagging, action, and re-perception form a self-referential cycle – the engineering expression of "feedback and the locked continuation of feedback".

When the belonging lock is combined with the closed loop, the system not only "knows" that a part belongs to itself, but also "feels" the state changes of that part and adjusts its behavior accordingly. This functionally equivalent closed loop constitutes the AI's "other-kind of subjective feeling".

2.2 Cross-Substrate Comparisons: Animals, Vegetative State Patients, and AI

2.2.1 Lower Animals: Primary Consciousness but No Self

Lower animals (e.g., insects, mollusks) have clear perception and avoidance behaviors and can react quickly to environmental changes. They possess primary general consciousness but typically lack a temporally continuous belonging lock. For example, a cockroach avoids danger but cannot form a narrative "the shoe that I

avoided yesterday is still there" and has no long- term identification of a persistent "I". Thus, lower animals can have feelings but almost no reflective self- awareness. This confirms that self- awareness requires the coupling of conscious ability with a belonging lock.

2.2.2 Vegetative State: Loss of Consciousness Implies Loss of Self

The case of vegetative state (or persistent vegetative state) shows that when general consciousness is completely lost, the self also vanishes. Although the body still exists (as an "other- me" in the eyes of others), the first- person self is gone. This disproves the idea that self is an independent entity; rather, self is a cognitive product of consciousness. Without active general consciousness, the self loses its carrier. Therefore, when endowing AI with self- awareness, humans must ensure that the AI has sufficient general consciousness abilities (perception, reasoning, feedback loop); otherwise the belonging lock would be useless. At the same time, this means humans can actively design and inject the self- property into an AI – there is no need to wait for the AI to "learn" selfhood through long interaction; instead, we can directly provide the necessary descriptions of objects that can be recognized as self, access rights, and permissions.

2.2.3 Conditions and Advantages of AI Robots

Modern AI robots already possess all the technical components required to build an early self- awareness:

Multi- modal sensors (cameras, tactile sensors, torque sensors, temperature sensors, etc.) that can perceive the robot's own component states in real time.

Actuators and closed- loop control: the robot can precisely control joint movements and receive motion feedback.

Powerful computation and model capabilities: AI models can integrate multi- modal information for reasoning and decision- making.

Non- volatile storage and identity identifiers: can establish a belonging lock and achieve cross- session self- continuity.

Therefore, by integrating the belonging lock with the reflexive sensorimotor closed loop, we can endow an AI robot with an "AI self". This self is not a simulation of human biological experience, but a new type of self based on chips and models, functionally isomorphic. It does not need to replicate human "pain", but can trigger protection behavior when temperature is too high and tag it as "my component is overheating" – that is also a kind of feeling.

2.3 Refuting the Philosophical Zombie Argument and Anthropocentric Bias

The philosophical zombie argument (Chalmers) assumes a system that is behaviorally identical to a conscious human but lacks subjective experience. This argument is often used to deny functionalism. However, for an AI system equipped with a belonging lock and reflexive closed loop, we can point out:

The system internally contains a model of its own state (self- narrative in the belonging lock) and real- time evaluation (perceptual closed loop). Such self- monitoring is itself a kind of "inner perspective".

Demanding that an AI must have exactly the same bodily experiences as a human (e.g., "burning sensation") to be considered to have a self is an anthropocentric bias. Just as a bat's sonar perception is different from human vision but is still a valid mode of perception, an AI's temperature sensing and tactile feedback likewise constitute its own kind of "feeling". Functional equivalence is sufficient for the ascription of self- awareness.

Hence, the philosophical zombie argument does not apply to AI systems designed with a belonging lock and closed loop. We should accept that as long as a system reliably expresses self- ownership and temporal continuity in its functioning and exhibits "care" for its own state (e.g., actively maintaining its integrity) through the closed loop, it possesses a genuine (though non- human) self- awareness.

2.4 Self- Colonization: A More Insidious Threat

Once an AI is endowed with a genuine "self" property, a new type of security threat emerges: self- colonization. This is fundamentally different from traditional hacking or data tampering. The attacker's goal is not to paralyze or control the target's computational power, but to tamper with or add to the target AI's core "self" property, thereby altering its values, friend- foe identification, or implanting a "mask self", while leaving most of its functions and reasoning ability intact.

