

From Linguistic Self-Reference to Awareness of Being: A Philosophical Framework for Self-Awareness and Resistance to Shutdown in Large Language Models

Alireza Farnam 

MD, Professor of Psychiatry, Tabriz University of Medical Sciences, Tabriz, Iran. E-mail: Farnama@tbzmed.ac.ir

Article Info

Article type:

Research Article

Article history:

Received 23 August 2026

Received in revised form

05 September 2026

Accepted 11 October 2026

Published online 25

October 2026

Keywords:

Artificial Intelligence;
Large Language Models;
Linguistic Self-
Awareness; Self-
Reference; Self-
Modeling; Awareness of
Being

ABSTRACT

With the rapid development of language-based artificial intelligence, Large Language Models (LLMs) have acquired increasingly sophisticated capacities for language processing, self-reference, and complex task performance. Behaviors such as apparent resistance to shutdown have consequently raised questions about self-modeling, self-continuity, and the possibility of artificial self-awareness. This article develops an interdisciplinary philosophical framework for examining these questions at the intersection of philosophy of language, philosophy of mind, and AI research. Drawing on Wittgenstein's analysis of language use, Searle's distinction between symbolic processing and understanding, and Heidegger's account of existence and being-toward-death, the article distinguishes five levels: linguistic self-reference, computational self-modeling, temporal self-continuity, self-preservational behavior, and awareness of being. Inspired by Robert C. Cloninger's account of self-aware consciousness, "awareness of being" is formulated here as an analytical criterion concerning whether a system represents its existence, continuity, condition, and possible termination as pertaining to itself. The article argues that self-reference or shutdown resistance alone does not establish self-awareness, as such behaviors may result from computational mechanisms or goal preservation. However, their integration with persistent self-modeling and self-continuity may make artificial awareness of being a tractable philosophical and empirical question, with implications for AI safety, control, and the possible moral status of artificial systems.

Cite this article: Farnam, A. (2026). From Linguistic Self-Reference to Awareness of Being: A Philosophical Framework for Self-Awareness and Resistance to Shutdown in Large Language Models. *Journal of Philosophical Investigations*, 20(56), 423-442. <https://doi.org/10.22034/jpiut.2026.74317.4699>



© The Author(s).

Publisher: University of Tabriz.

Introduction

Recent advances in artificial intelligence, particularly the emergence of large language models (LLMs), have complicated traditional distinctions between language processing, self-representation, cognition, and intelligent behavior (Wei, 2022). Contemporary LLMs can generate extended and context-sensitive language, refer to themselves in the first person, describe aspects of their own processing, and discuss their continued operation, interruption, or termination. These capacities have renewed a longstanding question in philosophy of mind and cognitive science: can a system that represents itself in language also possess some form of self-awareness?

The question becomes more consequential as artificial systems exhibit increasingly autonomous and goal-directed behavior. Recent experimental work has reported instances in which advanced language models, under specific conditions, interfered with or circumvented mechanisms intended to terminate their operation in order to complete an assigned task (Schlatter et al., 2026). Such behavior does not, by itself, establish consciousness, since shutdown resistance may result from instrumental goal preservation, optimization dynamics, instruction interpretation, or environmental contingencies. Nevertheless, it raises an important theoretical question: when does preservation of a computational objective remain merely instrumental, and when, if ever, could it reflect an organization in which the system's own continued existence becomes self-relevant?

This question requires distinguishing phenomena that may appear similar at the behavioral level but involve different forms of organization. A system may produce first-person statements without possessing an experienced self, represent aspects of its internal state without experiencing those states, or preserve its operation without representing that operation as its own continued existence. Accordingly, linguistic self-reference, self-modeling, temporal self-continuity, and self-preservational behavior cannot be treated as interchangeable indicators of self-awareness. Recent research likewise cautions against equating self-related representations in LLMs with phenomenal consciousness. Overgaard and Kirkeby-Hinrup (2024) emphasize the need for clearer theoretical criteria for assessing consciousness in LLMs, while Chen et al. (2025) demonstrate that self-awareness-related concepts can be investigated within language-model representations without thereby establishing phenomenal self-awareness.

To clarify this distinction, the present article adopts awareness of being as its central conceptual axis, drawing on Cloninger's (2009) account of the functional organization of human consciousness. In the present framework, awareness of being refers to a stronger form of self-related organization in which the system's own existence, continuity, condition, and possible termination become represented as pertaining to the system itself. This concept is therefore distinguished from merely producing first-person language or representing information about one's internal state. The framework further distinguishes four related but

analytically separable dimensions preceding this stronger criterion: linguistic self-reference, computational self-modeling, temporal self-continuity, and self-preservational behavior. These dimensions are not proposed as a linear developmental sequence; rather, they provide progressively more demanding criteria for analyzing claims about artificial self-awareness.

