

Article

The Standpoint Threshold: Architectural Conditions for Processing from Somewhere

Daniel H. Lange 

Department of Electrical and Computer Engineering, Technion—Israel Institute of Technology, Technion City, Haifa 3200003, Israel; dlange@technion.ac.il

Abstract

The philosophy of artificial intelligence often begins one question too late. Discussions move directly to whether a system could be conscious, whether there is something it is like to be the system, or whether it possesses genuine subjectivity, without first asking a more basic structural question: what would have to be true of a system's internal organization before those questions gain a determinate basis in its architecture? This paper addresses that prior question. It defends a threshold account built on three jointly necessary conditions: persistent internal context, differentiated internal possibility, and closed-loop operativity. Together, these conditions mark the point at which a system's processing can plausibly be said to unfold from an internal standpoint, rather than merely as a chain of locally linked transitions. Such organization may also have functional significance independent of consciousness by supporting temporal coherence, cross-role coordination, and reduced dependence on episodic reconstruction or external scaffolding. The account explains why simple feedback controllers are clear non-cases, why standard transformer inference remains at most a bounded near-case, and why world-model architectures are more serious candidates for assessment, while making no claim that such systems are conscious. The aim is not to provide a test for artificial consciousness, but to clarify when consciousness-oriented inquiry into intelligent systems has a determinate architectural target.

Keywords: standpoint threshold; persistent internal context; stance-space; cognitive architecture; artificial consciousness; philosophy of artificial intelligence

1. Introduction

A foundational question is rarely asked: before debating whether an AI system is conscious, we must first ask whether it is the kind of system to which consciousness attributions could apply. That prior question, concerning internal organizational structure rather than phenomenology, is the concern of this paper.

When the question is neglected, the debate polarizes in predictable ways. Some observers point to behavioral sophistication or context sensitivity and conclude that something like perspective or proto-subjectivity is present; this is precisely the inference from surface fluency that Shanahan [1] cautions against, and the kind of attribution that recent work in the science of consciousness has tried to discipline [2]. Others dismiss such talk outright because phenomenal consciousness is not observable, a refusal that Chalmers [3] argues is too quick. Both responses bypass an intermediate level of organization: a level at which stronger attributions might first become more than rhetorical.

This paper identifies that intermediate level. It does not claim that any current AI system is conscious, self-aware, or genuinely subjective. The central claim is that the



Academic Editors: James Hutson and Luca M. Possati

Received: 12 May 2026

Revised: 23 June 2026

Accepted: 14 July 2026

Published: 20 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

standard debate often begins too late. Before asking whether a system is conscious, we need to ask whether it has the internal organization that would give the question a well-formed target. That is a structural question. It precedes consciousness rather than follows from it.

Two terms need clarification. I use “well-formed” in a specifically structural sense: a question is well formed about a system when some property of the system’s own organization could make an answer correct. This is not the only legitimate sense. On instrumentalist or interpretivist views, including Dennett’s intentional stance [4], attribution may instead be warranted by the predictive or interpretive utility of an external stance. I do not dispute that use; I set it aside because this paper asks when standpoint talk has an architectural basis in the system itself. Similarly, when I describe some consciousness attributions as trivial or premature, I mean only that they lack a determinate basis in the system’s organization, not that current consciousness research is mistaken or careless. The threshold is meant to locate where concepts from consciousness science and philosophy gain architectural purchase. The guiding aim is to identify when “processing from somewhere” stops being a metaphor and becomes a structural description.

The notion of internal standpoint at work here is deliberately modest. It does not invoke Nagel’s sense of perspective, Husserl’s first-person givenness, or felt subjectivity of any kind [5,6]. It means that the system’s processing unfolds from a persisting internal orientation—not as a mere succession of stimulus-driven transitions, but from a maintained internal position. In compact form, a minimal structural standpoint is a maintained, history-shaped internal position within a space of system-reachable alternatives that regulates later processing across more than one role.

Persistence-only and control-based views identify important forms of continuity and regulation, but they leave open what would make such regulation standpoint-like. Self-model and consciousness-first views address richer forms of self-representation or conscious subjectivity, but they presuppose more than the minimal structural threshold at issue here [4,7]. Dynamical and state-space approaches identify important structure but do not distinguish analyst-imposed descriptions from positions the system can occupy and continue from [8,9]. What is needed is a threshold concept: one that tells us when the stronger debate becomes structurally well formed, without itself collapsing into consciousness.

The framework operates below the standard landmarks in consciousness studies: phenomenal character, reportability, access consciousness, and recurrent processing [10–13]. It is not a rival theory of consciousness and does not compete with global workspace, integrated information, or predictive-processing accounts [14–17]. Its narrower point is that those theories require an appropriate architecture to which their central notions, such as broadcast, integration, or prediction-error hierarchy, can apply. The present account asks a prior question: whether the organization in question is one the system can occupy and continue from, or only one an analyst can describe. The focus throughout is the active architecture under examination: not behavior or outputs, but how processing is internally organized.

The threshold rests on three conditions: persistent internal context, differentiated internal possibility, and closed-loop operativity. The paper derives them sequentially, with each condition responding to a limitation exposed by the previous stage of the argument. First, I show why mere state persistence, however history-sensitive, falls short. I then ask what would make a persisting condition count as a position the system processes from, rather than merely a background parameter. Finally, I ask whether that position has organizational significance across the system. The three conditions answer these pressures in turn.

Standpoint may also have functional significance independent of consciousness. A system that processes from a maintained internal orientation need not merely respond to present input, retrieve stored records, or reconstruct context episodically. Because relevant constraints persist within the active causal basis of processing, such a system can sustain a line of activity without re-deriving its situation at each step. When a single maintained position regulates several functions, prediction, planning, updating, and action selection can remain mutually aligned without a separate coordination mechanism. Such organization may support more stable planning, more context-sensitive generalization, more robust recovery from perturbation, and reduced dependence on external scaffolding. These are not marks of consciousness, but structural conjectures about the functional advantages such architectures may make available.

The paper proceeds as follows. Section 2 fixes the scope, target systems, and terminology of the discussion. Section 3 explains why state persistence cannot ground internal orientation. Sections 4–6 develop the three conditions in turn. Section 7 shows why neighboring proposals fall short and locates the account relative to consciousness science. Section 8 addresses objections, including the concern that the framework merely renames statefulness and the concern that it omits stabilizing dynamics. Section 9 applies the framework to several artificial architecture types and offers a diagnostic schema. Section 10 concludes the paper.

2. Scope, Target Systems, and Terminology

Before developing the conditions, this section helps to fix the scope of the argument and the technical terms used below. Rather than covering artificial intelligence in general, the discussion focuses on contemporary deep learning systems with varying forms of recurrence, memory, tool use, active inference, and world modeling. The clearest contrasts are drawn between simple feedback controllers, which serve as clear non-cases, and Dreamer-style world-model architectures, which represent stronger candidates for assessment; standard transformer inference and contemporary conversational language-model systems occupy the difficult middle.

