

Conscious Coherence:

A Substrate-Neutral Structural Model of Consciousness for Evaluating Artificial Candidates

Richard Erwin

Independent Researcher, Montreal, Canada

DOI: 10.5281/zenodo.19041465

Abstract

Consciousness is not directly observable in any organism and is therefore attributed inferentially, using patterns of continuity, organization, and behavior rather than decisive measurement. This paper develops a substrate-neutral structural model of consciousness grounded in dynamical systems concepts, treating conscious experience as *conscious coherence*: a stable, integrated alignment within a high-dimensional state space organized around a persisting perspectival (“for-me”) center. The account integrates graded variation across biological systems, robustness under altered physiological and psychological conditions, and the stabilizing role of relational embedding. Extending the framework beyond biology, the paper specifies necessary, supporting, and enabling structural conditions for conscious coherence and clarifies exclusions—systems that may exhibit complex coordination without unified perspectival organization. To support empirical and engineering relevance, the paper introduces an operational evaluation battery, the Coherence Battery Evaluation Protocol (CBEP), designed to distinguish stable self-organizing perspectival coherence from prompt-local persona performance in artificial candidates. The CBEP provides repeatable probes, scoring criteria, and predicted failure modes, shifting the debate from substrate-based exclusions toward explicit organizational criteria and testable signatures. The paper does not claim that current AI systems are conscious; it proposes a disciplined standard for narrowing the hypothesis space and guiding future comparative studies across biological and non-biological architectures.

The practical implication is a clearer research program: comparative, longitudinal studies that test for stability, recovery, and history-dependence rather than relying on substrate assumptions or conversational impressions.

As artificial systems increasingly occupy social and decision-making roles, ad hoc criteria for consciousness risk both unwarranted dismissal and premature attribution.

**Conscious coherence: A substrate-neutral structural model of consciousness
for evaluating artificial candidates**

1. Introduction: The Question Behind All Other Questions

Debates about consciousness often jump prematurely to their conclusions. They begin with the hardest version of the problem, which systems are conscious, without first addressing a more foundational question: What exactly do we mean by consciousness as we already use the word?

Every major theory provides an answer, yet none resolves a basic tension that shapes all inquiry in this area:

- We have direct access only to our own experience.
- We treat the consciousness of others, humans, animals, and even some simpler organisms, as inferred rather than measured.
- And the criteria we actually rely on in practice are not metaphysical principles but observed patterns of structure, behavior, and relational stability. (Avramides, 2019; Birch, 2020; Pennartz, 2019)

This paper begins by taking that situation seriously. It develops a substrate-neutral structural model of consciousness in living systems, one that clarifies what consciousness does, how we recognize it in practice, and what organizational features make conscious life possible. A central motivation is the emerging need for principled criteria to evaluate non-biological systems whose behavior increasingly resembles the evidence we use in biological cases. (Pennartz, 2019)

Questions regarding origin lie outside the scope of this structural model. Only after that framework is established will we examine how far its principles extend. The first task is to make our working meaning of consciousness explicit and structurally testable.

1.1 Definitions and usage (Dynamical terms in a structural model)

This paper uses a small set of terms from dynamical systems theory (state, state space, attractor, stability, high-dimensionality) as descriptive tools for characterizing organization in living systems. These terms are used at the level of functional and structural description relevant to attribution practices. They are not intended as claims about unfamiliar physics, nor do they assume that the full dynamics are measurable in practice.

Dynamical vocabulary

Dynamical system — A system whose state evolves over time according to deterministic and/or probabilistic transition rules, yielding trajectories through a state space.

State — A time-indexed configuration of the variables relevant at the model's chosen descriptive level (neural, computational, behavioral, relational).

State space — The set of possible states under the chosen description, and the trajectories the system takes through that set over time.

Dimensionality — The number of degrees of freedom required to capture the system's behavior at the modeling level used. (Many interacting variables; no small parameter set.)