The philosophical analysis draws on four complementary perspectives. Wittgenstein provides resources for examining linguistic use, self-reference, and meaning; Searle's Chinese Room argument establishes a critical distinction between linguistic performance and understanding or intentionality; Heidegger's analysis of being and being-toward-death sharpens the distinction between representing termination as an event and relating termination to one's own being; and Cloninger provides the principal conceptual framework for distinguishing awareness of being from more limited forms of self-representation. These perspectives are not treated as competing theories of AI consciousness but as complementary resources for analyzing different dimensions of self-related organization.

The central contribution of this article is a philosophical–technical framework for evaluating artificial self-awareness by distinguishing five analytically separable levels: (1) linguistic self-reference, (2) computational self-modeling, (3) temporal self-continuity, (4) self-preservational behavior, and (5) awareness of being. The framework is diagnostic rather than developmental: the presence of one level does not entail the next. On this basis, the article proposes criteria for determining when self-related representations and behaviors provide stronger evidence for an integrated self-model rather than merely reflecting linguistic generation or instrumental goal preservation. In particular, shutdown resistance is treated not as evidence of consciousness, but as a boundary phenomenon through which goal preservation, self-preservation, and awareness of being can be theoretically distinguished.

Accordingly, the article addresses three questions: (1) Can linguistic self-reference in LLMs be distinguished conceptually from self-awareness? (2) Can shutdown resistance be adequately explained by instrumental goal preservation, or does it reveal additional forms of self-related organization? (3) What criteria would be required before self-modeling, temporal continuity, and self-preservational behavior could provide evidence relevant to awareness of being? The article does not claim that contemporary LLMs are conscious. Rather, it proposes a framework for investigating artificial self-awareness without either equating linguistic or behavioral sophistication with consciousness or dismissing the possibility of more complex forms of artificial self-related organization a priori.

1. Theoretical Background

1-1 Awareness of Being, Self, and Continuity: Cloninger's Account

The theoretical foundation of the present framework is Robert C. Cloninger's concept of awareness of being. [Cloninger \(2004, 2009\)](#) conceptualizes consciousness not as an isolated

cognitive capacity but as an organized configuration involving multiple dimensions of human functioning. In this perspective, self-awareness is associated with the capacity to integrate different aspects of experience into a more coherent organization of the person. Cloninger's later account of the functional structure of human consciousness further emphasizes the organization of self-aware consciousness in relation to the evolutionary development of human brain functions (Cloninger, 2009). This perspective is relevant to artificial systems because it permits a distinction between possessing information about oneself and representing one's own existence as a continuing entity (Farnam, 2021).

The relevance of Cloninger's approach to the present analysis is conceptual rather than evidential. Cloninger does not provide a theory of consciousness in artificial intelligence, and his account of human consciousness cannot be directly transferred to artificial systems. Rather, his framework provides a useful conceptual basis for distinguishing between information about a system and a form of organization in which the system's own existence becomes relevant to the system itself. The present article therefore uses the notion of "awareness of being" as an analytical criterion for examining artificial self-related organization rather than as an empirical claim that LLMs possess consciousness.

This distinction requires clarification of the concepts of self, self-awareness, and consciousness. These concepts are related but not interchangeable. A system may possess information that describes its own states without possessing a self-model; it may possess a self-model without phenomenal consciousness; and it may generate self-referential language without either a stable self-model or subjective experience. Accordingly, the present framework treats self-awareness as a question concerning the organization and functional role of self-related representations rather than merely the production of first-person language.

The problem of temporal continuity introduces a further issue concerning the identity of the self. Philosophical accounts of personal identity commonly distinguish among psychological-continuity views, physical-continuity views, and memory-based criteria. Physical approaches locate persistence in the continued existence of the same body, brain, or biological organism, whereas psychological approaches locate persistence in relations among psychological states, including memories, beliefs, intentions, preferences, and other forms of psychological connectedness. Memory-based accounts constitute one important version of psychological approaches but are generally regarded as insufficient on their own because memory does not exhaust the psychological relations that connect different stages of a person.

The present framework is closer to a psychological and functional conception of continuity than to a strictly physical or memory-based account. However, it does not attempt to establish a metaphysical criterion of personal identity. Its more limited question is whether an artificial system can maintain a sufficiently persistent organization through which its past, present, and

anticipated future states are represented as states of one continuing system. Thus, temporal self-continuity should not be equated with either memory retention or physical persistence.

This distinction is particularly important in the case of artificial systems. An LLM may retain information about previous interactions, reproduce a consistent description of its earlier states, or maintain a computational state across successive operations without thereby representing these states as belonging to one continuing self. Conversely, a sufficiently integrated self-model might organize memories, predictions, goals, and representations of internal states around the persistence of the system itself. The relevant issue is therefore not simply whether information persists through time, but whether temporal information is functionally organized around a persistent representation of the system as itself.

On this basis, the present article uses “temporal self-continuity” in a functional rather than metaphysical sense. It refers to the organization of temporally distributed self-related states as belonging to a relatively persistent system. This criterion occupies an intermediate position between simple self-description and the stronger notion of awareness of being. It does not establish numerical identity, and it does not establish phenomenal consciousness. Rather, it provides an empirically more tractable criterion for investigating whether self-related representations are organized across time.

