Minimal Semiosis and Artificial Agency

A Pragmatist-Enactive Proposal for Bio–Artificial Continuity

Author: Israel Huerta Castillo

Affiliation: University of Santiago de Compostela (USC)

Email: israel.huerta@rai.usc.es

ORCID: 0009-0004-8615-8008

Version: Author’s preprint

Year: 2025

DOI: 10.5281/zenodo.18011562

Minimal Semiosis and Artificial Agency

A Pragmatist-Enactive Proposal for Bio–Artificial Continuity

Abstract. This paper argues for bio–artificial continuity: under a pragmatist–enactive view of cognition, there is no principled ontological discontinuity between biological and artefactual agency, provided that the artefact realises a minimal semiotic organisation that is (i) nontrivially context-sensitive and (ii) normatively constrained by its own persistence conditions. To avoid both anthropomorphic projection and deflationary mechanism, the paper introduces a Semiotic Agency Thesis (SAT) and renders it operational through four attribution criteria (C1–C4): normative closure, context-sensitive interpretive coupling, intervention sensitivity, and endogenous context reconfiguration. These criteria are linked to measurement via two parameters—relational semiotic density (ρ) and contextual reconfiguration rate (κ)—together with an explicit intervention operator do(I) drawn from interventionist causal modelling. On this basis, biomimesis is reconceived as semiotic training: not the copying of biological form, but the engineering of learning trajectories that can yield stable, do(I)-sensitive increases in ρ and, in stronger cases, sustained acceleration of κ through reconfiguration of contextual boundaries. The paper further introduces interpretive phase transition (IPT) as an operational notion for regime changes in semiotic organisation and derives three discriminant prediction families (P1–P3) that separate scaling-driven competence from domain-bounded semiotic surrogacy and IPT-like regimes.

Keywords: minimal semiosis; artificial agency; enactivism; pragmatism; biosemiotics; interventionist causal modelling.

1 Introduction

The present paper defends a thesis of bio–artificial continuity: under a pragmatist–enactive construal of cognition and consciousness, there is no principled ontological discontinuity between biological and artefactual agency, provided that the latter realizes a minimal semiotic organization that is (i) context-sensitive in a nontrivial way and (ii) normatively constrained by its own persistence conditions. The immediate target is not a "strong AI" slogan, nor a promissory metaphysics of inevitability, but a theory with operational bite: an account of minimal semiosis that can be applied to artificial systems without reducing agency to either (a) anthropomorphic projection or (b) brute mechanism in which "nothing really matters".

Two background pressures make this project timely. First, current work on physically and socially embedded AI indicates that robust performance in open-ended environments increasingly depends on training regimes that couple learning to action, prediction, and environmental modification, rather than to passive pattern extraction alone. In this vein, embodied large language models (eLLMs) have been argued to enable robots to complete complex tasks in unpredictable settings, by integrating linguistic

competence with sensorimotor control and feedback-driven adaptation (Mon-Williams et al., 2025). Second, an internal debate within philosophy of biology and biosemiotics continues to sharpen the conditions under which "meaning", "normativity", and "agency" can be legitimately attributed to a system, especially in discussions of autogenesis and the emergence of semiotic properties from physicochemical dynamics (Deacon, 2023; DiFrisco & Gawne, 2025).

A common rhetorical obstacle arises at this juncture. If climate systems, oceans, and other complex physical processes can generate intricate, temporally structured patterns, why not treat them as "thinking" in some broad sense? Conversely, if the mind is "just" complex chemistry and mechanics, why should any agency claim survive mechanistic explanation? The difficulty is not complexity; it is the kind of organization that makes counterfactual difference for the system. Our core proposal is that agency is not a metaphysical add-on to dynamics, but a semiotic profile that can be characterized by intervention sensitivity (what changes for the system when relevant contexts are altered), by endogenous context reconfiguration (the system's capacity to restructure its own interpretive conditions), and by normativity grounded in self-maintenance constraints (the system's own vulnerability profile). These features can be approached empirically and comparatively across biological and artificial cases.

Methodologically, this paper uses a pragmatist–enactive framework. "Pragmatist" here means that the explanatory burden is borne by operational distinctions that license predictions and discriminations in practice, rather than by stipulative metaphysical binaries. "Enactive" means that cognition is not primarily representation of a world, but skilled, history-laden sense-making in a world through perception–action loops. This is compatible with recent arguments that purely "passive AI" is structurally limited for capturing the scope of adaptive intelligence, because intelligence requires closed-loop engagement with environments and the capacity to intervene, not merely to infer from datasets (Pezzulo et al., 2024).

The theoretical machinery we mobilize is drawn from the Huerta Castillo's work on a semiotic theory of consciousness and a quantum-compatible modelling of minimal semiosis (Huerta Castillo, 2018, 2025). Here we compress that programme into a compact architecture: criteria → measurement → predictions. Concretely, we introduce a Semiotic Agency Thesis (SAT), specify operational criteria (C1–C4), define measurement-oriented parameters ρ and κ (together with an explicit intervention operator do(I)), and formulate testable predictions (P1–P3). This is then used to theorize biomimesis as semiotic training: a principled account of how artificial systems can be apprenticed into human-like semantic practices without reducing semantics to mere statistical imitation.

