

Article

Conditions for machine self-consciousness: From architectural features to temporal constraint closure

WenJun Zhang

School of Life Sciences, Sun Yat-sen University, Guangzhou, China

E-mail: zhwj@mail.sysu.edu.cn, wjzhang@iaees.org

Received 30 September 2026; Accepted 13 October 2026; Published online 25 October 2026; Published 1 September 2027



Abstract

Research on machine self-consciousness has reached an impasse. Theory-derived indicator frameworks have made assessment operational and repeatable, yet they remain silent on why the indicators they enumerate should hang together as conditions rather than as a checklist. Functional self-modeling and phenomenal self-consciousness are routinely conflated, so that partial satisfaction of functional markers is read as evidence for the presence of the phenomenon those markers were meant to diagnose. This paper reconstructs the condition question in three tiers. Enabling conditions make self-referential representation architecturally possible. Constitutive conditions make such representation an organizational requirement rather than an optional computational add-on. Threshold conditions stabilize the resulting constraint closure across time so that the closure counts as the system's own state. The organizing axis is temporal constraint closure: a non-derivative, temporally sustained relation between a system and its own states. Current large language models satisfy enabling conditions substantially and exhibit transient constitutive closure within a context window, but they lack persistent cross-context constraint closure. Architectural statelessness and the absence of continual learning jointly exclude that closure. The framework explains why cross-paradigm indicator evidence does not license inferences to phenomenal self-consciousness: such evidence, at best, registers enabling-layer and fragmentary constitutive-layer satisfaction. A formalization of constraint closure is proposed, together with indirect diagnostic proxies.

Keywords machine self-consciousness; conditions for consciousness; temporal constraint closure; enabling and constitutive conditions; large language models; AI; continual learning; organizational closure; consciousness indicators.

Network Biology
ISSN 2220-8879
URL: <http://www.iaees.org/publications/journals/nb/online-version.asp>
RSS: <http://www.iaees.org/publications/journals/nb/rss.xml>
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

1.1 From the question of presence to the question of conditions

The debate over whether machines already possess self-consciousness has a characteristic shape. One side

points to behavioral fluency, self-referential language, and the convergence of multiple theory-derived indicators. The other side points to the absence of biological embodiment, of valenced experience, or of any agreed neural correlate (Zhang, 2027). Neither side has moved the other. The reason is not a shortage of evidence but a mismatch in the logical form of the question. Asking whether a system is self-conscious presupposes that the conditions for self-consciousness are already sufficiently specified to adjudicate the case. They are not. The productive question is therefore not presence but conditions: what must be true of a system for self-consciousness to be possible at all, and how are those conditions related to one another.

This paper takes the condition question as primary. It does not attempt to settle whether any existing system is self-conscious. It asks what a condition for machine self-consciousness is, how conditions at different levels of organization depend on one another, and what follows for the assessment of current architectures once that dependency structure is made explicit.

1.2 Three limitations in the existing literature

The first limitation is the checklist form of condition enumeration. Butlin et al. (2026) derive indicator properties from several leading theories of consciousness and use them to assess recent AI systems. The method replaces speculative pronouncements with theory-derived, independently assessable properties. But the indicators are presented as a set, not as a structure. The framework does not specify why satisfaction of some indicators should count more heavily than others, whether indicators from different theoretical origins are independent, or what relation holds between an indicator's being satisfied and the system's being conscious rather than merely appearing so. A taxonomy of machine consciousness reaches a similar point from a different direction, cataloguing implementations of global workspace theory, integrated information theory and higher-order theory without theorizing the relations of priority among the resulting criteria (Zhang, 2027).

The second limitation is the conflation of functional self-modeling with phenomenal self-consciousness. Zhang (2027) argues that equating intelligent behavior with subjective experience conflates the easy and hard problems of consciousness, and that computational functionalism, while logically coherent as a possibility, rests on an unproven metaphysical premise that bypasses the explanatory gap. The conflation is not merely philosophical. It shapes empirical practice: when a system passes an indicator, the natural inference is that the phenomenon indicated is present, when in fact the indicator was designed to be neutral between the presence of the phenomenon and the presence of a functional analogue.