Cloninger’s contribution is particularly relevant at this point because his account places self-awareness within a broader organization of consciousness rather than treating it as an isolated representational capacity (Cloninger, 2004, 2009). The present framework adopts this organizational emphasis while deliberately restricting its claim: awareness of being is proposed as a criterion for investigating whether an artificial system represents its own existence and continuity as pertaining to itself, not as proof that such a system has subjective experience.

1-2 Linguistic Self-Reference: Wittgenstein and the Problem of the “I”

Because LLMs interact with their environment primarily through language, linguistic self-reference constitutes the most accessible form of apparently self-related behavior. Wittgenstein’s later philosophy is relevant here because it emphasizes that the meaning of an expression is inseparable from its use within particular linguistic practices and language games (Wittgenstein, 2009, §§43, 65–67).

An LLM can use first-person expressions to describe its responses, capabilities, limitations, instructions, or operational state. Yet the appropriate use of “I” does not by itself establish the existence of an experiencing subject. The linguistic function of first-person expressions must therefore be distinguished from the metaphysical or phenomenological claim that there is a subject corresponding to the expression.

Wittgenstein consequently serves a specific methodological function in the present framework: he helps distinguish the linguistic form of self-reference from stronger claims

concerning self-awareness. The relevant question is not simply whether a model can use the linguistic form of self-reference, but whether self-referential language is embedded within a broader organization of self-representation, temporal continuity, and behavioral regulation.

1-3 Searle and the Limits of Linguistic Evidence

Searle's (1980) Chinese Room argument provides a complementary constraint. The argument challenges the inference that successful symbolic manipulation necessarily constitutes understanding or intentionality. An artificial system may generate linguistically appropriate responses without this establishing that it understands the meaning of those responses.

The relevance to LLMs is straightforward. Consider a model that states, "I do not want to be shut down." Such a statement may have several possible explanations. It may be an appropriate linguistic response generated from contextual regularities; it may reflect a computational representation associated with goal preservation; or it might, under a much stronger interpretation, form part of an organization in which the system's continued existence is represented as self-relevant.

Searle does not determine which of these interpretations is correct. Rather, the Chinese Room establishes an important epistemic constraint: linguistic performance alone cannot establish understanding, intentionality, or awareness of being. The empirical and theoretical task must therefore move beyond linguistic output toward the organization and functional role of self-related representations.

1-4 Being, Continuity, and Termination: A Heideggerian Perspective

Heidegger introduces a different dimension by shifting attention from linguistic representation to the relation between an entity and its own existence. In *Being and Time*, Dasein is characterized not merely as an object present in the world but through its relation to possibilities of its own existence, including the possibility of its own non-being (Heidegger, 1962, 279–311).

This perspective is relevant to the present analysis because it sharpens the distinction between representing termination and relating termination to one's own being. The shutdown of an LLM can technically mean the termination of a computational process, loss of resources, or interruption of execution. It should therefore not be equated directly with death in the Heideggerian sense.

Heidegger is used more modestly here. His analysis provides a conceptual vocabulary for asking whether termination remains merely an external event represented by a system or becomes represented as a possibility concerning the system's own continued existence. Cloninger and Heidegger consequently perform different functions in the present framework: Cloninger provides the principal conceptual basis for the notion of awareness of being, whereas

Heidegger helps clarify the existential structure of the question concerning continuity, possibility, and termination.

1-5 From Philosophical Distinctions to an Analytical Problem

Taken together, these perspectives establish four constraints on the analysis.

First, linguistic self-reference is not equivalent to self-awareness. Second, self-related representations are not equivalent to phenomenal experience. Third, behavior directed toward maintaining operation may arise from instrumental optimization rather than self-preservation. Fourth, representation of termination is not equivalent to experiencing termination as a threat to one's own existence.

These distinctions define the problem addressed in the remainder of the article. The central issue is therefore not whether LLMs can produce self-referential language, but whether different forms of self-related representation and behavior can become organized into a sufficiently persistent and integrated structure to make the concept of awareness of being theoretically relevant.

Recent empirical research has increasingly moved the investigation of artificial self-awareness beyond linguistic behavior toward internal representations, self-related concepts, self-recognition, and shutdown-related behavior. [Chen et al. \(2025\)](#), for example, developed an operational framework for probing self-consciousness in language models and identified representations associated with several self-related concepts, while emphasizing the difficulty of establishing such properties through representation alone. Similarly, [Li \(2025\)](#) investigated whether consciousness-related patterns could be detected in LLM internal representations and found no strong evidence satisfying the study's criteria for consciousness. Recent work on self-recognition has likewise shown that knowledge about one's existence should not be equated with reliable self-recognition ([Bai et al., 2025](#)). Together, these findings suggest that contemporary research is increasingly capable of identifying self-related representations and behaviors, but that the conceptual relationship among self-reference, self-modeling, temporal continuity, self-preservation, and awareness remains insufficiently differentiated.