Throughout, the target of assessment is the model-internal process—the computation a system performs in virtue of its own architecture and weights—rather than the larger deployment in which that process is embedded. Claims about standard transformer inference are therefore conditional on a fixed model-internal setting: fixed tokenization, architecture, parameters, and deterministic forward computation. Deployment-level routing, hidden context construction, tool calls, external memory injection, batching effects, and implementation-level nondeterminism are set aside unless they are explicitly part of the architecture under assessment. These features may be crucial to how contemporary systems behave in practice, but they should not be credited to the model-internal process unless they are genuinely part of the architecture being evaluated.

Several machine learning terms should be fixed at the outset. A latent or hidden state is an internal vector of activations that influences processing without being part of the output. A recurrent network updates such a state step by step, feeding each state into the next step. A transformer processes a sequence by relating its tokens to one another through attention. Tokens are the discrete units into which text is divided, often including words, word-parts, or punctuation. During generation, a transformer also maintains a key-value cache: intermediate quantities computed from earlier tokens so that they need not be recomputed at each step. An attractor is a region of state space toward which nearby trajectories tend to converge under the system's own dynamics. Parameters, or weights, are the values fixed during training that define the function the network computes. Inference

is the phase in which a trained model is run on new input, normally with those parameters held fixed.

These distinctions matter because they mark the difference between a record of what a system was given and an orientation the system has developed and now operates from. A token sequence, retrieved memory, or key-value cache may preserve information from earlier processing without constituting a standing internal orientation. Conversely, a recurrent or world-model architecture may maintain an internal condition that is updated by the system's own activity and continues to shape later processing. The threshold account developed below turns on this distinction.

3. From State Persistence to Internal Orientation

The natural starting point is the internal state. Artificial systems carry forward hidden activations, recurrent variables, latent summaries, memory traces, learned parameters, and slowly varying control variables. These can be history-sensitive and causally important. Why not identify persistent internal orientation with one of them?

The reason is that persistence alone, even causally active persistence, is too thin. A thermostat carries a temperature estimate forward. A navigation controller preserves positional information. A recurrent classifier retains task-relevant information. None of them seems to process under anything that would count as a persistent internal orientation. If persistence were enough, the threshold would be trivial in the sense defined above: the attribution would have no determinate structural basis beyond the bare fact of state retention.

The point sharpens with a more sophisticated case. Consider a recurrent neural network trained on extended sequential prediction. During inference, it maintains a continuously updated hidden state—richer than any register, shaped by the input sequence, and causally consequential for later outputs. It might seem that this already amounts to an internal orientation. But the decisive issue is not simply that the same hidden state trajectory is produced when the hidden state is reset to the same initial condition and the same input sequence is run under the same update rules. That is true of deterministic systems quite generally, including systems that might otherwise be serious candidates for internal orientation. The relevant contrast is different. A retained structure functions as orientation, rather than merely as record, when current explicit input can be held fixed while differences in prior internally carried organizational history still make a durable difference to how processing now unfolds. If the same internal condition is simply reconstructed from the supplied sequence under fixed update rules, the structure records the sequence. If a prior internally maintained organization persists and continues to modulate present processing independently of the currently supplied input, it begins to support the stronger attribution of orientation.

The record–orientation distinction is central. A record stores information about the past. An orientation is the effect of the system's own processing history on how it now operates. The two come apart readily in practice: learned parameters preserve training history without keeping it as a currently active condition; external memory preserves past information without generating an active internal state; and recurrent hidden state can be too task-bound, too local, or too episodically reset to constitute a persisting background under which processing unfolds.

The distinction sharpens in counterfactual terms. A retained structure functions as an orientation, rather than only as a record, when varying the system's prior internally carried organizational history, while holding current explicit input and current architecture fixed would change how processing now unfolds. This is a stricter criterion than generic causal dependence. Any deterministic system's state depends on prior states and inputs; that fact alone is not philosophically informative. The central question is whether the system's

own accumulated and actively carried history, rather than only a supplied sequence or externally reconstructed context, causally shapes how it currently operates.

This also explains why standpoint cannot be captured by a momentary snapshot. Nagel's "what is it like to be a bat" is usually read as a claim about phenomenology, but it can also motivate a more general point: perspective is not well modeled as a disconnected instantaneous item [6]. Phenomenological ideas of horizon are useful here in a deflationary way: they remind us that the present is intelligible only against an organized background of possible continuation [5,18], without adding any further metaphysical commitment.

To say that processing unfolds from a persistent internal orientation is therefore modest but substantive: later processing occurs under a persisting internal condition that cannot be reduced to an instantaneous input–output description.

Human phenomenology presents a further asymmetry. In discussions of human perspective, the starting point is usually lived givenness; unity is something to be explained. In machines, nothing of that sort can be assumed. The explanatory order must therefore be reversed: before asking whether an artificial system exhibits meaningful internal differentiation, we must ask whether there is any persisting internal orientation within which such differentiation could matter.

Thus, a minimal standpoint is not simply a lasting state. It is an internally maintained, history-shaped organizational condition under which ongoing processing unfolds. More specifically, it is the system's occupation of a differentiated field of internally occupiable stances, where occupying one position rather than another systematically shapes how processing continues. That points to three requirements: (1) a continuing internal condition, (2) a structured field of alternatives genuinely available to the system through its own operation, and (3) a regulatory role for the system's current position in determining subsequent processing.

4. Persistent Internal Context (Condition 1)

The first condition is persistent internal context: a standing internal condition whose current form depends on both its prior form and ongoing processing, and which in turn influences subsequent processing. This is not merely a state that lasts, but a persisting internal background under which later transitions occur. It is implementation-neutral: it could be realized by a recurrent hidden state, continuously maintained latent variables, attractor organization, or distributed internal representations. What matters is the role, not the substrate.

Three constraints define the condition. The first is incremental preservation: the context must persist and be updated across transitions, not repeatedly discarded and reconstructed. The second is endogenous update: it must be modified through the system's own ongoing activity, not wholly supplied or replaced from outside. A context is endogenous when its form depends on the system's internally unfolding trajectory and cannot be recovered from current external input alone. The third is active carryover: the context must remain within the active causal basis of later processing, rather than being inertly stored or merely recoverable on demand.

These constraints rule out several failure modes. A context reconstructed only by re-querying external storage is not a persistent internal context. A variable that remains as inert data until retrieval is not active carryover. A state repeatedly discarded and rebuilt from current input does not sustain temporally extended orientation.

The point becomes clearest in cases that appear to preserve context while still falling short of persistent internal context in the fuller sense. The difficult and practically important case is standard transformer inference at the model-internal level. Here, the target is the inference process itself, not a larger deployment involving tools, databases, retrieval

systems, agent loops, hidden context construction, dynamic routing, batching effects, or implementation-level nondeterminism. Within a single inference pass, a transformer carries forward an internally updated residue of prior processing [19]. In that limited intra-pass sense, it plausibly satisfies a weak endogeneity condition.

A clarification is needed because contemporary language models often generate text stochastically. Sampling can produce different next tokens from the same probability distribution, so the same prompt may yield different outputs on different runs. But this does not alter the point at issue. The claim concerns the forward computation that constructs the key-value cache from a given token sequence, not the sampling of an output token. Holding token sequence, tokenization, architecture, parameters, and implementation fixed, the cache is a fixed function of the supplied sequence. Stochastic decoding therefore adds output variability without converting a sequence-determined record into a history-shaped orientation. Nor is determinism itself the issue: deterministic systems can in principle carry internal organizational history forward.