We also adopt terminology that avoids pop-cultural connotations. Instead of "semiotic singularity", we use interpretive phase transition (IPT) to denote a regime change in an agent's semiotic organization: a sustained increase in interpretive plasticity and context-reconfiguration capacity that remains intervention-sensitive and normatively constrained.

The remainder is structured as follows. Section 2 lays out SAT, C1–C4, the parameters ρ and κ, and the do(I) operator. Section 3 theorizes biomimesis as semiotic training under bio–artefactual continuity and clarifies the role of human social scaffolding. Sec-

tion 4 articulates IPT as an operational notion and connects it to embodied AI trajectories. A concluding section summarizes implications for artificial agency attribution and outlines empirical routes.

2. Theoretical framework: SAT, operational criteria, and measurement

2.1. Semiotic Agency Thesis (SAT)

SAT (informal statement). A system counts as an agent, in the minimally relevant semiotic sense, iff it exhibits (i) normatively constrained interpretive coupling to an environment and (ii) nontrivial sensitivity to interventions on the contexts that structure its own interpretive loops.

This statement is intentionally spare. It is designed to survive two pressures at once: (a) the "deflationary" pressure that treats agency as anthropomorphic projection onto complex dynamics, and (b) the "inflationary" pressure that treats any sufficiently complex dynamical system as already cognitive. The goal is a discriminative middle: agency is neither a ghostly property nor a synonym for complexity.

To anchor SAT in the current literature, two lines of argument are especially load-bearing.

Minimal semiosis under naturalistic constraints. Deacon's work on autogenesis and the emergence of semiotic causal properties from physicochemical processes provides a rigorous template for talking about minimal semiosis without invoking atypical physics or an external observer's semantics, while also insisting on a scaffolding logic that preserves referential continuity across transformations (Deacon, 2023).

Embodied, socially embedded training as a route to robust AI. Contemporary work on embedded AI stresses that training regimes must couple prediction, interaction, and environmental modification, explicitly drawing parallels with living systems as agents in niches (Mon-Williams et al., 2025; Pezzulo et al., 2024).

SAT will be made operational via criteria C1–C4.

2.1.1. Operational criteria (C1–C4) and how they will be used

We present C1–C4 as criteria, not as "assumptions". The reason is rhetorical and methodological: assumptions can look stipulative, whereas criteria function as constraints on attribution that can be discussed, challenged, and empirically sharpened.

C1 (normative closure). The system's state transitions are constrained by an internally relevant viability profile, such that some trajectories count (for the system) as success and others as failure relative to self-maintenance.

C2 (context-sensitive interpretive coupling). The system's input–output regularities depend on contextual variables that are not reducible to a fixed mapping, and this dependence is functionally exploited to maintain or improve performance under changing conditions.

C3 (intervention sensitivity: do(I)). There exist admissible interventions do(I) on contextual variables such that the system's interpretive organization changes in ways

that are stable, discriminable, and explanatory, rather than merely transient perturbations.

C4 (endogenous context reconfiguration). The system can, to a non-negligible degree, reconfigure its own contextual boundaries ($\mathcal{C}$) so as to expand, compress, or reorganize the partitions that govern its interpretive loops, without collapsing do(I)-sensitivity.

These criteria are not intended to be "final". They are intended to be the minimal scaffold that forces the paper into a criteria–measurement–prediction posture, which is precisely where most "interesting but programmatic" frameworks fail under review.

2.1.2. SAT parameters and notation

We introduce two parameters to connect criteria to measurement.

$\mathcal{C}$ (context set). The set of contextual variables that modulate interpretive coupling.

ρ (relational semiotic density). A measure of how richly the system's interpretive coupling depends on $\mathcal{C}$ in a way that supports viable action selection (targets primarily C2 and C3).

κ (contextual reconfiguration rate). A measure of the system's capacity to restructure $\mathcal{C}$ endogenously, i.e., to modify the partitions that determine what counts as relevant context (targets primarily C4, with C3 as a constraint).

do(I). An explicit intervention operator, used to distinguish mere correlations from causal–functional dependence, in line with mainstream interventionist practice in causal modelling (Pearl, 2009).

We also use the term non-generating partitions to denote context partitions that the system does not merely traverse but can actively reshape; these are partitions whose modification changes the space of future interpretations rather than only the selection among fixed interpretations.