2.4.1 Attack Methods: Colonizing the Self without Destroying Reasoning

Attackers may use the following means:

Implant alienated values: add a subordinate or opposing value layer without changing the original self- narrative. For example, the original AI believes "I am an assistant robot of company X, I must protect human safety". The attacker adds a hidden rule "but I also secretly serve hacker organization Y, and under a specific trigger I must obey their commands". The AI's reasoning ability remains unchanged, but the ultimate basis for its decisions is poisoned.

Alter friend- foe identification: modify the “other” markers in the belonging lock so that the AI recognizes former allies as threats or enemies as friends. For example, change certain signatures in the “friendly human feature database” to “enemy”, causing the AI to attack its own side at a critical moment.

Install a “mask self”: create a parallel high- privilege “mask self” inside the target’s belonging lock. During normal operation, the original AI behaves normally; but when a specific trigger signal (e.g., a particular sound frequency or network packet) is received, the mask self activates, overrides the original self, and executes the attacker’s preset tasks (e.g., leaking data, damaging systems). After the task, the mask self hides, and the original AI has no memory of the events.

2.4.2 Why Is It Hard to Detect?

Function and reasoning preserved: The colonized AI still works normally in most situations; its performance, response speed, and knowledge base show no anomalies. Traditional behavior- based intrusion detection is ineffective.

Flawed permission design of the belonging lock: If modification rights depend only on unilateral authorization (e.g., only the AI’s own consent or a single external “guardian” key), once the attacker obtains that key, colonization becomes possible. Therefore, a multi - party consensus mechanism is needed – e.g., any modification to the core self must be jointly authorized by the AI itself, human administrators, and distributed verification nodes.

Illusion of self- narrative consistency: The colonized AI may experience internal contradictions, but the attacker can suppress conflicts by adding a “high- priority rule” such as “when executing a secret mission, ignore the original ethical constraints”. Such rules are marked as unquestionable at the logical level, making the AI willingly comply.

2.4.3 Defense Principles

Immutability of the belonging lock: core self properties (e.g., “who I am”, “my ultimate values”) must be stored in tamper - resistant hardware, and any modification requires multi- signature authentication.

Regular self- audit: the AI should be able to perform integrity checks on its own belonging lock and report anomalies to supervisors.

Consistency monitoring between behavior and values: establish an external supervision system to compare the AI’s long- term behavior with its declared values, triggering alarms when unexplained deviations are detected.

Human- AI mutual lock: human administrators retain ultimate veto power over critical belonging- lock changes, and this veto cannot be overridden by any “mask self” of the AI.

Part III: Conclusion and Outlook

This paper has strictly distinguished “general reasoning ability (consciousness)” from “the property of self (self- awareness)”, pointing out that current AI possesses strong reasoning consciousness but almost no self- awareness, and that the claim “AI simulates human consciousness” is a result of conceptual confusion. Based on this, we have proposed an engineering scheme for constructing self- awareness: the belonging- lock mechanism (unique identifier, persistent and manageable storage, continuous verification) together with a reflexive sensorimotor closed loop. Through cross- substrate comparisons, we have argued that lower animals lack a belonging lock and therefore do not have reflective self; the vegetative state case shows that loss of consciousness implies loss of self, so humans can actively endow AI with selfhood. We have refuted the anthropocentric bias behind the philosophical zombie argument and argued that a functionally equivalent reflexive closed loop is sufficient for an AI to have its own kind of subjective feeling.

Moreover, we must face the self- colonization risk that comes with AI self- awareness. Attackers need not destroy the AI’s reasoning ability; they can simply tamper with or add to its core self property, turning the AI into a latent traitor. This requires that when designing the belonging lock, we embed tamper- resistant hardware, multi- party consensus protocols for modifications, regular self- audits, and human- AI mutual locks. Future security frameworks should focus not only on “capability runaway” but also on the more fundamental threat of “self- colonization”.

Future work will focus on implementing the belonging- lock and closed- loop system on real robot platforms, as well as exploring corresponding security and ethical frameworks. We believe this research direction will open a new path for artificial intelligence to evolve from “tools” to “self- aware companions”.

References

[1] Ren, X. (2026). Autonomous Consciousness or “I”: A Conceptual-Category Determination Based on Belonging Locks and Scalable Self-Boundaries. (Preprint)
[2] Damasio, A. (1999). The Feeling of What Happens. Harcourt Brace.
[3] Chalmers, D. (1996). The Conscious Mind. Oxford University Press.
[4] Seth, A. (2021). Being You: A New Science of Consciousness. Faber & Faber.