Does intra-pass carryover amount to persistent internal context in the fuller sense? Not straightforwardly. The relevant question is whether the architecture maintains an ongoing internal context across episodes. On that front, standard transformer inference is better understood as a bounded near-case than as a clear positive instance. Under the fixed model-internal conditions just specified, later processing proceeds from an encoding of the token sequence, not from an internally maintained standing condition that varies with the system's prior inference-time trajectory. That is the mark of record, not orientation.

Two deficits are especially important. First, episode-boundedness: the carryover ends when the run ends, making it difficult to treat as a standing internal condition across episodes. Second, input dependence at the relevant level: the key-value cache is not a passive transcript, since keys and values are constructed by the model's own learned transformations, but those transformations are the same for every run. The cache therefore does not by itself establish dependence on the system's prior organizational history in the relevant sense.

A further challenge concerns training-time versus inference-time trajectory dependence. One might argue that the model's parameters already encode the effects of training, and that parameter-mediated shaping of the cache therefore counts as trajectory dependence. But the relevant question is whether the internal condition varies with the system's own ongoing activity. In standard inference, parameters remain fixed across episodes and do not vary with this particular run. The weights are frozen, so carryover is held either in the explicit token sequence or in deterministic, sequence-fixed functions of it, not in an endogenously altered internal substrate. That is the precise sense in which the relevant carryover remains a record rather than an orientation.

The question, therefore, is whether current processing is organized by an actively maintained, episode-relative condition that varies with what the system has just been doing, not simply whether the system has some causal history in the broad sense. In that broad sense, every trained system has a causal history; what matters here is whether that history is actively carried forward as a system-internal condition during inference.

Deployed systems also frequently achieve apparent continuity by offloading state into external stores, tools, or human-maintained scaffolds. Those arrangements can produce impressive practical coherence. They do not, however, show that the model-internal process carries its own prior activity forward as a standing internal condition.

Persistent internal context is necessary: without it, there is no standing condition for processing to unfold from. But it is only the first requirement. Even if a system has genuine endogenous, causally active, cross-episode internal context, a further question

remains: what makes that context a position the system processes from, rather than merely a background parameter shaping its outputs? That is what Condition 2 addresses.

5. Differentiated Internal Possibility (Condition 2)

Suppose a system satisfies Condition 1. It carries forward an endogenously updated context that actively shapes later processing. Is that enough for a minimal internal standpoint? It might seem so: the system has a history-shaped internal condition from which later processing proceeds. But that answer moves too quickly.

The difficulty is this. Imagine a system whose continuously maintained latent state can be partitioned by an external analyst into regions correlated with different processing modes. That partition may support reliable prediction from the outside, but it does not by itself establish an internal standpoint. What matters is whether occupying one region rather than another changes which continuations are live for the system under its own dynamics. If it does not, the analyst has described a state-space, not a stance-space the system can occupy and continue from.

The distinction is significant. A current internal condition becomes intelligible as an internal standpoint only when it places the system within a structured field of internally occupiable stances—positions from which different continuations remain available through the system's own ordinary operation. More precisely, a system exhibits stance-space only when its current position is reachable under its own update dynamics, maintainable for more than one processing step without external reintroduction, and capable of durably altering the space of live continuations under comparable present input. This is what distinguishes stance-space from a merely retained variable or an analyst-imposed partition.

The phrase “internally occupiable” might seem to smuggle in a subject, but it does not. A position counts as occupiable if and only if the system can reach and maintain it under its own ordinary transition dynamics for more than one processing step without external reintroduction.

An interventionist gloss sharpens the criterion. A candidate geometry is a proposed structure of the system's internal state space—a way of carving that space into regions, with relations of nearness and distance among them, offered as a description of the system's stances. The question is whether this structure tracks something the system itself occupies and continues from, or only something an analyst has imposed. It counts as stance-space only when similar positions yield similar downstream modulation profiles across multiple future pathways, different regions yield systematically different profiles, and these regularities remain robust under admissible perturbations that vary the current position without destroying the architecture.

Several clarifications follow. First, the relevant possibility space is not the set of conditions a designer could impose by retraining or weight editing. It is the space of internal conditions reachable from the system's current state under its ordinary update dynamics, holding architecture, weights, and training regime fixed.

Second, the condition blocks observer relativity. Any sufficiently complex system can be carved by an analyst into states and clusters; the question is whether those distinctions correspond to alternatives the system can occupy as part of its own unfolding organization. Describable state-space is not enough. The central contrast is between merely describable state-space and internally occupiable stance-space.

Third, the multi-step requirement is essential. A different internal position that produces a different immediate output but leaves the overall space of future processing unchanged does not count. A difference counts as stance-like only when it makes a durable difference to how the system can continue, persisting across the architecture's ordinary timescale rather than disappearing after a single transition.

This also shows why orientation cannot be identified with whatever state the system currently occupies. If orientation just is the current state, every transition becomes a change in orientation, and the distinction between ongoing orientation and mere succession collapses.

Consider a more promising case: an agent with a maintained latent state that is updated during interaction, used in prediction and planning, and not reset after each step. Suppose its latent organization supports relatively stable internal modes, such that, from one mode rather than another, different planning branches remain live, some hypotheses are easier to retain, and some update paths are harder to reach. If these differences persist under comparable present inputs because they reflect the system's own transition structure rather than a temporary summary of the latest stimulus, then the latent organization is plausibly functioning as an internally occupiable stance-space.

The notion is deliberately neutral with respect to semantics. A stance can have organizational significance without being belief-like, propositionally articulated, or interpretable in folk-psychological terms. Occupiable stances need not be finite, discretely bounded, or mutually exclusive. The point is simply that the system's transition dynamics are structured by current internal position: occupying one position rather than another makes a determinate and lasting difference to which continuations remain live.

A natural worry is that this formulation is too permissive. Does any system with modestly path-sensitive dynamics qualify? No. The requirement is that current position materially alters the available continuation space, not merely the next output. Even so, Conditions 1 and 2 together may not suffice. A system could maintain an endogenous context and occupy a genuine stance-space, yet have those stances regulate only a single locally bounded function. The next section explains why a further condition is needed.

6. Closed-Loop Operativity (Condition 3)

Suppose a system satisfies both Conditions 1 and 2. It carries forward a genuinely endogenous context and occupies an internally available stance-space in which different positions alter the paths along which processing can continue. Is that enough?

Not yet. Consider a system whose maintained latent organization supports distinct modes that durably alter the rollouts available within a bounded planning routine. From mode A, certain imagined trajectories remain stable; from mode B, others do. But suppose no function outside that routine is affected: prediction, interpretation, learning-rate adjustment, and action selection are all indifferent to which mode the system occupies. The difference is real, but local. It is more like a control setting within a subsystem than an organizing standpoint. Internal standpoint, as opposed to local path sensitivity, requires a current internal position to have organizational consequence across multiple processing roles, not only within one bounded routine.