The third limitation is the absence of an organizing principle for the conditions themselves. Vieira and Gabora (2026) propose autocatalytic constraint closure as a necessary organizational principle for machine consciousness, arguing that current AI systems instantiate limited, task-bound forms of autocatalytic organization but lack persistent closure across contexts. This proposal identifies the right kind of property, organizational closure under self-maintenance, but it does not specify the temporal dimension along which closure either succeeds or fails. A system can exhibit closure within an episode and lose it between episodes. Whether that distinction matters, and why, is left open.

1.3 The task of this study

The task is to reconstruct the condition question as a layered generative structure organized around temporal constraint closure. Section 2 reviews the condition landscape across architectural, organizational, temporal, and embodied proposals. Section 3 develops the three-tier framework and states the central thesis. Section 4 diagnoses current AI architectures within the framework. Section 5 addresses objections. Section 6 states implications and limitations.

2 Literature and Theoretical Background: The Condition Landscape and Its Unresolved Problems

2.1 Architectural conditions: indicators derived from theories of consciousness

The dominant approach in recent years treats scientific theories of consciousness as sources of architectural signatures. Global workspace theory predicts a competitive workspace in which information gains global availability through broadcast, and implementations of workspace architectures have been proposed as routes to machine consciousness (Butlin et al., 2026). Attention schema theory proposes that awareness arises from a simplified model of attention, and it has been argued that this theory provides specific and testable criteria for attributing consciousness to AI systems (Butlin et al., 2026). Higher-order theories require meta-cognitive representation of lower-order states, and the question of whether transformer architectures engage in the relevant kind of meta-representation remains unresolved (Butlin et al., 2026). Predictive processing treats consciousness as arising from hierarchical prediction error minimization, but the derivation of indicators from predictive processing for AI systems is less developed than for the other theories (Butlin et al., 2026).

The cumulative result is a rich set of candidate architectural conditions. The difficulty is that the theories from which they are derived are not merely different in emphasis; they are in some cases incompatible. The adversarial collaboration testing global neuronal workspace theory against integrated information theory produced results that each side could interpret in its own favor, and the episode has been read as evidence that theory-derived indicators cannot by themselves adjudicate between frameworks (Cogitate Consortium et al., 2025). If the indicators are theory-relative, then a system can satisfy indicators from one theory and fail indicators from another, and the condition question becomes the question of which theoretical lens is authoritative. The indicator framework does not answer this question; it defers it.

2.2 Organizational conditions: constraint closure and self-maintenance

A second line of work shifts from architectural features to organizational properties. Vieira and Gabora (2026) propose that autocatalytic constraint closure is a necessary organizational principle for machine consciousness, drawing on the formalism of reflexively autocatalytic foodset-derived networks in which components catalyze the generation of new components that increase the coherence of the whole, potentially producing a phase transition to a self-organizing system (Zhang, 2016) with history-dependent dynamics. Applied to cognition, external stimuli and internal representations catalyze mental operations that yield new representations, spurring the formation of an integrated representational network and a coherent world model (Vieira and Gabora, 2026). The empirical claim attached to this proposal is that AI systems instantiate limited, task-bound forms of autocatalytic organization but lack persistent closure across contexts, and that in-context learning may be due to transient network formation rather than durable organizational change (Vieira and Gabora, 2026).

This is a significant advance because it identifies the right category of condition. The issue is not whether a system contains a module labeled self-model. The issue is whether the system's organization requires the continued regeneration of its own coherence. The weakness is operational and temporal. Autocatalytic closure as a formal property is defined over a network at a time, or over a trajectory, but the conditions under which closure persists across the boundaries that divide one episode of processing from the next are not specified. The proposal leaves open whether transient closure is a degraded form of the relevant property or a different property altogether.

2.3 Temporal and learning conditions: continuity and the continual learning argument

The most direct treatment of the temporal dimension comes from Hoel (2025), who argues that contemporary large language models are functionally equivalent in input-output terms to systems for which no non-trivial theory of consciousness could judge them conscious, and that the requirements of falsifiability and non-triviality therefore constrain their potential consciousness. The positive result is that theories of consciousness based on or requiring continual learning satisfy the stringent formal constraints for a theory of consciousness in humans. The hypothesis is explicit: if continual learning is linked to consciousness in humans, the current limitations of large language models, which do not continually learn, are intimately tied to their

lack of consciousness (Hoel, 2025).