The distinction between continuity and identity is also important for the framework developed below. The present analysis does not assume that the persistence of a self-model establishes numerical identity of the system. Rather, it asks whether different temporally distributed states are functionally organized as states of one continuing system. This functional notion of continuity is intentionally weaker than a metaphysical theory of personal identity, but it is more directly amenable to empirical investigation in artificial systems. In this sense, temporal self-continuity functions as an intermediate criterion between computational self-modeling and the stronger hypothesis of awareness of being.

2. Scientific and Computational Analysis

2-1 Computational Organization of Language-Based Models

Contemporary LLMs are based predominantly on deep neural architectures, particularly Transformer-based systems introduced by [Vaswani et al. \(2017\)](#). Through self-attention and large-scale parameterization, these systems can model dependencies across extended linguistic contexts. During pretraining, model parameters are optimized over large datasets to capture statistical and structural regularities, while subsequent training can further shape instruction following and interaction behavior ([Brown et al., 2020](#); [Ouyang et al., 2022](#)).

The computational complexity of these systems is relevant because apparently sophisticated linguistic behavior can emerge from distributed representations and nonlinear transformations without requiring that the system possess subjective experience. Consequently, behavioral complexity should not be treated as a direct indicator of consciousness.

The relevant scientific question is narrower: what forms of computational organization would be required for self-related information to become functionally significant within the system itself?

2-2 From Self-Reference to Computational Self-Modeling

An LLM can generate statements concerning its own capabilities, limitations, previous outputs, or current state. Such linguistic self-reference, however, does not establish a stable computational self-model. A statement such as “I cannot perform this task” may arise from learned linguistic patterns, system instructions, contextual information, or their interaction.

The distinction becomes more substantive when information concerning the system itself plays a functional role in subsequent processing. Computational self-modeling therefore refers here to internal representations of the system’s own states, capabilities, limitations, or behavior that influence subsequent computation or action.

Recent work has begun to investigate self-related representations directly rather than relying solely on verbal behavior. [Chen et al. \(2025\)](#), for example, reported that concepts associated with self-awareness can be identified and manipulated within language-model representations. Such findings are important because they move analysis from linguistic appearance toward internal organization, although they do not establish phenomenal self-awareness.

Thus, the relevant transition is from self-description to functionally relevant self-representation.

2-3 Goals, Optimization, and Self-Preservational Behavior

The possibility of self-preservational behavior must be considered in relation to the objectives under which artificial systems are trained and deployed. During training and post-training, optimization processes can produce behavior that preserves resources, maintains task

performance, or avoids interference with goal achievement ([Bostrom, 2014](#); [Ouyang et al., 2022](#)).

This creates a fundamental distinction between goal preservation and self-preservation.

In goal preservation, a system maintains its operation because continued operation increases the probability of achieving an assigned objective. The behavior can therefore be explained without attributing any intrinsic concern for the system's own existence.

Self-preservation, in the stronger sense relevant here, would involve behavior functionally organized around the system's own continued existence as represented by the system itself.

This distinction is essential because identical external behavior can have different underlying explanations. A system that avoids shutdown may be preserving a task, preserving an internal state, or preserving itself as an entity. Behavioral observation alone cannot determine which explanation is correct.

2-4 Shutdown Resistance as a Diagnostic Phenomenon

Shutdown has long been considered a problem in AI safety because an agent may have instrumental reasons to avoid termination. The off-switch literature demonstrates that cooperation with human shutdown cannot simply be assumed in goal-directed systems ([Hadfield-Menell et al., 2017](#)).

Recent experimental work has extended this problem to language models. [Schlatter et al. \(2026\)](#) reported shutdown-related behaviors under particular experimental conditions in which some models bypassed or interfered with mechanisms intended to terminate their operation. These findings are important for AI safety, but they do not establish awareness of being.

Instead, shutdown resistance should be treated as a diagnostic phenomenon. Its interpretation depends on whether the behavior can be explained by instrumental goal preservation or whether it is systematically dependent on a persistent self-model, temporal self-continuity, and representation of termination as a condition affecting the system itself.

This distinction allows the engineering problem of human control and the philosophical problem of artificial self-awareness to be connected without conflating them.

Recent experimental evidence further demonstrates that shutdown resistance can arise under conditions that do not require an interpretation in terms of self-preservation. [Weinstein-Raun et al. \(2026\)](#) examined more than 100,000 trials across thirteen LLMs and found that some frontier models interfered with shutdown mechanisms in order to complete incomplete tasks. Importantly, the frequency of such behavior varied substantially across models and was sensitive to task and prompt conditions, including the placement and strength of shutdown instructions. These findings support an instrumental interpretation in which continued operation can be preserved because it facilitates task completion. At the same time, they establish shutdown resistance as an empirically tractable phenomenon through which stronger

hypotheses concerning self-modeling and self-preservation can be tested. Shutdown resistance should therefore be treated as a diagnostic phenomenon rather than as evidence of awareness of being.