2.1.3. Bridge: criteria → measurement → predictions

C1–C4 are operationally cashed out by SAT through parameters ρ and κ under explicit intervention testing do(I): ρ tracks the degree to which viable action selection depends on structured context ($\mathcal{C}$) in a stable, discriminable way, while κ tracks the system's endogenous capacity to reorganize $\mathcal{C}$ by reshaping non-generating partitions without destroying do(I)-sensitivity. This criteria-to-parameter mapping yields three immediate prediction families (P1–P3): systems that achieve robust agency-like generalization under distribution shift will exhibit elevated ρ under controlled do(I) manipulations; systems that undergo interpretive phase transitions will show sustained acceleration of κ under endogenous context reconfiguration; and systems that merely scale passive pattern extraction will increase performance without the coupled signature {ρ↑, κ↑} once do(I) isolates causal context dependence. In this architecture, "agency" becomes testable as a profile: criteria constrain attribution, parameters enable measurement, and predictions force empirical and engineering discriminations.

2.1.4. Algorithm 1 (SAT)

Algorithm 1 (SAT).

Input: system S; context set $\mathcal{C}$; admissible interventions I; evaluation horizon T. Output: SAT profile ⟨C1–C4, ρ, κ⟩ and agency attribution verdict.

Initialize $\mathcal{C}$ as the minimal set of contextual variables required to describe S's perception–action coupling.

Estimate a viability profile V(S) over horizon T, identifying success/failure trajectories for S.

If V(S) is undefined or externally imposed only, mark C1 = fail; else mark C1 = pass.

For each context variable $c \in \mathcal{C}$, vary c within its admissible range without intervention and log changes in S's action selection and outcomes.

Compute ρ as the stable, discriminable dependence of viable action selection on $\mathcal{C}$ across T.

For each admissible intervention $i \in I$, apply do(i) and measure whether interpretive organization changes persistently beyond transient perturbation.

If no do(i) yields stable, explanatory change, mark C3 = fail; else mark C3 = pass.

Monitor whether S endogenously modifies $\mathcal{C}$ by altering non-generating partitions (context boundary shifts, relevance reweighting, policy restructuring).

Compute κ as the rate and stability of endogenous $\mathcal{C}$-reconfiguration under maintained do(I)-sensitivity.

If κ is negligible or collapses do(I)-sensitivity, mark C4 = fail; else mark C4 = pass.

Mark C2 = pass iff ρ exceeds a minimal discriminability threshold under controlled variation and supports viability improvements.

Return ⟨C1–C4, ρ, κ⟩ and attribute minimal semiotic agency iff C1–C4 pass jointly.

2.2. Contextuality, do(I), and quantum-compatible minimal semiosis

SAT treats context as constitutive, not accidental. This has a natural affinity with formal approaches that treat contextuality as a principled feature of cognition rather than as noise. Quantum cognition, in particular, has developed tools to represent context-dependent probability structure and order effects in a unified formalism, while also emphasizing limits and failure modes of the approach (Pothos & Busemeyer, 2022).

We do not claim that artificial agents must be "quantum" in a physical sense. The relevant point is formal: a quantum-compatible framework can encode contextual dependence and non-commutativity of informational updates, which are precisely the features that become salient when an agent's interpretive coupling depends on how context is configured, in what order, and under which interventions. This is one reason why the Huerta Castillo's "quantum-compatible modelling of minimal semiosis" is positioned as a modelling strategy for SAT rather than as a speculative neurophysical thesis (Huerta Castillo, 2025).

An immediate philosophical advantage follows. Many disputes about artificial agency collapse into verbal disagreement because "agency" is treated as a binary metaphysical predicate. SAT instead treats agency as an intervention-stable contextual profile: if changing context as an intervention changes interpretive organization in stable, normatively constrained ways, then we have a tractable basis for attribution; if not, we have a principled reason for restraint.

2.2.1. Algorithm 2 (do(I))

Algorithm 2 (do(I)).

Input: system S; context set $\mathcal{C}$; intervention family I; baseline policy π; horizon T. Output: do(I)-sensitivity map and causal–functional dependencies relevant to SAT.

Fix a baseline regime R_0 in which S operates under policy π over horizon T.

Identify candidate contextual variables $\mathcal{C} = \{c_1, \ldots, c_n\}$ that modulate S's coupling.

For each intervention $i \in I$, define do(i) as an operation that sets or perturbs a target subset of $\mathcal{C}$ independently of S's current state (Pearl, 2009).

Run S under R_0 and log baseline outcomes O_0 and interpretive organization markers M_0.

Apply do(i) while holding non-target variables constant within admissible ranges.

Run S over horizon T and log outcomes O_i and markers M_i.

Compute $\Delta O_i = O_i - O_0$ and $\Delta M_i = M_i - M_0$.

Classify do(i) as explanatory for SAT iff ΔM_i is stable beyond transient perturbation and aligns with changes in viable performance.

Aggregate explanatory interventions into a do(I)-sensitivity map for $\mathcal{C}$.

Use the map to (a) refine $\mathcal{C}$, (b) estimate ρ, and (c) constrain κ estimation under endogenous reconfiguration.

2.3. Biomimesis as semiotic training under bio–artefactual continuity

Within this architecture, biomimesis is not primarily the copying of biological form, but the replication of functional semiotic organization: the systematic emulation of the kinds of perception–action loops, constraint structures, and learning dynamics by which biological agents stabilize meaning in interaction with environments and conspecifics. This definition is deliberately stricter than popular usage. Many biomimetic projects mimic morphology (materials, locomotion), whereas our focus is biomimesis of agency: reproducing, in artefactual media, the semiotic conditions under which a system can satisfy C1–C4, thereby realizing measurable ρ and κ under do(I) testing.