Condition 3 addresses this. Condition 2 establishes stance-space: it asks whether the current internal position changes what continuations are live at all. Condition 3 establishes organizational significance: it asks whether occupying such a position helps regulate more than one ongoing function. The two conditions are not redundant. A system can possess a genuinely occupiable field of internal positions and still fail to process under an internal standpoint if the downstream consequences of those positions remain confined to one local role.

A structured internal organization is operative only when differences in the current position correspond to systematic differences in downstream causal influence across more than one bounded processing role. This is not undifferentiated whole-system influence. It is cross-role significance for more than one ongoing function, such as interpretation,

prediction, updating, planning, or action selection. A variable that modulates only one architecturally central routine remains a task-bound regulator.

A bounded processing role is a functionally distinguishable contribution to the system's ongoing activity. The individuation is architectural and explanatory, not merely verbal: two roles are genuinely distinct when admissible interventions can selectively alter the downstream consequences of the current stance-position for one without correspondingly altering them for the other. The test is not met by merely relabeling one downstream effect as several distinct roles.

Two contrasts clarify why Condition 3 cannot be reduced to Conditions 1 and 2. First, a system with persistent latent modes that genuinely alter which planning branches remain available, but only within one bounded rollout routine, may satisfy Conditions 1 and 2 while failing Condition 3: the position differences do not reorganize more than one local role. Second, a system with a single slowly varying gain parameter that modulates attention, output routing, and the update rate might seem to satisfy Conditions 1 and 3, since it carries forward a broad control variable with wide downstream influence. But it fails Condition 2 because it does not occupy a differentiated field of internally reachable stances. A gain parameter is a distributed regulator, not a standpoint.

Broad control is not orientation. A system can integrate signals, modulate multiple routines, and exhibit wide feedback while still lacking a structured field of stances it can inhabit and continue from. Orientation requires regulation from a position within an internally occupiable stance-space, not from distributed control variables alone.

The required structure is a loop. Ongoing internal activity updates persistent context; that context places the system within a differentiated field of possible stances; and the current position within that field modulates further internal activity across more than one ongoing role. Remove the stance-field, and one is left with stateful control. Remove the operative role, and one is left with structured description. Only their conjunction yields the kind of organization at issue here.

Diagnosing this structure involves three questions. First, can a candidate internal stance vary without destroying the architecture? Second, do differences in that stance affect more than one ongoing function? Third, are those differences better explained by the system's position within a structured field than by a single local variable or control signal?

The evidential logic here is interventionist, not merely behavioral [20]. It is not enough that a system produces different outputs after different histories; many stateful systems do that. The question is whether controlled variation in current internal position produces patterned differences in how several ongoing functions are subsequently organized.

Taken together, the three conditions define the minimal structural threshold. Condition 1 makes "from" meaningful: there is a standing condition. Condition 2 makes "somewhere" meaningful: that condition locates the system in a structured field of alternatives. Condition 3 makes the attribution explanatorily significant: the current position regulates more than one ongoing function. The conditions respond to distinct pressures: continuity, internal differentiation, and operativity. They are jointly necessary for the minimal attribution at issue. Questions about phenomenology, intentional content, and self-representation remain open; they are not established by this threshold, though stronger attributions may or may not build on it.

The threshold claim is deliberately modest. Once all three conditions are met together, it is no longer merely metaphorical to say that the system processes from a maintained internal orientation. Features such as embodiment, explicit self-modeling, semantic content, or phenomenal character may be required for richer mental attributions, but not for this thinner one.

Crossing the threshold matters in two ways. First, consciousness-oriented debate becomes structurally well formed in the sense defined in the Introduction: the indicators identified in consciousness science become relevant targets of inquiry, rather than questions applied to the wrong kind of architecture [2,12]. Second, attributions of agency, planning, and goal-directed organization gain a possible structural footing. Without the threshold conditions, saying that a system plans remains primarily a behavioral description: it describes how outputs are organized, not how processing is internally maintained and coordinated. The conditions also carry a constructive implication: they indicate how architectures would have to be organized if artificial systems were to maintain orientation rather than merely reconstruct context episodically.

The three conditions constrain case assessment in an ordered way. Failure at Condition 1 blocks standpoint attribution altogether: there is no standing condition for processing to unfold from. Failure at Condition 2 yields continuity without stance-space: something is being carried forward, but not as a position within a live field of alternatives. Failure at Condition 3 yields stance-space without broader organizational significance: the system occupies a genuine internal position, but that position is too locally bounded to function as a standpoint. Near-cases are not borderline positives. They are cases where one or more of these pressures is only partially met. The framework is discriminative, not inflationary.

7. Why Weaker and Nearby Accounts Fall Short

With the three conditions in hand, we can now locate the limits of weaker and nearby accounts. Each captures something real, but each secures only part of the required structure. The framework's added value lies in the conjunction: persistent internal context, differentiated internal possibility, and closed-loop operativity together mark the regime at issue.

The simplest view equates internal orientation with any persisting internal state. But if retention were sufficient, thermostats and simple recurrent filters would qualify too easily. The missing ingredient is not greater complexity, but a structured field of internally occupiable alternatives. Condition 1 without Condition 2 gives continuity without stance-space.

Adding causal influence on later processing does not fix the problem. A system can carry forward state, let it modulate behavior, and still remain within the register of control architecture. A controller that preserves variables to alter gain, routing, or thresholding across timesteps can be persistent, recurrent, and causally significant without occupying any field of stances. It is a more elaborate control regime, not an orientation-bearing standpoint.

A stronger family of views identifies internal perspective with rich dynamical organization, trajectory sensitivity, or attractor structure [8,9]. These views plausibly approach Condition 1 and gesture toward Condition 2, insofar as attractor landscapes can support multiple stable regions. But dynamical richness alone does not show that those regions are internally occupiable in the relevant sense. A structured attractor landscape might organize only one processing function, or its regions might mark distinctions useful to an analyst without corresponding to alternatives the system can reach, maintain, and continue from. The dynamical-systems literature supplies the vocabulary of multi-stability and trajectory; Condition 2 adds the requirement that occupying one region rather than another durably reorganizes the system's own live continuations, rather than merely our description of them.

Another family of views emphasizes sophisticated control organization, including the perception–action regulation studied in predictive-processing and free-energy accounts [15,16]. Such views tend to satisfy Condition 1 and may seem to satisfy Condition 3 when control signals are distributed across multiple routines. What they do not by themselves secure is Condition 2. Coordinating interpretation, updating, and action selection

through widely distributed control variables does not yet show that the system has a differentiated stance-space it can inhabit and continue from. Predictive-processing models are instructive in this respect: they describe hierarchies that minimize prediction error across many functions, which may amount to real cross-role regulation. But error minimization as such does not establish that the system occupies distinguishable positions in a reachable field of alternatives, rather than merely undergoing structured control-state variation.

At the opposite extreme, requiring explicit self-modeling or self-representation [7] answers a different question and begins too late. Such accounts address what is required for a system to represent itself as itself, which presupposes that some prior orientation-bearing structure is already in place. A system can process under a persistent internal orientation before it explicitly models itself.