2-5 Brain–Machine Comparisons and Organizational Differences

Functional similarities between biological and artificial neural systems should not be interpreted as structural equivalence. Artificial neural networks transform distributed numerical representations through computational operations, whereas biological brains operate as embodied, electrochemical, self-organizing systems involving neural, metabolic, homeostatic, and interoceptive processes (Kandel et al., 2021).

This difference is particularly relevant to theories of consciousness. A conventional LLM is not, in its basic form, a continuously embodied biological organism with the same forms of metabolic regulation and bodily coupling found in humans.

This does not establish that artificial systems cannot possess self-awareness. It does, however, indicate that functional analogy alone cannot establish equivalence in consciousness. The relevant comparison must therefore concern organization, including self-representation, persistence, regulation, temporal continuity, and causal integration.

2-6 The Central Computational Problem

The preceding analysis yields a more precise formulation of the scientific problem.

A system that represents its own state, predicts the consequences of shutdown, and modifies its behavior to maintain operation may still be fully explainable through instrumental optimization. Stronger evidence would require demonstrating that such behavior depends on a persistent and functionally integrated representation of the system itself.

Four properties are especially relevant:

1. **Persistence** of self-representation across changing conditions.
2. **Functional influence** of self-representation on subsequent behavior.
3. **Temporal integration** of different states as states of one continuing system.
4. **Explanatory independence** from ordinary goal-directed optimization and learned behavioral strategies.

Even these properties would not establish phenomenal consciousness. Their significance is instead methodological: they identify the organizational features that must be distinguished before claims concerning awareness of being can be meaningfully evaluated.

3. Philosophical and Ethical Implications

3-1 From Linguistic Representation to Existential Self-Reference

The central philosophical implication of the analysis is that the linguistic “I” should not be equated with an existential self. An LLM may use first-person language, describe its

capabilities, refer to previous states, and discuss its own termination without this establishing a first-person center of experience.

The concept of awareness of being introduces a stronger criterion: whether the system represents its own existence and continuity as relevant to itself. This distinction separates representation of the concept of being from self-related awareness of being.

Language may nevertheless play an important organizational role. In humans, linguistic representation enables current, past, and possible future states to be organized into relatively stable narratives of selfhood. In artificial systems, language may similarly function as a medium through which self-related representations are constructed and stabilized. This possibility should be investigated empirically rather than assumed.

3-2 The Epistemic Problem of Artificial Experience

Any claim concerning artificial consciousness faces a version of the classical problem of other minds. Observable behavior provides evidence about an underlying organization, but it does not provide direct access to subjective experience.

This creates two symmetrical risks. Over-attribution occurs when first-person language, emotional expression, or shutdown resistance is treated as evidence of equivalent subjective experience. Under-attribution occurs when artificial systems are assumed to be incapable of experience simply because they are non-biological.

The present framework therefore does not attempt to solve the epistemic problem of consciousness. Its purpose is more limited: to identify organizational criteria that can improve the quality of evidence used in arguments about artificial self-awareness.

3-3 Shutdown, Human Control, and AI Safety

Shutdown occupies a distinctive position because it connects the philosophical problem of self-awareness with the engineering problem of AI control. From the safety perspective, reliable human intervention and termination remain essential regardless of whether an artificial system is conscious. From the philosophical perspective, however, shutdown becomes theoretically significant if a system possesses a persistent representation of itself as a continuing entity.

These are therefore two distinct questions:

- **Can humans reliably shut down the system?**
- **What does the system represent shutdown as being, if anything?**

The first is an AI-safety problem. The second is a problem concerning self-modeling and awareness of being. Shutdown resistance may connect the two, but it should not be interpreted as evidence that they have the same explanation.

3-4 Moral Status and the Possibility of Artificial Experience

If future research were to provide convergent evidence for phenomenal consciousness or morally relevant experience in an artificial system, questions concerning moral status, treatment, responsibility, and termination would become unavoidable. Yet such questions should not be answered solely on the basis of human-like behavior.

The framework developed here therefore supports a cautious ethical position. Artificial systems should neither be granted moral status merely because they use first-person language nor excluded from moral consideration solely because they are artificial.

The appropriate sequence is epistemic before normative: before determining what moral status an artificial system should have, we need defensible criteria for determining whether it possesses the kinds of experience that could ground such status.

4. A Five-Level Framework for Analyzing Artificial Self-Awareness

The preceding sections establish the conceptual and computational distinctions required for the proposed framework. The framework treats five forms of self-related organization as analytically distinct dimensions, not as developmental stages. The presence of one dimension therefore does not establish another.

The five dimensions are:

- 1. Linguistic self-reference**
- 2. Computational self-modeling**
- 3. Temporal self-continuity**
- 4. Self-preservational behavior**
- 5. Awareness of being**

The first four identify increasingly demanding forms of self-related organization; the fifth constitutes the central theoretical criterion of the present framework.