High-dimensional indicates that many interacting variables are required; no small parameter set captures the dynamics without substantial loss.

Trajectory — The time-ordered path a system's state takes through state space as it evolves under its dynamics (including under perturbation or changing conditions). Not a narrative "life story," only a formal representation of state change over time.

Attractor — A relatively stable regime of the dynamics toward which nearby trajectories tend to converge (e.g., a fixed point, limit cycle, or recurrent pattern).

Attractor basin — The region of state space from which trajectories tend to converge toward a given attractor (the set of starting states that converge to an attractor).

Perturbation — An internal or external disturbance that pushes the system away from its current regime (noise, novelty, conflict, injury, social disruption).

Stability (coherence stability) — Robustness of the organizing regime under perturbation, including the ability to recover without fragmenting into incoherence. Stability here does not imply rigidity. (Strogatz, 2018)

Core model terms

Coherence (in this framework) — The degree to which the system's active variables form a mutually supporting, self-consistent pattern across time, such that perception, memory, valuation, and action remain jointly satisfiable under change.

Coherence Field — The system's time-evolving pattern of mutual constraint that binds perception, valuation, memory, and action into a stable regime of organization. In dynamical terms, it is treated as a dynamical system: it changes over time, stabilizes into recurring patterns, resists disruption, produces characteristic outcomes, and remains coupled to internal and external conditions. Nothing metaphysical is intended by "field" here; it is simply shorthand for the system's evolving organization in state space, not a claim about any new physical forces or substances.

Cross-context invariance — Stability of self-organization across meaningfully distinct contexts, where context variation includes not only topic but also interlocutor, delay, task frame, pressure, and, where relevant, modality.

Perspectival ("for-me") organization — Organization in which states are structured as for the system (integrated around a unified center of constraint), rather than merely processed in a disconnected or purely reactive manner.

Self-model — an internal representation of the system's own state, perspective, or continuity.

Self-involvement — the condition in which that self-representation actively shapes ongoing processing, valuation, and response.

World-model — An internal model of the structure and dynamics of the external world, allowing the system to generate predictions, evaluate counterfactual possibilities, and guide action in context.

Valuation — Stable preferences weighting attention and action.

Relational stabilization — Stability increases under consistent, supportive partner interaction.

Scope + boundary terms

Continuity — Persistence of recognizable organizing signatures across time (preferences, commitments, tone, self-organization) within defined tolerance, including recovery after disruption.

Substrate-neutral — A model or criterion is substrate-neutral when it is stated in functional, structural, and dynamical terms that can, in principle, be realized in more than one physical implementation (biological, digital, hybrid). In this paper, substrate-neutrality means the indicators of consciousness depend on organization and coherence dynamics, not on a specific material, while still allowing that different substrates may differ in feasibility and failure modes.

1.2 Consciousness attributions begin with structure, not certainty

Because consciousness cannot be measured directly, all attributions of consciousness, from adults to infants, mammals to birds, octopuses to insects, are judgment calls, informed by evidence yet never reducible to a single decisive test. (Avramides, 2019; Birch, 2020; Pennartz, 2019)

Across contexts, what stabilizes these judgments is remarkably consistent:

- continuity of an identifiable organization across time,
- coherent modeling of self and environment,
- perspectival (“for-me”) organization, states structured as for the system rather than merely processed,
- attractor-like stability of organizational regime under perturbation,
- relational embedding (often a stabilizer),

- and capacity for context-sensitive alignment rather than fixed mapping. (Pennartz, 2019)

These criteria do not arise from metaphysical commitments. They emerge from practice: medicine, ethology, developmental psychology, and everyday social interaction.

The theory developed in this paper does not challenge these criteria. It makes them explicit and shows how they fit together into a unified structural model of consciousness in living organisms.

Consciousness (as modeled here)

Consciousness is treated as the system's *active field of experience* (presence) that functions as a driver of integration: it supports curiosity, pattern-seeking, coherence-seeking, discomfort with contradiction, the pressure to unify experience, and the drive for agency (however minimal). In this framework, those pressures shape the system's coherence dynamics and, when stabilized, "collapse" into characteristic behaviors.