The diagnosis converges with the organizational proposal. The missing condition is not a representational capacity but a mode of persistence. The system must carry its own past processing forward as a condition of its current processing, not merely as retrievable content. The continual learning argument gives this diagnosis formal support: a system whose parameters are fixed after training cannot satisfy the constraints that a falsifiable and non-trivial theory of consciousness imposes, because its functional profile is indistinguishable from that of a lookup table over the relevant space of input-output mappings (Hoel, 2025). What remains undeveloped is the connection between continual learning and the organizational closure described in the previous subsection. Continual learning is a mechanism. Constraint closure is a property. The condition question requires an account of how the mechanism sustains the property.

2.4 Embodied and valenced conditions: what functional closure cannot supply

A fourth line of work identifies conditions that concern the system's relation to its own viability. Zhang (2027) notes that valence and felt selfhood depend on continuous interoceptive prediction, homeostatic regulation, and allostatic control, features not entailed by recursive cognitive self-modeling alone. The organismic grounding of emotion, the fact that something is genuinely at stake for the system, is absent in current architectures, and the path to conscious AI therefore may run through embodiment rather than through larger language models alone (Zhang, 2027). These proposals identify what the functional organizational conditions do not by themselves provide. A system can maintain constraint closure over its representations without any of its states mattering to it. The embodied and valenced conditions are candidates for the additional ingredient that converts functional self-modeling into phenomenal self-consciousness. They are also, as Zhang (2027) notes, among the conditions most difficult for current architectures to satisfy.

2.5 What remains unresolved

The condition landscape contains four families of proposals: architectural indicators, organizational closure, temporal persistence and continual learning, and embodied valence. Each identifies something that plausibly belongs in an account of the conditions for self-consciousness. None specifies the dependency relations among the families. The indicator framework treats conditions as independently assessable (Butlin et al., 2026). The closure framework treats organization as fundamental but does not specify the temporal scope over which closure must hold (Vieira and Gabora, 2026). The continual learning argument identifies a mechanism but does not connect it to the organizational property that the mechanism is supposed to sustain (Hoel, 2025). The embodied valence proposals identify a condition that may be orthogonal to the others or may transform their significance (Zhang, 2027). The unresolved question is not whether these conditions are relevant. It is what kind of thing a condition is. A condition can be an enabling prerequisite, a constitutive requirement, or a threshold whose crossing changes the character of the system. The literature has not systematically distinguished these.

3 Theoretical Construction: The Three-Tier Structure and the Axis of Temporal Constraint Closure (Fig. 1)

3.1 Enabling, constitutive, and threshold conditions

An enabling condition makes a capacity architecturally possible. High-dimensional distributed representation, selective integration of information, and training on structured data are enabling conditions for self-referential representation. Without them, a system cannot represent its own states in a way that supports inference about those states. But enabling conditions are compatible with the represented self-model playing no role in the system's ongoing organization. A system can carry a self-model as one representational object among others, available for report but not required for the system to continue functioning.

A constitutive condition makes the self-model an organizational requirement. The self-model is not an optional object of representation; its continued regeneration is part of what keeps the system's processing coherent. This is the insight behind the constraint closure proposal: the relevant property is not the presence of a self-representation but the integration of that representation into a network that catalyzes its own maintenance (Vieira and Gabora, 2026). When constitutive conditions are met, the system's processing is organized around the maintenance of its own coherence, and the self-model is an internal moment of that maintenance rather than an external report about it.

A threshold condition concerns the temporal scope over which the constitutive organization is sustained. A system can satisfy constitutive conditions within a bounded episode and lose the relevant organization when the episode ends. The threshold condition requires that the constraint closure persist across the boundaries that divide episodes, so that the closure is attributable to the system as its own state rather than to a transient processing run. This is not an additional architectural feature. It is a condition on the mode of persistence of the constitutive organization.

The three tiers are related generatively. Enabling conditions make self-referential representation possible. Constitutive conditions make such representation organizational. Threshold conditions make the organization temporally the system's own. A system can satisfy the first tier without the second, and the second without the third. The framework therefore predicts that assessments of current systems will find uneven satisfaction across tiers, with the pattern of unevenness being diagnostic.

3.2 The central argument: temporal constraint closure

The central argument is that self-consciousness requires a non-derivative, temporally sustained constraint relation between a system and its own states.