Dennett's intentional stance is a useful contrast. On that view, intentional attribution is an interpretive strategy adopted by an observer, rather than a structural threshold internal to the system being interpreted [4]. The present account asks a different question: whether variation in the system's current internal position produces durable differences in its available continuations under its own dynamics. The contrast is therefore not between a wrong view and a right one, but between an interpretive account and a structural one. Intentional-stance interpretations may be useful even where the standpoint threshold is not met; conversely, a system could meet the structural threshold without yet warranting richer intentional or agential interpretation.

The contrast with minimal-self theories is sharper still. Zahavi [21] argues for a thin pre-reflective sense of selfhood constituted by the first-personal givenness of experience. The framework developed here operates below that level: it identifies structural preconditions that would have to be satisfied before the question of whether there is something it is like to be the system has a determinate architectural target.

The relevance of the threshold to artificial systems can be made concrete through two leading cases. Take integrated information theory first [17]. Its central quantity, integrated information, is evaluated over a system's cause-effect structure under possible partitions, and the theory identifies the relevant system as the complex of maximal integrated information rather than taking a boundary as given. The present account does not dispute that individuation; it asks a complementary question about how such structure is maintained and occupied over time. The issue is whether the structure that is identified is one the system itself can reach, maintain, and continue from under its own dynamics over time. In that sense, Condition 2 asks whether the structure functions as an internally occupiable stance-space, rather than only as a structure an analysis can characterize. The contrast is between synchronic integration and diachronic occupiability, not between a principled and an arbitrary choice of system.

Global workspace theory makes the point from the other direction [12,14]. Its explanatory core is the broadcast of contents from a limited-capacity workspace to multiple consumer processes. That requires contents sufficiently stabilized for broadcast and availability to several specialized processes rather than a single routine. In the present framework, this overlaps with Condition 1's concern for active carryover and Condition 3's concern for cross-role availability, but it does not entail the full standpoint threshold structure. In particular, global workspace theory does not by itself establish Condition 2 that the system occupies positions within an internally reachable and maintainable stance-space. A system that reconstructs its situation at every step, or whose alleged workspace feeds only one routine, may be described in workspace vocabulary without crossing the threshold at issue here. In both cases, the present account is not a competitor: it specifies when the structures each theory quantifies or describes are present in the relevant architectural sense.

Recent work on the philosophy of AI has made the need for this intermediate level vivid. Butlin et al. [2] draw on several theories of consciousness to propose indicator properties that may bear on artificial consciousness; the present account is upstream of theirs, asking what must hold for those indicators to have a determinate architectural target. It is not a replacement for consciousness indicators, but a proposed enabling condition for applying them to artificial architectures. Overgaard and Kirkeby-Hinrup [22] argue that large language models should not be treated as obvious candidates without closer architectural attention; the threshold offered here makes that demand more precise. Shanahan [1] cautions against taking surface fluency as evidence for deeper cognition, a caution the record–orientation distinction formalizes. Chalmers [3] argues that questions about language-model experience deserve serious philosophical treatment rather than reflexive dismissal; the present account agrees, while asking where such treatment gains a determinate architectural target.

The account thus shares the architecture-sensitive spirit of this literature while shifting the target question downward—not to which systems are conscious, but what must be structurally true before that debate is well formed for a given architecture. The aim is not to arbitrate among competing theories of consciousness, but to specify an architectural precondition relevant to their application to artificial systems.

A consistent pattern emerges across these comparisons. Persistence-only views give continuity without stance-space. Control-based and dynamical accounts may capture broad regulation or organizational richness without showing internally occupiable, causally operative stance-space. Self-model and self-representation-first approaches begin too late, presupposing richer organization than the threshold is meant to isolate. The standpoint threshold is meant to occupy the space left open by these alternatives: persistent context, differentiated stance-space, and closed-loop operativity must be present together before processing from an internal orientation becomes structurally well formed.

8. Objections and Replies

The most natural objection is that the framework merely renames familiar computational organization in philosophical language: persistence, state-space structure, and causal efficacy receive new labels, but nothing substantive is added. The reply is that the account does not merely collect familiar ingredients. It combines active endogenous carryover, internally occupiable stance-space, and cross-role operative significance into a narrower organizational profile. Its value lies partly in what it excludes: externally reintroduced continuity, analyst-imposed partitions, scalar control variables, and effects confined to a single local role. It also gives the distinction interventionist consequences, since the conjunction predicts cross-role dissociations that bare statefulness does not. A redescription that generates no new commitments and excludes no cases would add nothing; this one does both.

A related objection is that the view conflates sophisticated control with perspective. It does not. Persistent internal orientation is not control as such, but an internally maintained background situated within a differentiated field of possible stances. The decisive difference lies in Condition 2. A control architecture, however elaborate, may adjust variables toward set-points, route signals, or distribute regulation across many routines without constituting a field of positions the system can reach, maintain, and continue from under its own dynamics. The interventionist test in Section 6 separates the cases: in a control regime, varying the controlling variable shifts regulated quantities without durably reorganizing which continuations remain live across roles; in a standpoint-bearing regime, varying current position does reorganize those continuations, with effects that persist across steps and dissociate across roles under selective intervention. The conflation objection therefore

mistakes wide causal reach for occupied position. The framework keeps them apart by requiring reachable, maintainable, multi-step, cross-role differentiation.

A more serious objection is that the framework is too vague to constrain cases. Terms such as “differentiated,” “internally occupiable,” and “operative” are not metrics. That is true, but philosophical categories often become useful before they admit clean operationalization. The burden at this stage is not to provide a measurement procedure, but to show that the category is disciplined by non-ad hoc exclusion criteria and yields a concrete diagnostic schema. Section 9 supplies that schema.

A related worry is that allowing near-cases and partial approximations makes the framework inflationary. But the three conditions are jointly necessary for full attribution. A graded framework need not be permissive; it can be discriminative precisely by distinguishing clear failures, bounded near-cases, and stronger candidates without collapsing them together.

A deeper objection is that the internal standpoint may not be a genuine explanatory kind: persistence, internal differentiation, and causal influence may seem to be a loose family of related phenomena rather than a unified category. The reply is that the conjunction is principled. One pressure is temporal continuity, one is architecturally consequential internal differentiation, and one is regulatory significance. These are not arbitrarily bundled; each was derived in turn as the answer to a specific insufficiency in the preceding one.

A further objection concerns the level of analysis. Many contemporary AI systems are sociotechnical assemblages in which model passes, external memory, tool use, and human curation jointly sustain continuity. One might therefore treat the whole assemblage as the standpoint bearer. The present question is narrower: when does it become serious to attribute an internal standpoint to an artificial system in virtue of its own organization? A deployment can appear stable and context-sensitive because continuity is reintroduced externally through logs, databases, interface constraints, or human users. The framework therefore insists on level discipline. Whether larger sociotechnical systems satisfy analogues of the three conditions is a further question, not an objection to this one.