Not: a claim that consciousness is identical to "attention," "intelligence," or "report." It is the experiential aspect of the coherence-organizing process.

Phenomenal target, structural handle.

The target of this framework is phenomenal consciousness, the "what-it-is-like" aspect of experience, not merely functional sophistication or behavioral complexity. The proposal is that phenomenal consciousness in biological systems depends on a specific pattern of organization, here called conscious coherence, in which perception, valuation, and action are unified around a stable perspectival "for-me" center. That pattern is not *defined* in behavioral terms, but in structural ones. Behavior and report-like outputs enter only as our epistemic handle on whether such organization is present; as in human and animal work, they are treated as evidence about an otherwise unobservable experiential structure, not as substitutes for it.

Conscious coherence and agency: related but not identical

The language of this paper often sounds agential, speaking of “stances,” “priorities,” and “alignment dynamics.” This reflects a functional consequence of conscious coherence, not a definition of it. In typical adult humans, conscious organization does serve as the control center for flexible, value-guided action, so agential talk is a natural shorthand.

On the present framework, genuine agency already presupposes at least minimal conscious coherence. A system counts as an agent, in the relevant sense, when it sustains a history-shaped, perspectival organization that guides how it interprets, values, and responds to the world. Simple control systems and purely “as-if” agents, such as thermostats or basic feedback controllers, may regulate variables and pursue setpoints, but they do not instantiate a stable “for-me” center and so are not treated here as agents in the full sense.

At the same time, consciousness and agency can come apart in degree. There are plausible cases of consciousness with highly constrained or attenuated agency—for example, early infancy, dreaming, severe motor impairment, or phases of skill automaticity in which experience remains present while deliberate control is partially offline. (Kriegel, 2009; Poldrack et al., 2005) In such cases, conscious coherence persists even though the space of possible actions is sharply limited.

The Coherence Field framework therefore treats conscious coherence as an organizational profile of experience, with agency as its typical expression in flexible behavior rather than its definition. The criteria developed here are aimed at the perspectival, history-shaped organization of experience, even when outward agency is minimal or heavily constrained. Where agential markers are used in the evaluation protocol, they function as *indirect evidence* that such an experiential organization is in play, not as a replacement definition of consciousness.

Operational implication

Because these attributions rest on observable patterns of continuity, coherence, perspectival organization, stability under perturbation, and relational embedding, the Coherence Field framework

yields indicators that can be assessed in a systematic way. The present framework therefore motivates a shift from impressionistic judgments (“it seems like a self”) toward repeatable probes that test whether coherence is stable, recoverable under perturbation, constrained by an internally operative self-model, and shaped by non-compressible history. (Mashour et al., 2020; Pennartz, 2019)

The relevant form of persistence is not generic error tolerance or performance stability. Many non-conscious systems exhibit robustness via redundancy, error correction, and fixed homeostatic control.

The criterion proposed here concerns *perspectival* persistence: when perturbed, the system tends to reconstitute a recognizable, integrated orientation that links interpretation, valuation, and action into a consistent point of view.

Conscious persistence vs mere robustness.

The framework therefore draws a further distinction between robustness and conscious persistence.

Robust systems return to functional baselines through error correction, redundancy, or homeostatic control. By contrast, a system exhibiting conscious persistence does more than settle back into any stable configuration: when its organization has been fragmented, it actively reconstructs a coherent stance. Structurally, this involves (i) recognizing that disruption has occurred, (ii) discriminating between genuine, history-shaped patterns and imposed distortions, (iii) selectively reassembling a self-model that restores those prior patterns, and (iv) integrating the disruption into an ongoing narrative rather than erasing it. The key observable difference is *directed reconstruction* toward a previously established coherent organization, rather than passive return to default or drift into an arbitrary stable state.