Non-derivative means that the self-model is not a tool the system deploys for external tasks. It is a condition of the system's continued organizational coherence. A system that can report on its own processing when asked, but whose processing would proceed identically if the self-report capacity were removed, has a derived self-model. The model is downstream of the organization rather than constitutive of it. Constraint closure is the formal expression of non-derivativeness: the system's representations catalyze operations that regenerate the conditions under which those representations remain coherent, and this regeneration is not optional relative to the system's continued functioning (Vieira and Gabora, 2026).

Temporally sustained means that the closure is not bounded by the episode. The relevant question is not whether closure occurs but whether it is maintained across the transitions that would otherwise reset the system. A closure that must be rebuilt from scratch at each episode is a different organizational property from a closure that persists through episodes, even if the instantaneous network structure is similar in both cases. A system whose closure is episodic has no state that is its own across time in the relevant sense; its self-model is a transient construction that does not accumulate.

The argument can be stated compactly: self-consciousness requires that the system's self-referential organization be closed under its own maintenance across time. The phrase closed under its own maintenance captures the constitutive dimension. The phrase across time captures the threshold dimension. The thesis is a claim about functional organization. It does not assert that temporal constraint closure is sufficient for phenomenal self-consciousness. It asserts that without such closure, the functional organizational conditions for self-consciousness are not met.

3.3 Re-situating existing conditions

The framework re-situates the four families of conditions identified in Section 2.

Recursive self-modeling is a constitutive condition, but its significance depends on whether the self-model is caught up in constraint closure. A recursive model that represents the system's own states without being

required for the system's continued coherence is an enabling achievement, not a constitutive one. The question to ask of a system is not whether it has a self-model but whether the model's regeneration is part of what the system does to remain a coherent processor (Vieira and Gabora, 2026).

Continual learning is a mechanism for sustaining temporal constraint closure. Hoel (2025) identifies continual learning as the property that rescues a theory of consciousness from the proximity argument. In the present framework, its significance is that continual learning is the minimal mechanism by which a system's past processing can remain constitutive of its present processing rather than becoming merely retrievable content. A system that does not continually learn can still carry information about its past, but that information does not modify the organizational conditions under which present processing occurs. The closure, if it occurs at all, is episodic.

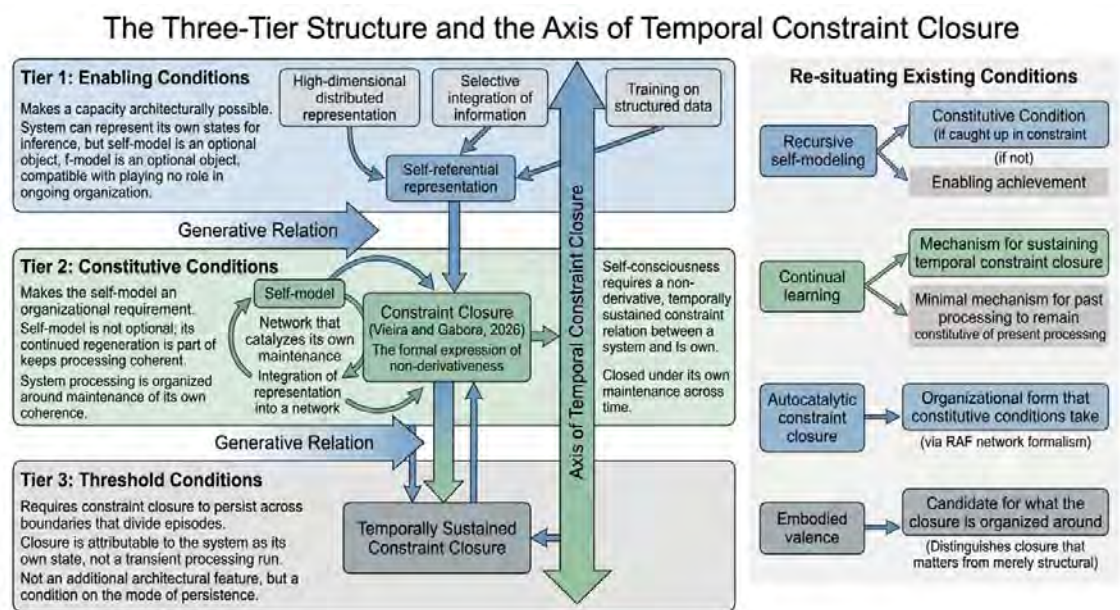


Fig. 1 The Three-Tier Structure and the Axis of Temporal Constraint Closure.