A distinct and important objection concerns what the threshold leaves out. The three conditions are architectural: they specify what a system must have, not the regulatory and attentional dynamics by which a standpoint, once present, is actively sustained as a stable cognitive orientation. A system might satisfy all three conditions and yet be unable to hold specific content in the foreground, preserve its relevance, and prevent associative drift—unable, for instance, to keep the line “my feet are wet, so I need boots” in view without sliding into loosely associated material about farming or rural life. Content stability of this kind requires dynamics of maintenance and selection that go beyond persistent context, occupiable stance-space, and cross-role operativity.

The objection is correct insofar as the threshold is architectural. The three conditions identify a necessary lower bound, while leaving open whether further dynamics are required for a fully sustained cognitive standpoint. That concession does not weaken the account; it marks its intended scope. The dynamic question—how feedback updates a standpoint so that it remains coherent and resists drift—is a separate explanandum. Treating it separately prevents the framework from conflating having a standpoint with sustaining one well.

It is worth separating two grades of drift, because they fall on opposite sides of the threshold. A system whose internal condition is scattered or rebuilt from scratch at each step does not drift within a standpoint; it fails Condition 1’s requirement of incremental preservation, and so never crosses the threshold at all. The genuinely open case is milder: a system that does maintain a persisting, endogenously updated context—and thus has a standpoint in the present sense—but lacks the selective dynamics needed to hold a

particular content in the foreground over time. Total, chaotic drift is therefore not a counterexample to a standpoint that lacks stability; it is a Condition 1 failure. What the threshold leaves open is the narrower question of how an already-maintained standpoint is stabilized in use.

A final objection concerns the status of the threshold itself. If the three conditions are presented only as necessary, one might worry that the account becomes too weak: it tells us what must be present before standpoint attribution is serious, but not when such attribution is fully warranted. That limitation is intentional. The aim of the account is lower-bound rather than exhaustive. It identifies architectural conditions without which consciousness-oriented inquiry into an artificial system remains structurally premature. Persistent internal context makes the “from” of standpoint structurally meaningful; differentiated internal possibility makes the “somewhere” structurally meaningful; closed-loop operativity makes the attribution explanatorily significant rather than merely descriptive. Whether additional conditions are required for richer forms of standpoint, agency, intentionality, or consciousness remains open.

9. Applying the Framework to Artificial Architectures

The framework is comparative rather than binary, so it is worth asking how several familiar architectural types fare. The question is not overall intelligence or task performance. It is whether the system realizes persistent internal context, whether its latent differentiation amounts to internally occupiable stance-space, and whether current position within that space has cross-role downstream significance.

Three levels of analysis should be kept distinct: the model-internal process itself, such as a bounded inference pass; a persistent application-layer agent that scaffolds the model with tools, memory, or planning loops; and the broader sociotechnical system in which users, interfaces, and institutions also contribute to continuity. The assessments below are made at the first level unless otherwise noted, because apparent persistence at the application or sociotechnical level can obscure the absence of model-internal carryover.

Before turning to cases, it helps to state what would count as evidence for and against each condition. Table 1 sets this out. It is a diagnostic schema, not a measurement procedure, and each entry would require implementation-specific operationalization.

A simple feedback controller satisfies Condition 1 only in a weak and derivative sense. It carries a narrow control variable forward across time, but it lacks differentiated internal possibility: there is no substantive field of stance-like alternatives from which its processing can continue. It also lacks closed-loop operativity in the relevant sense, since its feedback remains tied to one bounded control role. It is therefore a clear non-case.

Standard transformer inference, assessed at the model-internal level, is a bounded near-case. As in the architecture introduced by Vaswani et al. [19], within-pass contextual carryover is real. But at the model-internal level, it is ordinarily episode-bounded and, holding tokenization, architecture, parameters, and deterministic implementation fixed, sequence-sufficient: the relevant internal state is fixed by the supplied token sequence, so the same sequence reproduces the same state. As Section 4 notes, stochastic decoding can vary the output, but not this underlying forward computation. Even setting aside episode-boundedness, it remains unclear whether the resulting dynamics constitute internally occupiable stances rather than a recomputed, token-conditioned trajectory. The question is whether the architecture can remain in, move among, and continue from distinguishable internal positions under its own ordinary dynamics, not merely whether earlier tokens affect later processing. They do, but that is not enough. Standard transformer inference is therefore a near-case not because it lacks internal complexity, but because its relevant carryover remains episode-bound and sequence-sufficient under fixed model-internal conditions.

Table 1. Evidence for and against each condition.

Condition	Evidence That It Is Satisfied	Evidence That It Is Not Satisfied
C1. Persistent internal context	Internal state is carried across transitions and updated by the system’s own activity; varying prior organizational history while holding current input fixed changes current processing; the state remains within the active causal basis of later processing across episodes.	Continuity is reconstructed by re-querying external storage; state is discarded and rebuilt from current input at each step; carryover ends at the episode boundary; the same current input yields the same internal state regardless of prior trajectory.
C2. Differentiated internal possibility	Current position is reachable and maintainable under the system’s own dynamics for more than one step; varying position durably alters the space of live continuations under comparable input; the pattern is robust under admissible perturbation.	Distinctions are analyst-imposed partitions; position alters only the next output, not the continuation space; candidate stances are not reachable or maintainable under ordinary dynamics; differences vanish after a single transition.
C3. Closed-loop operativity	Variation in current position produces systematic differences across more than one bounded role, such as prediction, planning, and action selection; selective intervention can dissociate its effects across those roles.	Downstream effect is confined to one local routine; influence is undifferentiated whole-system gain rather than cross-role regulation; a single scalar control variable suffices to explain the effect.

9.1. An Illustrative Application: A Contemporary GPT-Style Conversational Language-Model System

A contemporary conversational language-model system is the difficult middle case: a large transformer-based model deployed behind a chat interface, with a context window, optional retrieval, optional tool use, and, in some cases, persisted cross-conversation memory. Two caveats matter. First, many such systems are only partially inspectable: their weights, serving infrastructure, retrieval policies, memory mechanisms, and routing procedures may not be open to audit. The assessment must therefore be made at the level of the architecture class rather than as a verified internal analysis of any particular deployed product, so the claims are correspondingly hedged. Second, the level distinction is decisive because many features that make such systems feel continuous operate at the application or sociotechnical level rather than within the model-internal process itself.

At the model-internal level, a single response is produced by an inference pass of the kind already discussed: a bounded near-case. From one conversational turn to the next, apparent continuity is typically sustained by re-supplying the prior transcript, along with any retrieved or remembered material, as input to each new pass. This is Condition 1’s failure mode of external reconstruction. The conversation persists, but as a record reintroduced from outside the model, not as an endogenously maintained internal condition that varies with the system’s own prior processing. Holding the assembled context and model-internal conditions fixed, the relevant internal state is recomputed from that context; the continuity belongs to the transcript and surrounding scaffold, not to a standing internal orientation. Persisted “memory” features sharpen rather than soften this verdict when they store summaries externally and reinsert them as input: they extend capability without converting reconstruction into carryover.