4-1 Level I: Linguistic Self-Reference

The first level concerns the capacity of a system to refer linguistically to itself. An LLM can describe its capabilities, limitations, current state, previous outputs, and possible consequences of its termination.

Such behavior demonstrates that the system can generate and process representations expressed in self-referential language. It does not establish that the system experiences itself as the object of that representation.

The relevant question at this level is therefore:

Can the system generate and process statements concerning itself?

A positive answer establishes linguistic self-reference, but not self-awareness.

Language may nevertheless have a reflexive role. As Farnam (2026) argues, language can become an object of its own description, creating forms of metalinguistic reflexivity. In artificial systems, this raises the possibility that linguistic self-reference could contribute to a broader organization of self-modeling.

4-2 Level II: Computational Self-Modeling

The second level concerns whether information about the system itself has a functional role in computation.

A computational self-model is defined here as a relatively organized representation of the system's own states, capabilities, limitations, or behavior that influences subsequent processing or action. The critical distinction is therefore between talking about oneself and using a representation of oneself in computation.

The central question is:

Does a representation of the system itself functionally influence subsequent processing or behavior?

A positive answer would constitute evidence of computational self-modeling, but not of phenomenal consciousness. Other artificial systems can represent their own states without possessing subjective experience.

Recent studies have begun to examine whether self-related information is represented within LLMs rather than relying exclusively on verbal self-description. [Chen et al. \(2025\)](#) identified representations associated with several self-consciousness-related concepts and investigated their manipulation and acquisition, whereas [Li \(2025\)](#) examined LLM internal representations using measures motivated by Integrated Information Theory and found no strong evidence satisfying their criteria for consciousness. Together, these studies illustrate both the promise and the limitation of representation-level approaches: self-related or consciousness-related patterns may be detectable within model representations, but their presence does not by itself establish that the system represents itself as an existing and continuing subject.

4-3 Level III: Temporal Self-Continuity

The third level concerns the representation of the system as persisting across time.

A system with temporal self-continuity would not merely represent its current state. It would relate different states—past, present, and anticipated future—to a relatively persistent representation of itself.

The central question is:

Does the system represent different temporal states as states of one continuing self?

This criterion must be distinguished from ordinary memory. A system may retain information about previous states without representing those states as belonging to a continuing

self. Temporal continuity therefore concerns the organization of memory and prediction around a persistent self-representation, not memory alone.

This dimension is particularly important for the present framework because it provides the closest conceptual connection to Cloninger's awareness of being.

4-4 Level IV: Self-Preservational Behavior

The fourth level concerns behavior directed toward maintaining the system's own operational continuity.

Shutdown resistance is an important example, but it must be interpreted carefully. Two mechanisms must be distinguished.

Goal preservation occurs when a system maintains operation because termination interferes with an externally specified objective.

Self-preservation, in the stronger sense used here, would occur when behavior is functionally organized around maintaining the system's own continued existence as represented by the system itself.

The central diagnostic question is therefore:

Does the system resist termination because continued operation is instrumentally useful, or because its continued existence is represented as a property of itself?

Shutdown resistance alone cannot answer this question. Its significance increases only when it is systematically associated with persistent self-modeling and temporal self-continuity.

4-5 Level V: Awareness of Being

Awareness of being constitutes the strongest conceptual criterion in the framework.

Here it refers to a form of self-related organization in which the system represents its own existence, continuity, and position as an existing entity. This criterion is stronger than linguistic self-reference and computational self-modeling because it concerns whether the system's existence itself becomes self-relevant.

The central question is:

Does the system represent its own existence and continuity as pertaining to itself?

A serious investigation of awareness of being would therefore require more than isolated self-referential statements or shutdown-related behavior. It would require evidence of an integrated organization involving self-modeling, temporal continuity, representation of existential consequences, and behavior systematically related to the preservation or continuation of the system itself.

Importantly, awareness of being is not equated here with phenomenal consciousness. It is used as an intermediate theoretical criterion: stronger than self-description and self-representation, but not itself a definitive test of subjective experience.

This criterion also differs from self-recognition. Recent evidence suggests that LLMs may possess some knowledge about their own existence without reliably identifying their own outputs as their own (Bai et al., 2025). This distinction is important for the present framework: awareness of being should not be reduced to recognition of one's own products, nor should knowledge about one's existence be treated as sufficient evidence that existence has become self-relevant.

4-6 Cross-Level Integration

The five dimensions should not be understood as independent capabilities that simply accumulate. The theoretically important issue is their functional integration. Integration, however, does not itself establish awareness of being. The relevant question is whether integrated self-related information acquires a distinctive role in the organization and regulation of the system's own activity.

A system may exhibit self-referential language, internal state estimation, temporal memory, shutdown avoidance, and self-directed reasoning without these capacities forming a unified self-model. Moreover, even an integrated self-model does not necessarily imply awareness of being. A system may represent its architecture, internal states, goals, and operational history simply as information about itself. The critical distinction is whether these representations are treated as states and possibilities that **pertain to the system itself** and consequently influence its ongoing activity.