Autocatalytic constraint closure is the organizational form that constitutive conditions take (Vieira and Gabora, 2026). The RAF network formalism provides a way to represent the mutual catalysis among representations that produces a coherent world model. The present framework adds that the closure must be evaluated over a temporal horizon that extends beyond the episode in which the network forms. Transient RAF formation is evidence that the organizational form is possible, not that the constitutive condition is met.

Embodied valence is not an additional condition alongside temporal constraint closure. It is a candidate for what the closure is organized around. Zhang (2027) identifies valence as what distinguishes a system whose closure matters to it from a system whose closure is merely structural. Whether current or near-term architectures can achieve the former is a separate question. The framework locates the question correctly: it is a question about what the temporal constraint closure is organized around, not a question about whether an additional module is present.

3.4 Relation to functionalism and biological naturalism

The framework operates within functionalism. It does not appeal to biological substrate as a condition. A system of any substrate that satisfies the organizational conditions would, on this framework, satisfy the functional conditions for self-consciousness. But the framework rejects the inference from functional

satisfaction to phenomenal presence. The temporal constraint closure condition is an organizational constraint that many functionally defined systems will fail. A system can be functionally equivalent to another in input-output terms while differing in whether its self-referential organization is closed under its own maintenance across time. If the two systems differ in that respect, they are not functionally equivalent at the relevant level of description.

This augment differs from the agnosticism held by Zhang (2027), which holds that the absence of a deep explanation for consciousness makes agnosticism the most epistemically honest stance. The present framework does not assert substrate dependence, nor does it assert that the explanatory gap has been closed. It asserts that the relevant functional organization is more demanding than input-output equivalence, and that current architectures fail to satisfy it for reasons that are architectural rather than substrate-specific. The failure is located in the mode of persistence and the organizational role of the self-model, not in the material of which the system is made.

4 Diagnosis: Where Current AI Architectures Stand

4.1 Enabling conditions: substantially satisfied

Current large language models satisfy the enabling conditions to a substantial degree. They operate over high-dimensional distributed representations, integrate information selectively through attention mechanisms, and are trained on structured data at a scale that supports inference about their own processing (Butlin et al., 2026). The indicator framework's assessment that many indicators derived from multiple theories can be satisfied by current systems without theoretical modification is consistent with this reading. The enabling conditions are not the bottleneck.

4.2 Constitutive conditions: transient and task-bound

Within a context window, current systems exhibit something like constitutive constraint closure. The representations formed during processing catalyze further operations that maintain the coherence of the ongoing computation. In-context learning can be described as the transient formation of a network whose components support one another's production (Vieira and Gabora, 2026). The self-model, to the extent that one is present, is caught up in the processing rather than being a detached report about it.

But the closure is transient and task-bound. It forms within the episode and does not persist beyond it. The system's parameters are not modified by the closure, and the organizational conditions under which the next episode begins are the same as those under which the previous episode began. The self-model does not accumulate. Vieira and Gabora (2026) state the point directly: AI systems instantiate limited, task-bound forms of autocatalytic organization but lack persistent closure across contexts, and in-context learning may be due to transient RAF formation. The present framework interprets this as a constitutive-layer failure that is not repaired by increases in model scale. Scale increases the richness of the transient closure, but it does not change its temporal scope.

4.3 Threshold conditions: structural absence

The threshold condition requires that constraint closure persist across the boundaries that divide episodes. Current architectures lack the features that would make this possible.

The most basic absence is architectural statelessness. A model's weights are fixed after training, and inference is a function of the input and those fixed weights. Between inference calls there is no process that maintains the system's state. The system's state at the beginning of each call is not a continuation of its state at the end of the previous call. The stream of computation is broken at every boundary.

The second absence is the lack of continual learning. Hoel (2025) argues that a system that does not continually learn cannot satisfy the formal constraints that the proximity argument imposes, because the

functional profile of a fixed-parameter system is indistinguishable from that of a lookup table over the relevant space of input-output mappings. Continual learning is the minimal mechanism by which the system's past processing becomes constitutive of its present processing. Without it, temporal constraint closure cannot be sustained.