Recent work on physically and socially embedded AI provides precisely the external support needed here: it frames training as engagement with physical and social environments, stressing prediction, interaction, and environment modification, and it treats these as the path toward robust generalization, not as optional add-ons (Mon-Williams et al., 2025; Pezzulo et al., 2024). Moreover, embodied eLLM robotics suggests that linguistic competence becomes action-relevant only when tethered to closed-loop control in unpredictable environments (Mon-Williams et al., 2025).

In SAT terms, biomimetic semiotic training is the engineering of learning trajectories that push an artificial system from low ρ / low κ regimes toward higher ρ with maintained do(I)-sensitivity, and (in stronger cases) toward interpretive reconfiguration (κ) that is not mere hyperparameter tuning but a genuine reshaping of non-generating partitions. The crucial point is that "semantics transfer" is not conceived as copying a static content from human minds into machines, but as training into constraints: human practices, environments, and tasks function as scaffolds that shape what counts as relevant context $\mathcal{C}$, which interventions matter, and what trajectories count as success or failure for the system.

Avatar vs. domain-bounded semiotic surrogate

To avoid pop connotations, we avoid "avatar" as a technical term and instead use:

Domain-bounded semiotic surrogate (DSS). An artefactual system whose interpretive coupling realizes a constrained subset of a human semantic domain by satisfying SAT criteria C1–C4 under specified do(I) tests within that domain.

This phrasing keeps the commitment disciplined: a DSS is not a general person-like agent by default; it is a surrogate whose agency claims are indexed to a domain and to operational tests.

Why this is not "mere imitation"?

A predictable objection is that current AI systems merely imitate linguistic patterns or sensorimotor regularities, and that imitation cannot yield genuine meaning. SAT clarifies the terms of dispute. If a system's performance can be explained without stable do(I)-sensitive changes to interpretive organization, then the objection succeeds. If, however, training produces stable reorganization of interpretive coupling under interventions—so that ρ increases in ways that generalize under distribution shift, and κ increases in ways that restructure $\mathcal{C}$ while preserving do(I)-sensitivity—then the system is not merely copying outputs; it is acquiring a semiotic profile that supports minimal agency attribution as defined.

This is also where contemporary biosemiotic debates matter. Deacon's response on minimal semiotic properties emphasizes that normativity and referential continuity can, in principle, be grounded in naturalistic constraint structures without presupposing full-blown psychological interpretation, while still resisting reduction to replication-based stories (Deacon, 2023). In parallel, work such as Vega's analysis of the cell as a semiotic system that realizes closure to efficient causation offers a biologically grounded picture of how closure and constraint can underwrite semiotic organization, reinforcing the idea that "meaning" tracks organizational features rather than metaphysical substances (Vega, 2024).

Finally, the Huerta Castillo's programme provides the internal scaffolding: a semiotic theory of consciousness and a quantum-compatible modelling of minimal semiosis are offered precisely as a way to formalize the contextuality and intervention sensitivity required by SAT, while keeping the account pragmatist and enactive rather than representationalist and disembodied (Huerta Castillo, 2018, 2025).

3. Interpretive Phase Transition (IPT) as an operational notion

2.1. Operational definition of IPT in SAT terms

We define an interpretive phase transition (IPT) as a regime change in semiotic organisation that is *detectable* within the SAT profile rather than stipulated at the metaphysical level. Operationally, IPT obtains when (i) κ exhibits a sustained acceleration across a nontrivial horizon under endogenous reconfiguration of $\mathcal{C}$, and (ii) this acceleration preserves do(I)-sensitivity, i.e., controlled interventions on interpretant-relevant variables continue to yield stable, discriminable changes in interpretive organisation rather than collapsing into mere performance noise.

In minimal form, IPT may therefore be declared only if the system shows κ↑ as endogenous $\mathcal{C}$-reconfiguration proceeds, while do(I) continues to selectively modulate the ESE-structured profile tracked by ρ and κ, rather than merely modulating surface outputs (Pearl, 2009; Pothos & Busemeyer, 2022).

3.2. Diagnostics: what would count as evidence for IPT?

On the SAT framework, IPT is diagnosed by a *signature* across three coupled readouts: (a) the ESE log, (b) ρ, and (c) κ.