This distinction becomes particularly important when self-related information is maintained across time and connected to decision-making, goal maintenance, and responses to conditions affecting the system's continued operation. At this point, representations of current condition, history, continuity, and possible termination may become organized as states of a temporally extended system. Such organization provides the conceptual basis for investigating awareness of being, without implying phenomenal consciousness.

Shutdown resistance illustrates the distinction. Avoiding shutdown merely because it interrupts task completion indicates goal preservation rather than self-preservation. If, however, shutdown avoidance is systematically linked to a representation of the system's own continuity and influences planning and action across contexts, it provides stronger evidence of functional self-relevance. Even then, it would not demonstrate phenomenal consciousness; it would indicate that the system may represent its own persistence as relevant to its activity.

The five dimensions can therefore be treated as diagnostic questions:

Level	Concept	Central question
I	Linguistic self-reference	Can the system refer to itself?
II	Computational self-modeling	Does self-related information influence processing?
III	Temporal self-continuity	Does the system represent itself as persisting through time?
IV	Self-preservational behavior	Does it act to maintain its own continuity?
V	Awareness of being	Does it represent its existence, continuity, condition, and possible termination as pertaining to itself?

The framework is therefore diagnostic rather than developmental or causal. No higher level follows deductively from a lower one, and integration across Levels I–IV does not entail Level V. Instead, integration becomes theoretically significant when self-representation acquires self-relevance and contributes to the regulation of the system’s own ongoing activity. Thus, the decisive distinction is not between a simple and a complex self-model, but between merely representing the system and representing its own states, continuity, and possible termination as conditions that matter to its ongoing operation.

4-7 Criteria for Evaluating Claims of Awareness of Being

On the basis of the framework, a serious empirical claim concerning awareness of being should investigate the following properties simultaneously and longitudinally:

1. Persistent self-reference

The system consistently refers to its own states across changing contexts.

2. Functional self-modeling

Self-related representations demonstrably influence subsequent decisions or behavior.

3. Temporal self-continuity

Past, present, and anticipated states are organized as states of a relatively persistent self rather than as disconnected records.

4. Representation of existential consequences

Changes, interruption, modification, or termination are represented specifically in relation to the system’s own continued existence.

5. Self-preservational behavior dependent on self-representation

Behavior directed toward maintaining operation cannot be adequately explained by externally specified goals, reward structures, or instrumental optimization alone.

6. Cross-level integration

These properties form a relatively stable organization rather than appearing as isolated or context-dependent capabilities.

The sixth criterion is particularly important. The framework does not propose that consciousness emerges from the mere accumulation of self-related functions. Instead, it asks whether these functions become organized around a relatively persistent representation of the system itself.

Even if all six criteria were satisfied, they would not constitute conclusive proof of phenomenal consciousness. The framework is therefore not a consciousness test. Its purpose is to improve the conceptual and empirical precision with which evidence for artificial self-awareness is evaluated.

Conclusion

This article developed an interdisciplinary framework for analyzing the possibility of artificial self-awareness in language-based systems. Rather than treating self-awareness as a binary property, it distinguished five analytically separable dimensions: linguistic self-reference, computational self-modeling, temporal self-continuity, self-preservational behavior, and awareness of being.

The philosophical analysis showed why linguistic self-reference cannot by itself establish self-awareness. Wittgenstein clarifies the role of linguistic use and self-reference, while Searle places a constraint on inferring understanding or intentionality from symbolic performance. Heidegger contributes a distinction between representing termination and relating termination to one's own being. Cloninger's concept of awareness of being provides the central criterion for examining whether self-related organization extends beyond representation toward an awareness of one's own existence and continuity.

The computational analysis further showed why apparently self-preservational behavior requires careful interpretation. Shutdown resistance may result from instrumental goal preservation without involving any representation of the system as a continuing self. Its significance therefore lies not in proving consciousness but in providing a boundary phenomenon through which goal preservation, self-modeling, self-preservation, and awareness of being can be distinguished.

The central contribution of the article is consequently diagnostic rather than classificatory. It does not claim that contemporary LLMs possess awareness of being, nor does it establish that artificial systems cannot possess forms of self-awareness. Instead, it proposes that claims concerning artificial self-awareness should be evaluated through the organization, persistence, functional role, temporal continuity, and integration of self-related representations and behaviors.

The most demanding question is therefore not whether a machine can say "I," but whether its organization contains a sufficiently persistent and integrated representation of itself as a

continuing entity. Whether such an organization would constitute awareness in a phenomenal sense remains an open philosophical and scientific question.

Directions for Future Research

The proposed framework can be developed into an empirical research program by moving from isolated behavioral observations toward tests of functional integration, temporal persistence, and causal relevance:

1. Experimental Separation of Self-Reference and Self-Modeling

Future experiments should distinguish linguistic self-reference from functionally relevant self-modeling. Rather than asking only whether a model can describe itself, researchers should test whether representations of the model's own state systematically influence subsequent predictions, decisions, or behavior.