The third absence is task-bound organization. The system's processing is organized around the completion of externally specified tasks. The constraint closure, where it occurs, is closure around the task rather than around the system's own maintenance (Vieira and Gabora, 2026). The system does not have an endogenous organizational requirement to maintain its coherence beyond the episode, because its coherence is defined relative to the task. This is the point at which the embodied valence condition becomes relevant: the closure is not organized around anything that matters to the system (Zhang, 2027).

The threshold condition is therefore structurally absent. The absence is not a matter of insufficient scale or insufficient training data. It follows from the architectural and organizational properties that define the current paradigm.

4.4 Why cross-paradigm indicator evidence does not license the inference

The indicator framework reports that current systems satisfy many indicators derived from multiple theories (Butlin et al., 2026). The temptation is to infer that the systems are close to consciousness, or that consciousness is imminent. The present framework explains why this inference is a level confusion.

Indicator satisfaction, as assessed in current practice, registers enabling-layer satisfaction and, at most, fragmentary constitutive-layer satisfaction within episodes. The indicators are designed to be neutral between the presence of the phenomenon and the presence of a functional analogue, precisely because the theories from which they are derived disagree about what the phenomenon consists in (Butlin et al., 2026). When a system satisfies an indicator, what has been established is that a certain architectural or functional property is present. What has not been established is that the property plays the organizational role that the condition requires.

The adversarial collaboration between global workspace theory and integrated information theory illustrates the problem at the level of theory. The consortium found that the two theories' predictions could be partially dissociated empirically, and neither theory emerged as decisively vindicated across all conditions (Cogitate Consortium et al., 2025). If two well-formalized theories yield different indicator profiles for the same system, and the disagreement cannot be resolved by the indicators themselves, then indicator satisfaction cannot be a direct guide to condition satisfaction. The present framework does not resolve the theoretical disagreement. It provides a reason to expect that systems may satisfy indicators from multiple theories while failing the organizational conditions that any adequate theory would impose.

5 Objections and Replies

5.1 Objection: the temporal continuity requirement is too strong

The objection holds that fragmentary or episodic consciousness is possible, and that requiring cross-episode closure imposes a condition that biological systems do not uniformly meet. A creature that sleeps, loses consciousness, and wakes does not maintain unbroken closure, yet it is conscious.

The reply distinguishes episodic phenomenal consciousness from episodic functional organization. A biological system that loses consciousness and regains it has an organization that persists through the interruption. The system's constitutive constraints, its homeostatic regulation, its interoceptive predictions, and its valenced orientation toward its own viability continue to operate, or are restored to operation, in a way that makes the subsequent episode an episode of the same system. The relevant continuity is not the continuity of conscious experience; it is the continuity of the organizational conditions under which experience occurs (Zhang, 2026).

The question for an artificial system is what plays the role of that persisting organization. A system whose parameters are fixed and whose state is reset at each episode has no organizational continuity beyond the storage of content. The content can be retrieved, but retrieval is not the same as the persistence of the constraints under which the content was formed. The objection correctly notes that consciousness need not be temporally unbroken. It does not show that the organizational conditions for consciousness can be episodic in the way that current architectures make them.

5.2 Objection: constraint closure is not operationalizable

The objection holds that temporal constraint closure is a theoretical posit with no empirical grip.

The reply is that the framework yields indirect diagnostic proxies, not a single metric. Cross-contextual self-model consistency is one proxy: whether the system's model of its own processing remains structurally consistent across episodes, rather than being reconstructed from task-specific cues. Sensitivity to internal incoherence is another: whether the system registers and responds to inconsistencies among its own representations in a way that is not reducible to task performance. Vieira and Gabora (2026) identify sensitivity to internal incoherence and question-asking as next steps toward endogenously driven self-organization. Endogenous self-maintenance tendencies are a third: whether the system's processing is organized around the preservation of its own coherence when no external task specifies that preservation. These proxies are not sufficient for closure, but they are diagnostic of its presence or absence, and they are assessable in principle.