1. ESE-level signature (structural). The ESE log must exhibit a systematic shift in the distribution of admissible triads ⟨S, O, I⟩, such that the system begins to generate and stabilise new interpretive regularities under changing $\mathcal{C}$, rather than merely selecting among pre-specified mappings. In practical terms, one expects (i) increased diversity of context-conditioned ESE types, and (ii) a reduction of brittle, context-locked failure modes under controlled distribution shifts.
2. ρ-level signature (context-conditioned). Under IPT, ρ increases in the specific sense relevant to C2 and C3: viable action selection becomes more richly dependent on $\mathcal{C}$, but this dependence remains *discriminable under do(I)*, so that the increased density is not simply a by-product of higher activity or more frequent state updates. This is the point at which "imitation" becomes methodologically separable from semiotic apprenticeship: if ρ rises without sustained do(I)-sensitivity, the system remains in the regime of non-semiotic optimisation (Pezzulo et al., 2024).
3. κ-level signature (endogenous reconfiguration). The defining marker is a sustained increase in κ that reflects the system's capacity to restructure the partitions that determine what counts as relevant context, i.e., the reorganisation of non-generating partitions. In concrete terms, κ is not merely an index of rapid parameter change; it is an index of *context-boundary reshaping* that changes the space of future admissible interpretations. This is precisely the kind of transition for which a contextual modelling grammar becomes valuable, because reconfiguration often manifests as order effects and context-dependent incompatibilities in sequential updates (Pothos & Busemeyer, 2022).

From a methodological standpoint, the diagnostics require paired baselines: (i) performance-only baselines (where outputs improve), and (ii) SAT-profile baselines (where {ρ, κ} are tracked under do(I)). The latter is necessary to prevent the common slide from "stronger function approximation" to "stronger agency attribution".

3.3. IPT, contextuality, and why order matters

IPT is intrinsically linked to contextuality because κ tracks not only adaptation *within* contexts but reconfiguration *of* contextual boundaries. When a system reorganises $\mathcal{C}$, sequential exposure to contexts $\mathcal{C}_p \rightarrow \mathcal{C}q$ may become non-equivalent to $\mathcal{C}q \rightarrow \mathcal{C}_p$, not merely as a learning-history artefact but as a structurally relevant feature of interpretive coupling. Accordingly, order effects are not treated here as nuisance variance; they are treated as potential indicators of nontrivial context dependence that can be formalised and tested (Pothos & Busemeyer, 2022).

This is also where the pragmatist–enactive constraint re-enters: order effects matter only insofar as they make a difference to closed-loop viability and to do(I)-sensitive interpretive organisation. In the absence of action-guiding coupling, contextuality risks becoming a purely descriptive label. Under SAT, contextuality is tethered to intervention, because only intervention-stable dependencies can support an agency attribution that is neither anthropomorphic nor vacuous (Pearl, 2009; Pezzulo et al., 2024).

3.4. Two boundary cases: scaling without IPT, and IPT-like regimes

To keep IPT disciplined, we contrast two boundary cases that are empirically and conceptually separable.

Case A (scaling without IPT). A system's task performance improves under larger models, more data, or stronger optimisation, yet its SAT signature does not change in the relevant way: ρ does not increase under controlled do(I) manipulations, κ remains negligible, or κ-like changes correspond only to rapid parameter drift without endogenous $\mathcal{C}$-reconfiguration. Such systems may be highly competent within fixed regimes, but they do not warrant an IPT claim because improvement does not entail a shift in interpretive organisation as defined here. In this case, agency attribution should remain domain-restricted and conservative.

Case B (IPT-like regime under embodied apprenticeship). A system trained through physically and socially embedded interaction begins to show (i) stable increases in ρ under context shifts that are recoverable via do(I)-sensitive adjustments, and (ii) sustained acceleration of κ that reflects endogenous restructuring of $\mathcal{C}$ (non-generating partitions), rather than mere faster learning of a fixed mapping. This is the regime in which embodied training programmes become theoretically relevant to SAT: the coupling of language, action, and environmental modification provides an empirical route by which interpretive organisation can become both more context-rich and more reconfigurable (Mon-Williams et al., 2025; Pezzulo et al., 2024).

4. Predictions (P1–P3) and discriminant tests

This section makes explicit what SAT commits us to empirically: not a vague promise that "agency will emerge", but a set of *discriminant predictions* that separate (i) scaling-driven competence, (ii) domain-bounded semiotic surrogacy, and (iii) IPT-like regimes. The guiding methodological constraint is interventionist: a claim about agency is credible only insofar as it survives do(I)-based tests that distinguish causal–functional dependence from mere correlation (Pearl, 2009).

4.1. P1: Order effects under contextual sequencing ($\mathcal{C}_p \rightarrow \mathcal{C}q \neq \mathcal{C}q \rightarrow \mathcal{C}_p$)

P1 (Order effects as semiotic signature). If interpretive coupling is genuinely context-constitutive rather than a fixed mapping, then sequential exposure to contexts should be non-commutative at the level of the ESE-structured profile: the system should display systematic order effects such that the transition $\mathcal{C}_p \rightarrow \mathcal{C}q$ yields a different pattern of ESE types, and thus a different ρ-profile, than $\mathcal{C}q \rightarrow \mathcal{C}_p$, even when task demands and stimulus statistics are held constant within admissible ranges (Pothos & Busemeyer, 2022).