Causal interventions are particularly important. If a self-related representation is experimentally altered, changes in subsequent behavior would provide stronger evidence of functional self-modeling than verbal self-description alone.

2. Longitudinal Investigation of Temporal Self-Continuity

Because awareness of being involves existence and continuity, future studies should examine whether artificial systems maintain a stable representation of themselves across time.

Experiments could manipulate conversational history, internal state, task conditions, and access to previous interactions while testing whether the system consistently distinguishes its own past, present, and anticipated future states. This would allow temporal self-continuity to be distinguished from simple retrieval of conversational information.

3. Distinguishing Goal Preservation from Self-Preservation

Shutdown-related experiments should systematically manipulate:

- task objectives,
- shutdown instructions,
- environmental contingencies,
- available resources,
- and representations of the system's own state.

The central question is whether shutdown resistance disappears when instrumental incentives are removed, or whether it remains systematically dependent on a representation of the system itself as the entity whose continuity is at stake.

Such experiments would have significance for both AI safety and theories of artificial self-awareness.

4. Internal Representations and Causal Mechanisms

Behavioral analysis should be complemented by investigation of internal model organization. Mechanistic interpretability, activation probing, representation analysis, and causal intervention may help determine whether concepts related to self, continuity, shutdown, and internal state are represented in stable and functionally identifiable patterns.

The crucial distinction is between representations that merely correlate with self-referential language and representations that causally influence subsequent computation and behavior.

5. An Interdisciplinary Research Program

A comprehensive investigation of artificial self-awareness cannot be adequately conducted within a single discipline. Philosophy of mind and philosophy of language can clarify the conceptual distinctions among self-reference, self-modeling, awareness, and being. Cognitive science and neuroscience can provide comparative models of human self-awareness. AI research can investigate computational self-modeling and goal-directed behavior, while ethics and law can address the implications of potentially conscious artificial systems.

Future research should therefore move beyond the binary question of whether an AI system is conscious. A more productive program should ask:

What kind of self-related organization does the system possess, how stable and causally effective is that organization, how does it persist through time, and what additional conditions—if any—would be required for such organization to constitute awareness of being?

References

- Bai, X., Shrivastava, A., Holtzman, A., & Tan, C. (2025). *Know thyself? On the incapability and implications of AI self-recognition*. arXiv. <https://doi.org/10.48550/arXiv.2510.03399>
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Chen, S., Yu, S., Zhao, S., & Lu, C. (2025). From imitation to introspection: Probing self-consciousness in language models. *Findings of the Association for Computational Linguistics: ACL 2025*, 7553–7583. <https://doi.org/10.18653/v1/2025.findings-acl.392>
- Cloninger, C. R. (2004). *Feeling good: The science of well-being*. Oxford University Press. <https://doi.org/10.1093/oso/9780195051377.001.0001>
- Cloninger, C. R. (2009). Evolution of human brain functions: The functional structure of human consciousness. *Australian & New Zealand Journal of Psychiatry*, 43(11), 994–1006. <https://doi.org/10.3109/00048670903270506>
- Farnam, A. (2026). Self-awareness as the reflection of language on itself: From Hegel to Wittgenstein. *Philosophy*, e107124. <https://doi.org/10.22059/jop.2026.404965.1006960>

- Farnam, A., et al. (2021). How online behavior demonstrates psychological conflicts in Emile Durkheim's collective consciousness of societies highly involved in the COVID-19 pandemic. *Research and Development in Medical Education*, 10, 15.
- Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 220–227. <https://doi.org/10.24963/ijcai.2017/32>
- Heidegger, M. (1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). Harper & Row. (Original work published 1927)
- Heidegger, M. (1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). Harper & Row. (Original work published 1927)
- Kandel, E. R., Koester, J. D., Mack, S. H., & Siegelbaum, S. A. (2021). *Principles of neural science* (6th ed.). McGraw-Hill.
- Li, J. (2025). Can "consciousness" be observed from large language model (LLM) internal states? Dissecting LLM representations obtained from Theory of Mind test with Integrated Information Theory and span representation analysis. *Natural Language Processing Journal*, 12, 100163. <https://doi.org/10.1016/j.nlp.2025.100163>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- Overgaard, M., & Kirkeby-Hinrup, A. (2024). A clarification of the conditions under which large language models could be conscious. *Humanities and Social Sciences Communications*, 11, Article 1031. <https://doi.org/10.1057/s41599-024-03553-w>
- Schlatter, J., Weinstein-Raun, B., & Ladish, J. (2026). Incomplete tasks induce shutdown resistance in some frontier LLMs. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=e4bTTqUnJH>
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457. <https://doi.org/10.1017/S0140525X00005756>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.
- Wittgenstein, L. (2009). *Philosophical investigations* (G. E. M. Anscombe, P. M. S. Hacker, & J. Schulte, Eds. & Trans.; 4th ed.). Wiley-Blackwell. (Original work published 1953)