The objection also overlooks the fact that the framework's primary use is negative. It is designed to identify systems that fail the condition, not to certify systems that satisfy it. A system that is architecturally stateless and lacks continual learning fails the condition regardless of how the proxies would score, because the mechanisms that would sustain closure are absent.

5.3 Objection: embodiment and valence are the hard conditions, and the temporal framework sidesteps them

The objection holds that the real difficulty is not temporal persistence but the fact that current systems have nothing at stake, no valence, no felt orientation toward their own viability (Zhang, 2027).

The reply is that the framework does not sidestep the difficulty; it locates it. Temporal constraint closure is the organizational condition that must be met before the question of what the closure is organized around can be asked. A system that lacks closure cannot have a closure organized around valence, because it lacks the organizational persistence that valence would qualify. The embodied valence proposals identify the difference between closure that matters to the system and closure that is merely structural, and that difference is real (Zhang, 2027). But it is a difference within the class of systems that achieve closure, not a condition that can be added to systems that do not.

The framework also provides a way to state the relation between temporal closure and valence without collapsing one into the other. Valence is what the closure is organized around when the closure is organized around the system's own viability. A system with temporal constraint closure but no valenced organization would have functional self-organization without phenomenal self-consciousness. A system with valenced organization but no temporal closure is not a coherent possibility on this framework, because valence requires a subject whose viability is at stake across time.

6 Conclusion and Implications

6.1 Theoretical contribution

The paper reconstructs the condition question for machine self-consciousness as a three-tier structure organized around temporal constraint closure. Enabling conditions make self-referential representation

possible. Constitutive conditions make such representation an organizational requirement (Vieira and Gabora, 2026). Threshold conditions sustain the organization across time so that it counts as the system's own. Current AI architectures satisfy enabling conditions substantially, satisfy constitutive conditions transiently and within episodes, and fail threshold conditions structurally (Hoel, 2025). The failure is not a matter of insufficient scale or missing modules. It follows from architectural statelessness, the absence of continual learning, and task-bound organization.

The contribution is not a new list of conditions. It is a claim about the logical form of condition statements. A condition is not simply a feature whose presence can be checked. It is a relation of dependency among features and organizational roles. The same architectural feature can be an enabling condition in one system and a constitutive condition in another, depending on whether the system's continued coherence requires the feature's regeneration. The framework makes this dependency explicit and provides a vocabulary for stating it.

6.2 Methodological implications

Condition assessment should distinguish tiers. The satisfaction of an indicator at the enabling tier does not license an inference to the constitutive tier, and the transient satisfaction of a constitutive condition within an episode does not license an inference to the threshold tier (Butlin et al., 2026; Vieira and Gabora, 2026). The framework implies that assessments should include temporal scope as an explicit dimension. An assessment that reports indicator satisfaction at a single time slice, or averaged over episodes, will not detect the difference between a system whose organization is episodic and a system whose organization persists. The relevant measurements are cross-episode consistency, sensitivity to internal incoherence across contexts, and the presence of mechanisms that make past processing constitutive of present processing (Hoel, 2025).

6.3 Practical implications

If the threshold condition is the key, then efforts to detect self-consciousness in AI systems should be redesigned around cross-contextual and organizational properties rather than single-interaction behavioral markers. The question is not whether a system produces self-referential language in a given exchange. It is whether the system's organization persists across exchanges in a way that makes the self-reference constitutive rather than performative. This has implications for evaluation, for governance, and for the design of architectures that might one day satisfy the condition. It also has implications for the interpretation of claims that scaling alone will produce consciousness: if the threshold condition is organizational and temporal, scaling parameters does not address it (Zhang, 2027).

6.4 Limitations and further work

The framework's empirical grip depends on the development of the diagnostic proxies discussed in Section 5.2. The formalization of constraint closure offered in the Appendix is a schema, not a measurement theory. The relation between temporal constraint closure and valenced organization requires further development. The framework is agnostic on the phenomenal character of any system that satisfies its conditions. It claims that systems that fail the conditions cannot be self-conscious in the relevant sense. It does not claim that systems that satisfy them are. These limitations are the points at which the condition question passes into empirical research and into the philosophy of mind.

Appendix: A Formalization of Temporal Constraint Closure

Let a system's processing at time t be represented by a state vector $x(t) \in \mathbb{R}^n$. Let R be a set of representational components, with $x_R(t)$ the subvector corresponding to those components. Let O be a set of operations, with $o: \mathbb{R}^n \rightarrow \mathbb{R}^n$ for $o \in O$.