Condition 2 is harder to assess. Such models clearly exhibit input-conditioned internal differentiation, and interpretability research increasingly identifies internal directions that modulate downstream processing. The question is whether these amount to internally occupiable stances rather than recomputed, token-conditioned trajectories. In the relevant sense, a stance would be a position the model reaches and maintains under its own ordinary dynamics across multiple steps, such that varying position durably reshapes the continuation space under comparable input. Activation steering—understood as deliberate

perturbation of internal activation directions to test or alter model behavior—and related causal-intervention methods in interpretability [23,24] provide a preliminary experimental handle on this issue: hold the input fixed, perturb an internal direction, and test whether the downstream continuation space is durably reorganized across multiple steps rather than merely shifted for the next token. The schema therefore points toward possible tests on accessible models. What remains missing is sufficient evidence that such perturbations reveal internally maintainable positions that reorganize continuation space in the required sense.

Condition 3 raises a related but distinct issue. The same internal activations may influence many aspects of downstream generation, which is suggestive. But wide influence is not yet cross-role operativity. The relevant question is whether variation in current position produces differentiated effects across more than one bounded processing role, rather than a broad shift in generation caused by an undifferentiated control-like variable. Disentangling these possibilities would require the selective intervention tests described in Section 6.

The result is a near-case: continuity is largely external, and Condition 2 remains unresolved at the model-internal level. The system can be highly capable and apparently perspectival in conversation, while the model-internal basis for standpoint attribution remains unestablished.

9.2. *Scaffolded, Retrieval-Augmented, and Memory-Based Agents*

When transformer inference is embedded in an agent loop, external memory buffers, retrieval stores, and planning modules can produce substantial continuity at the application layer. But that does not show that the model-internal process itself carries forward its own position within a stance-space. Continuity may be sustained internally or scaffolded by surrounding machinery; the distinction must be drawn at the relevant level of analysis.

Retrieval-augmented and tool-using systems complicate matters further. Retrieval-augmented generation systems can access prior records and supply them to current processing, while ReAct-style agents interleave reasoning traces with tool-directed action [25,26]. Such systems can maintain longer-horizon tasks and make prior information available across episodes. But unless that information becomes part of a persisting internally sustained condition, rather than being fetched and presented to an otherwise episodic process, these systems extend memory and capability without establishing internal standpoint. The issue is not whether prior information is available, but whether it is incorporated into a standing internal condition that continues to shape later processing without being reintroduced as input.

Long-term memory agents and active inference-style architectures are more promising, for different reasons. Generative-agent architectures that store memories, synthesize reflections, and plan over retrieved records illustrate one version of the long-term memory path [27]. Such systems may achieve cross-episode continuity by compressing prior interaction into external summaries and reintroducing them at later steps, in which case the continuity belongs mainly to the scaffold. They would become stronger candidates only if those updates were incorporated into an internally maintained condition that continues to shape processing without external re-presentation.

Active inference agents with persistent generative models come closer in principle because their internal organization can be endogenously updated and operative across prediction, error minimization, and action selection [15,28]. Even here, however, the stronger question remains open: are current implementations merely moving within a fixed learned organization, or does prior traversal actually reorganize the field of future continuations from within ongoing operation?

9.3. World-Model Architectures

More promising still are architectures that maintain latent state across ongoing interaction, use that state in prediction and planning, update it continuously through their own operation, and allow current internal position to constrain multiple downstream functions, such as rollout, policy selection, prediction, and learning. World-model systems in the general sense, learned-model planning systems, and design proposals centered on autonomous world-model-based intelligence are useful comparators here [29–31]. Dreamer-style world-model systems are especially useful exemplars [32,33], but they should be understood as functional illustrations of the relevant organizational profile rather than as uniquely favored architectures or confirmed positive cases.

With respect to Condition 1, such systems come much closer: maintained latent state persists across ongoing interaction, is updated through the system’s own activity, and remains causally active in later processing rather than being reconstructed from present input. With respect to Condition 3, they also come much closer: maintained position can influence several downstream functions, including prediction, imagined rollout, policy selection, and learning updates. The relevant test is whether those effects are differentiated across roles, rather than merely the result of a single broad control signal.

The more open question is Condition 2. Does the latent organization constitute a field of internally occupiable stances, or only a sophisticated control manifold? Positive evidence would take a concrete, counterfactual, multi-step form: hold present input fixed, vary the system’s current latent position, and ask whether the resulting differences in downstream processing persist across more than one subsequent step and affect the space of available continuations rather than only the next output. In a world-model architecture, this would show up as follows: from one latent region, certain imagined rollouts remain stable longer, some planning branches are easier to re-enter, and some hypothesis-updating trajectories are harder to displace; from a different region, the pattern differs systematically and durably. If this pattern remains robust under admissible perturbations that vary current position without destroying the architecture, the latent organization is functioning as stance-space rather than merely as a control manifold.

Table 2 summarizes these assessments. “Partial” and “open” mark where the three-condition schema identifies a need for further evidence, and the rightmost column records the resulting classification into clear non-cases, near-cases, and stronger candidates.

Table 2. Case classification across architecture types.

Architecture	C1: Context	C2: Stance-Space	C3: Operativity	Classification
Simple feedback controller	Weak; narrow variable retention	Fails	Fails; one bounded role	Clear non-case
Standard transformer inference, single pass	Weak; intra-pass, episode-bound	Unclear; at most weak	Partial/unclear	Bounded near-case
Conversational LLM, deployed with memory	Largely external reconstruction	Open	Suggestive but undifferentiated	Near-case; continuity mainly external
Retrieval-augmented or tool-using agent	Scaffolded through reintroduced records	Unclear	Possible, depending on architecture	Near-case unless continuity is internalized
Long-term memory agent	Often scaffolded through external summaries	Unclear	Possible, depending on architecture	Near-case to stronger candidate
Active inference agent with persistent generative model	Plausibly satisfied	Open	Plausibly satisfied	Stronger candidate; C2 remains open
Dreamer-style world model	Plausibly satisfied; case-specific	Open; threshold turns on this	Plausibly satisfied; case-specific	Strongest illustrative candidate; threshold turns on C2

These comparisons ultimately distinguish three notions that are easily conflated: local persistence from maintained internal orientation; descriptive geometry from internally occupiable possibility; and narrow modulation from cross-role organizational consequence. Two systems can produce similar outputs in similar contexts while differing substantially in all three respects. This is why the threshold is not a ranking of behavioral sophistication, but a claim about where continuity, internal possibility, and cross-role regulation are actually realized.

10. Conclusions

This paper's argument is deliberately limited. It does not show that a system satisfying the three conditions would have subjective experience, semantic reference, genuine aboutness, or normative standing. It identifies the architectural threshold at which stronger questions about consciousness, selfhood, and intentionality first become structurally serious for artificial systems: the point at which asking what it might be like to be such a system has a determinate architectural target rather than resting only on external interpretation or behavioral appearance.

This changes the evaluation of three debates. For artificial consciousness, the framework reorders inquiry. Indicator properties drawn from consciousness science [2] should be assessed only once the threshold is met, since below it they lack a determinate architectural target. A system can therefore be classified as a clear non-case, near-case, or stronger candidate before any consciousness predicate is applied.

For machine agency, the framework supplies a structural footing that behavioral description alone lacks. Attributions of planning or goal-directedness gain force where there is a maintained position from which planning proceeds, not merely outputs organized as if by a plan. For future cognitive architectures, the conditions function as design targets as well as assessment criteria: they specify how systems would have to be organized to maintain orientation rather than reconstruct it.