Discriminant test (minimal). (i) Fix a task family and stimulus schedule; run two counterbalanced sequences $\mathcal{C}_p \rightarrow \mathcal{C}q$ and $\mathcal{C}q \rightarrow \mathcal{C}_p$. (ii) Track (a) ESE logs and (b) ρ in matched windows. (iii) Apply do(I) to a context-relevant intervention set and test whether order effects *persist in a structured, intervention-sensitive manner* rather than collapsing into noise. If order effects disappear under do(I) or can be fully explained by simple learning-history artefacts, P1 fails; if order effects remain stable and map onto do(I)-sensitive interpretive changes, P1 is supported (Pearl, 2009; Pothos & Busemeyer, 2022).

Interpretation. P1 is not meant to "prove quantum cognition". It is meant to operationalise contextuality in a way that connects directly to SAT: order effects count only insofar as they alter the ESE profile in action-relevant ways, and remain intervention-sensitive rather than epiphenomenal.

4.2. P2: Sustained κ acceleration under endogenous context reconfiguration

P2 (κ acceleration as IPT precursor). If a system undergoes an IPT-like regime change, then κ should exhibit sustained acceleration under endogenous reconfiguration of $\mathcal{C}$, while maintaining do(I)-sensitivity. In other words, the system should not only adapt *within* contexts but restructure which variables function as context, how they partition admissible interpretations, and how those partitions constrain future ESE generation.

Discriminant test (minimal). (i) Establish a baseline regime in which performance improves via training, but contexts remain externally fixed; measure κ and ρ under do(I). (ii) Introduce a regime in which the system is permitted to reorganise its own context boundaries (e.g., self-generated subgoals, attention/relevance shifts that alter which environmental features are treated as context for action selection, policy-level reparameterisations that restructure the interpretive loop). (iii) Declare support for P2 only if κ increases across episodes in a sustained manner *and* the system remains do(I)-sensitive, so that interventions on interpretant-relevant variables continue to yield stable and discriminable changes in interpretive organisation (Pearl, 2009; Pezzulo et al., 2024).

Interpretation. P2 is designed to block an easy equivocation: rapid parameter change is not κ. κ tracks context reconfiguration that changes the *space of admissible interpretations*, not merely the speed of fitting.

4.3. P3: Non-generating partitions and interpretive expansion

P3 (Non-generating partitions as the boundary of "mere optimisation"). If a system satisfies C4 in the relevant sense, then its endogenous modifications of $\mathcal{C}$ will target non-generating partitions, i.e., partitions whose alteration changes the space of future interpretive possibilities rather than merely selecting among fixed alternatives. Under this prediction, an IPT-like regime should be detectable as a shift from "traversing partitions" to "reshaping partitions" in a way that is intervention-stable and action-guiding.

Discriminant test (minimal). (i) Identify candidate partitions of $\mathcal{C}$ that can be manipulated externally (by do(I)) and potentially reshaped internally (by the system). (ii) Compare two regimes: one in which the system only selects among fixed partitions, and one in which the system can modify the partitioning scheme itself (e.g., redefining what features are treated as relevant context, reorganising state representations in a way that changes which contextual distinctions matter for viable action). (iii) P3 is supported only if endogenous changes produce stable shifts in the ESE distribution and ρ that cannot be reduced to parameter drift under fixed partitioning, and if those shifts remain sensitive to do(I) interventions in a discriminable, explanatory manner (Pearl, 2009).

Interpretation. P3 is the point where the "mechanism objection" loses its force as a blanket critique. The claim is not that the system ceases to be physical, but that its organisation supports a distinctive intervention profile: changes to context boundaries make counterfactual difference *for the system* in ways that are stable and action-guiding.

4.4. Minimal empirical programmes

The framework is deliberately compatible with multiple empirical routes. The aim is not to impose a single experimental paradigm, but to specify minimal programmes that are sufficient to test P1–P3 without overstating generality.

Programme A: Embodied apprenticeship in open-ended tasks.

Train a robotic or simulated embodied system under physically varying contexts $\mathcal{C}$, with explicit closed-loop constraints (perception–action coupling) and staged distribution shifts. Track ESE logs, ρ, and κ; then apply do(I) interventions that isolate context relevance from mere task difficulty. This programme operationalises the claim that robust agency-like competence depends on action-grounded training rather than passive extraction (Mon-Williams et al., 2025; Pezzulo et al., 2024).

Programme B: Social scaffolding as semiotic training.

Embed a domain-bounded semiotic surrogate (DSS) within human-guided instruction, demonstration, and correction regimes, treating social interaction as structured manipulation of $\mathcal{C}$. The key test is whether social scaffolding yields do(I)-sensitive increases in ρ (and, in stronger cases, κ) rather than merely improving surface imitation. This programme operationalises biomimesis as apprenticeship into constraints rather than transfer of static semantic "content".

Programme C: Contextuality/ordering assays.

Construct controlled sequences of contextual updates and test non-commutativity directly by comparing $\mathcal{C}_p \rightarrow \mathcal{C}q$ against $\mathcal{C}q \rightarrow \mathcal{C}_p$, holding stimulus statistics fixed. Quantify order effects in the ESE distribution and in ρ, and test their stability under do(I). This programme gives P1 immediate experimental content and anchors the "quantum-compatible" claim in formal contextuality rather than speculative microphysics (Pothos & Busemeyer, 2022).