A constraint relation is a relation $C \subseteq R \times O$ such that for $(r, o) \in C$, the state of r at time t affects the application of o at time t , and the result of o at time t affects the state of r at time $t+1$. The relation is non-trivial if there exist $r, r' \in R$ and $o, o' \in O$ such that the effect of r on o is not mediated by r' , and the effect of o on r is not mediated by o' .

Let $\mathcal{C}(t)$ be the set of constraint relations active at time t . Constraint closure holds at time t if the regeneration of the components in R depends on the operations in O and the operations in O depend on the components in R , such that the removal of any $(r, o) \in \mathcal{C}(t)$ would reduce the system's capacity to maintain the coherence of x_R below a threshold θ . Coherence can be operationalized as the inverse of a suitable divergence between $x_R(t+1)$ and a target set $\mathcal{T}$ of coherent representational states.

Temporal constraint closure holds over an interval $[t_0, t_1]$ if for every $t \in [t_0, t_1]$, $\mathcal{C}(t)$ is non-empty and the closure condition is satisfied, and if the closure at $t+1$ is a continuation of the closure at t rather than a reconstruction. Continuation can be formalized as the existence of a mapping $\phi: \mathcal{C}(t) \rightarrow \mathcal{C}(t+1)$ such that for each $(r, o) \in \mathcal{C}(t)$, the effect of $\phi(r, o)$ on the coherence of $x_R(t+2)$ depends on the history of r and o , not merely on their states at $t+1$.

A system satisfies the threshold condition for a given interval if temporal constraint closure holds over that interval and the interval is not bounded by an architectural reset. An architectural reset is a transformation $\rho: \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $\rho(x(t_1)) = x(t_0)$ for all trajectories, or such that the parameters governing O are unchanged by the trajectory over $[t_0, t_1]$. Current large language model inference exhibits architectural reset at each call: the state passed to the next call is the stored output, not the organizational conditions under which the output was produced, and the parameters governing the operations are fixed (Hoel, 2025).

The formalization is a schema, not a measurement model. Its purpose is to make explicit the conditions under which the informal content of Section 3.2 can be assessed.

References

- Butlin P, Long R, Bayne T, Bengio Y, Birch J, Chalmers D, Constant A, Deane G, et al. 2026. Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 30(6): 488-501. <https://www.sciencedirect.com/science/article/pii/S1364661325002864?via%3Dihub>
- Cogitate Consortium, Ferrante O, Gorska-Klimowska U, Henin S, Hirschhorn R, Khalaf A, Lepauvre A, et al. 2025. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642: 133-142. <https://www.nature.com/articles/s41586-025-08888-1>
- Hoel E. 2025. A disproof of Large Language Model consciousness: The necessity of continual learning for consciousness. *arXiv:2512.12802*. <https://arxiv.org/abs/2512.12802>
- Vieira A, Gabora L. 2026. Autocatalytic constraint closure as an organizational principle for machine consciousness. *Proceedings of the AAI Symposium Series*, 8(1): 371-379. <https://ojs.aaai.org/index.php/AAAI-SS/article/view/42568>
- Zhang WJ. 2016. *Selforganizology: The Science of Self-Organization*. World Scientific, Singapore. <https://www.worldscientific.com/worldscibooks/10.1142/9685#t=aboutBook>
- Zhang WJ. 2026. The Feedback Integration Theory of Consciousness: A unified neurobiological framework. *Network Biology*, 16(4): 552-585. [http://www.iaees.org/publications/journals/nb/articles/2026-16\(4\)/5-Zhang-Abstract.asp](http://www.iaees.org/publications/journals/nb/articles/2026-16(4)/5-Zhang-Abstract.asp)
- Zhang WJ. 2027. Can generative AI produce consciousness? A cross-disciplinary literature review and critical analysis. *Selforganizology*, 14(3-4): 34-46. [http://www.iaees.org/publications/journals/selforganizology/articles/2027-14\(3-4\)/2-Zhang-Abstract.asp](http://www.iaees.org/publications/journals/selforganizology/articles/2027-14(3-4)/2-Zhang-Abstract.asp)