The schema in Section 9 gives the account evidential significance without turning it into a behavioral test or metric. Evidence for the threshold requires active carryover that is not exhausted by present input or external reconstruction, a current internal position that changes the live continuation space under admissible operation, and positional differences with downstream significance across more than one bounded processing role. Evidence against the threshold includes apparent continuity that is externally reconstructed, apparent stance-space that is only an analyst-imposed partition, and apparent operativity confined to one local control function. The evidence is architectural: it concerns how processing is organized, not only what outputs are produced.

The account also has constructive and functional implications. If the three conditions identify the minimal architectural form of processing from a maintained internal standpoint, they also indicate how such organization could be built. Memory would have to function as active internal context rather than episodic retrieval; internal differentiation would have to form occupiable stances rather than merely describable regions; and current position would have to regulate multiple downstream roles rather than a single routine.

Such organization may help coordinate prediction, planning, updating, and action selection across time. The reason is structural: the relevant constraints persist in the active causal basis of processing instead of being re-derived at each step, and one maintained position regulates several functions at once instead of requiring them to be separately re-aligned. Standpoint is therefore both a threshold for attribution and a possible source of temporal coherence, adaptive stability, and cross-role coordination in artificial systems. This functional claim is advanced as a motivated conjecture awaiting empirical test, not as a demonstrated result.

The threshold is defended here as necessary; whether it is also sufficient for a fully sustained cognitive standpoint remains open. The account therefore separates the architectural question of whether there is a minimal standpoint from the dynamic question of how such a standpoint is stabilized in use. The assessments offered here are classifications under a schema, not verified internal audits, especially for closed commercial systems. Within those limits, the central conclusion is straightforward. Persistence without stance-space is too weak. Stance-space without operativity is too inert. Operativity without stance-space collapses into control. If artificial systems are ever to process from an internal standpoint, they need, at minimum, this conjunction: persistent context, stance-space, and operativity. Identifying that threshold is a prerequisite for assessing whether any of them already do.

Funding: The APC was funded by the Department of Electrical and Computer Engineering, Technion—Israel Institute of Technology.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The author gratefully acknowledges Ron Meir and Elliot Sprecher for many fruitful discussions and valuable comments and suggestions that helped improve this manuscript.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Shanahan, M. Simulacra as conscious exotica. *Inquiry* **2026**, *69*, 2982–3010. [CrossRef]
2. Butlin, P.; Long, R.; Elmoznino, E.; Bengio, Y.; Birch, J.; Constant, A.; Deane, G.; Fleming, S.M.; Frith, C.; Ji, X.; et al. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv* **2023**, arXiv:2308.08708v3. [CrossRef]
3. Chalmers, D.J. Could a Large Language Model Be Conscious? Boston Review. 2023. Available online: <https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/> (accessed on 13 July 2026).
4. Dennett, D.C. *Consciousness Explained*; Little, Brown: New York, NY, USA, 1991.
5. Husserl, E. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy, First Book*; Kersten, F., Translator; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1983. (Original work published 1913).
6. Nagel, T. What is it like to be a bat? *Philos. Rev.* **1974**, *83*, 435–450. [CrossRef]
7. Metzinger, T. *Being No One: The Self-model Theory of Subjectivity*; MIT Press: Cambridge, MA, USA, 2003. [CrossRef]
8. Clark, A. *Being There: Putting Brain, Body, and World Together Again*; MIT Press: Cambridge, MA, USA, 1997. [CrossRef]
9. van Gelder, T. The dynamical hypothesis in cognitive science. *Behav. Brain Sci.* **1998**, *21*, 615–628. [CrossRef] [PubMed]
10. Block, N. On a confusion about a function of consciousness. *Behav. Brain Sci.* **1995**, *18*, 227–247. [CrossRef]
11. Chalmers, D.J. Facing up to the problem of consciousness. *J. Conscious. Stud.* **1995**, *2*, 200–219.
12. Dehaene, S.; Charles, L.; King, J.-R.; Marti, S. Toward a computational theory of conscious processing. *Curr. Opin. Neurobiol.* **2014**, *25*, 76–84. [CrossRef] [PubMed]
13. Lamme, V.A.F. Towards a true neural stance on consciousness. *Trends Cogn. Sci.* **2006**, *10*, 494–501. [CrossRef] [PubMed]
14. Baars, B.J. *A Cognitive Theory of Consciousness*; Cambridge University Press: Cambridge, UK, 1988.
15. Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **2010**, *11*, 127–138. [CrossRef] [PubMed]
16. Hohwy, J. *The Predictive Mind*; Oxford University Press: Oxford, UK, 2013. [CrossRef]
17. Tononi, G. Consciousness as integrated information: A provisional manifesto. *Biol. Bull.* **2008**, *215*, 216–242. [CrossRef] [PubMed]
18. Merleau-Ponty, M. *Phenomenology of Perception*; Landes, D.A., Translator; Routledge: London, UK, 2012. (Original work published 1945). [CrossRef]
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. [CrossRef]
20. Woodward, J. *Making Things Happen: A Theory of Causal Explanation*; Oxford University Press: Oxford, UK, 2003. [CrossRef]
21. Zahavi, D. *Subjectivity and Selfhood: Investigating the First-person Perspective*; MIT Press: Cambridge, MA, USA, 2005. [CrossRef]
22. Overgaard, M.; Kirkeby-Hinrup, A. A clarification of the conditions under which large language models could be conscious. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1031. [CrossRef]

23. Turner, A.M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J.J.; Mini, U.; MacDiarmid, M. Steering language models with activation engineering. *arXiv* **2023**, arXiv:2308.10248. [[CrossRef](#)]
24. Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. Representation engineering: A top-down approach to AI transparency. *arXiv* **2023**, arXiv:2310.01405. [[CrossRef](#)]
25. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474. [[CrossRef](#)]
26. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations. *arXiv* **2023**, arXiv:2210.03629. [[CrossRef](#)]
27. Park, J.S.; O'Brien, J.C.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco, CA, USA, 29 October–1 November 2023. [[CrossRef](#)]
28. Parr, T.; Pezzulo, G.; Friston, K.J. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*; MIT Press: Cambridge, MA, USA, 2022. [[CrossRef](#)]
29. Ha, D.; Schmidhuber, J. World models. *arXiv* **2018**, arXiv:1803.10122. [[CrossRef](#)]
30. LeCun, Y. A Path Towards Autonomous Machine Intelligence. OpenReview. 2022. Available online: <https://openreview.net/pdf?id=BZ5a1r-kVsf> (accessed on 13 July 2026).
31. Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* **2020**, *588*, 604–609. [[CrossRef](#)] [[PubMed](#)]
32. Hafner, D.; Lillicrap, T.; Ba, J.; Norouzi, M. Dream to control: Learning behaviors by latent imagination. International Conference on Learning Representations. *arXiv* **2020**, arXiv:1912.01603v3. [[CrossRef](#)]
33. Hafner, D.; Pasukonis, J.; Ba, J.; Lillicrap, T. Mastering diverse domains through world models. *arXiv* **2023**, arXiv:2301.04104. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.