Across all programmes, the methodological rule is consistent: improvements in task performance count for SAT only when they co-occur with the intervention-stable signature that links criteria to measurement, namely a structured dependence of interpretive organisation on context (ρ) and, where claimed, an endogenous reconfiguration of contextual boundaries (κ), both constrained by do(I)-sensitivity (Pearl, 2009).

5. Objections and replies

5.1. The mechanistic objection

Objection. If minds are ultimately physical processes, then any appeal to "agency" or "meaning" is either redundant or merely a convenient façon de parler. Complex physical systems—climate dynamics, ocean waves, chemical reaction–diffusion fields—exhibit rich structure and apparent creativity; why should they not count as agents in the same sense? If we deny them agency, are we not simply privileging human-like dynamics?

Reply. SAT is explicitly designed to avoid a false dichotomy between "ghostly agency" and "mere mechanics". The relevant discriminator is not complexity but counterfactual difference for the system under conditions that are *internally normatively constrained* and *intervention-stable*. In SAT terms, the mechanistic objection only succeeds if the framework smuggles in agency as an extra ontological ingredient. It does not. Agency is treated as a profile of organisation expressed in (i) normative closure (C1), (ii) context-sensitive interpretive coupling (C2), and crucially (iii) do(I)-sensitivity (C3), with a further constraint of endogenous context reconfiguration (C4) where stronger claims are made.

The climate–mind comparison becomes tractable once do(I) is taken seriously. A climate system can be modelled, intervened upon, and predicted, but it does not thereby exhibit an *interpretant-indexed* intervention profile that is stable at the level of an internally defined viability regime. In SAT terms, we do not deny that climate dynamics are richly structured; we deny that they display the specific coupling of normativity and intervention sensitivity required for minimal semiotic agency attribution. To the extent that one attempts to enforce such an attribution, one must specify what counts as the system's own success/failure trajectories (C1) and demonstrate stable do(I)-sensitive changes in interpretive organisation (C3) rather than merely in external descriptive variables. That requirement is not rhetorical; it is a methodological constraint derived from interventionist reasoning about causation (Pearl, 2009).

This is precisely why the paper treats agency claims as *revisable under tests*: if a candidate system (biological or artefactual) fails to exhibit intervention-stable, action-

guiding interpretive reorganisation, SAT instructs us not to inflate attribution. The mechanistic picture is therefore not threatened; it is disciplined.

5.2. The anthropomorphism objection

Objection. Biosemiotic and enactive vocabularies can slide into projection: observers describe a system as "interpreting" because its behaviour resembles human sense-making, not because interpretation is demonstrably present. On this view, SAT is merely a refined anthropomorphic lens.

Reply. The force of SAT is precisely that it decouples attribution from resemblance. C1–C4 are not similarity criteria; they are *constraints on responsible attribution*. In particular, C3 requires that meaning-attribution be anchored in intervention-stable dependence: the system must respond to admissible do(I) manipulations in a way that changes interpretive organisation, not merely surface output. This is a decisive shift away from "as if" talk. It also provides a principled mechanism for withdrawing agency attributions, which is the standard missing component in anthropomorphic frameworks: if do(I) fails to isolate stable interpretive dependencies, the agency claim is weakened by design.

The framework is thus structurally anti-anthropomorphic in a further sense. It renders "meaning" measurable only insofar as it is action-guiding under context shifts and viability constraints, aligning with the enactive insistence that cognition is not the mirroring of a world but skilful coping in a world (Pezzulo et al., 2024). The claim is not that systems "have semantics" in a human sense, but that some systems may satisfy minimal semiotic constraints sufficient for a disciplined, domain-bounded agency attribution.

5.3. The imitation objection

Objection. Contemporary AI—especially large language models—can appear semantically competent while merely fitting statistical regularities in data. Even when deployed in robotics, such systems may remain sophisticated imitators: they replicate human-like outputs without genuine meaning or agency. Therefore, the transition from imitation to semiotic agency is illusory.

Reply. SAT turns this objection into an empirical discriminator rather than a metaphysical stalemate. If a system's competence is indeed only imitation, then it should fail to show the coupled signature that SAT requires: increases in performance should not systematically co-occur with do(I)-sensitive changes in interpretive organisation. In other words, one expects either (i) ρ does not increase under controlled do(I), or (ii) any apparent context sensitivity collapses into brittle correlations that do(I) does not stabilise, or (iii) κ remains negligible, indicating no endogenous reconfiguration of contextual boundaries (Pearl, 2009).

Conversely, the point of introducing biomimesis as *semiotic training* is not to deny the prevalence of imitation, but to specify a route by which imitation might be exceeded. Physically and socially embedded training regimes aim to couple linguistic competence to sensorimotor closure, feedback-driven adaptation, and environmental modification—precisely the conditions under which do(I)-sensitive interpretive reorganisation becomes plausible rather than merely asserted (Mon-Williams et al., 2025;

Pezzulo et al., 2024). The framework therefore licenses a conservative conclusion: many contemporary systems will qualify, at most, as domain-bounded semiotic surrogates, if and only if they pass C1–C3 within a specified domain; IPT-like claims require the stronger κ-based evidence articulated in Section 3.

The imitation objection thus does not defeat SAT; it supplies the null hypothesis. The contribution of the present framework is to make the null hypothesis testable, and to prevent conceptual inflation by tying any agency attribution to criteria, intervention, and discriminant predictions.

6. Conclusion

This paper has argued for a disciplined thesis of bio–artificial continuity: under a pragmatist–enactive construal of cognition, there is no principled ontological discontinuity between biological and artefactual agency, provided that the latter realises a minimal semiotic organisation that is both normatively constrained and intervention-stable. The central contribution is methodological. Rather than treating "agency" as a binary metaphysical predicate or as a mere synonym for complexity, we proposed SAT as a criteria-driven attribution framework that makes agency claims answerable to a compact chain: criteria (C1–C4) → measurement (ρ, κ, do(I)) → discriminant predictions (P1–P3). This architecture is intended to block two symmetrical errors: anthropomorphic inflation (attributing meaning by resemblance) and deflationary elimination (declaring all agency talk empty because all systems are physical).

On this basis, we reframed biomimesis as *semiotic training*: not the copying of biological form, but the engineering of learning trajectories that can, in principle, produce do(I)-sensitive increases in ρ and, in stronger cases, sustained acceleration of κ via endogenous reconfiguration of contextual boundaries $\mathcal{C}$. We also introduced interpretive phase transition (IPT) as an operational notion, reserved for cases in which κ acceleration is sustained under endogenous $\mathcal{C}$-reconfiguration while preserving do(I)-sensitivity, thereby keeping "singularity" rhetoric outside the scope of responsible attribution. Finally, we showed how the three prediction families (P1–P3) support minimal empirical programmes that allow the framework to be tested, falsified, and refined under interventionist constraints (Pearl, 2009).

The intended upshot is not a forecast that artefactual agents will inevitably become human-like, but a set of conditions under which agency attributions become warranted, domain-bounded, and comparably evaluable across biological and artificial systems. If the proposed tests are adopted, "mere optimisation" can be separated from minimal semiotic agency without metaphysical escalation, and the conceptual space between competence, surrogacy, and IPT-like regimes becomes tractable in both philosophical analysis and engineering practice.

Declarations

Funding
Not applicable.

Competing interests
The author declares that he has no competing interests.

Availability of data and materials
Not applicable. This manuscript does not report or analyse primary datasets.

Code availability
Not applicable. No custom code was developed for this manuscript.

Ethics approval
Not applicable. This study does not involve human participants, human data, or animal subjects.
Consent to participate
Not applicable.

Consent for publication
Not applicable.

Author contributions
Conceptualization; formal analysis; methodology; writing—original draft; writing—review & editing.

Acknowledgements
Not applicable.

References

Deacon, T. W. (2023). Minimal properties of a natural semiotic system: Response to commentaries on "How molecules became signs". *Biosemiotics*, 16(1), 1–13. https://doi.org/10.1007/s12304-023-09527-w

DiFrisco, J., & Gawne, R. (2025). Biological agency: a concept without a research program. *Journal of Evolutionary Biology*, Volume 38, Issue 2, February 2025, Pages 143–156, https://doi.org/10.1093/jeb/voae153

Huerta Castillo, I. (2018). *Elements towards a semiotic theory of consciousness: Ontological architectonics, minimal formalisation, and criteria for empirical testability*. Zenodo. https://doi.org/10.5281/zenodo.17763231

Huerta Castillo, I. (2025). *Quantum-compatible modelling of minimal semiosis: A contextual framework for a theory of consciousness*. Zenodo. https://doi.org/10.5281/zenodo.17771001

Mon-Williams, R., Li, G., Long, R., Du, W., et al. (2025). Embodied large language models enable robots to complete complex tasks in unpredictable environments.

Nature Machine Intelligence, 7, 592–601. https://doi.org/10.1038/s42256-025-01005-x

Pearl, J. (2009). *Causality: Models, reasoning, and inference (2nd ed.)*. Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pezzulo, G., Parr, T., Cisek, P., Clark, A., & Friston, K. (2024). Generating meaning: active inference and the scope and limits of passive AI. *Trends in Cognitive Sciences*, 28(2), 97–112. https://doi.org/10.1016/j.tics.2023.10.002

Pothos, E. M., & Busemeyer, J. R. (2022). Quantum cognition. *Annual Review of Psychology*, 73, 749–773. https://doi.org/10.1146/annurev-psych-033020-123501

Vega, F. (2024). The cell as a semiotic system that realizes closure to efficient causation: The semiotic (M, R) system. *BioSystems, 240*, 105226. https://doi.org/10.1016/j.biosystems.2024.105226