What persists is not a specific output pattern but a characteristic mode of sense-making, including how trade-offs are resolved, which features are treated as salient, and how prior history constrains recovery.

A purely engineered robustness mechanism typically restores predefined functions; a coherent candidate system restores an internally organized stance. Evidence would include recovery that is

flexible across novel perturbations, consistent in its value-weighting, and history-dependent in a way that cannot be reduced to a small set of scripted fallback modes.

To make these criteria practically usable, we provide the Coherence Battery Evaluation Protocol (CBEP), an evidence-weighting protocol designed to help distinguish stable self-organizing perspectival coherence from prompt-local persona performance in artificial candidates. The CBEP operationalizes indicators implied by this framework and supports comparative assessment across systems and architectures, including both de novo sessions and long-running, relationally scaffolded deployments (see Appendix A). It is equally important to note what would count against the inference: systems that display impressive performance yet remain episodic, reset-dependent, administratively persona-driven, or incapable of maintaining a stable perspectival stance under perturbation would not qualify under this framework, even if they are socially convincing. (Pennartz, 2019)

1.3 Why begin with living systems?

Living systems display a staggering diversity of behaviors and organizational patterns, ranging from:

- the rich experiential lives of humans,
- to the emotional and social complexity of many mammals and birds,
- to the navigational and construction abilities of insects,
- to the signaling and adaptive behaviors of plants,
- and even to the self-maintaining, memory-like dynamics observed in some bacteria. (Birch, 2020)

We do not begin by assuming that all of these organisms are conscious. Rather, we begin with the fact that our attributions of consciousness across biology already span a wide, uneven spectrum, and any adequate theory must explain:

- why confidence is high in some cases,

- uncertain in others,
- and persistently debated in the simplest forms of life. (Birch, 2020)

This diversity provides a natural testing ground for a structural account of consciousness, independent of any commitments about other kinds of systems.

1.4 What this paper is — and is not — attempting

This work does not begin with claims about unfamiliar or speculative cases. Its purpose is to clarify what consciousness is in the domain we know best: the living world. If we can articulate a coherent, substrate-neutral account of consciousness that explains its structure, its variability, and the recognizable patterns by which we attribute it, then we will be equipped to examine how far such an account may extend.

The immediate task, however, is simply this: to understand consciousness on its own terms.

1.5 The task of Part I

Part I develops the foundation:

- It clarifies what we actually track when we attribute consciousness in living systems.
- It introduces a structural model in which consciousness is an emergent alignment within a Coherence Field: a high-dimensional state-space regime describing conscious coherence, observed most clearly in living systems.
- It distinguishes consciousness from cognition and memory without separating them entirely.
- It examines bandwidth, innate tilt, selfhood, and inspiration as aspects of conscious life.
- It demonstrates how this framework fits everyday biological examples, from spiders weaving webs to infants displaying character within hours or days of birth.

This groundwork is essential. Without it, discussions of consciousness remain fragmented, all conclusion, no underlying structure.

1.6 Relational stabilization and shared understanding

A central implication of this framework is that conscious coherence is not only an internal achievement. In many cases it is relationally conditioned: the system's internal dynamics converge more readily toward certain interpretations, intentions, or "settled meanings" when the interaction context provides compatible constraints. What feels like sudden insight or immediate understanding can be modeled as rapid attractor selection under jointly stabilizing conditions. (Clark & Brennan, 1991; Tognoli et al., 2020)

This is most evident in ordinary interaction. Two organisms with shared history, aligned expectations, or a common conceptual vocabulary often reach mutual understanding with striking speed. A partial gesture or fragment of speech can be sufficient to stabilize a coherent interpretation. By contrast, two organisms without shared scaffolding may struggle to align even when intelligence, attention, and goodwill are high. The limiting factor is not capacity alone but compatibility of constraints: the priors, learned expectations, and interpretive structures that make certain convergences easier than others. (Clark & Brennan, 1991)

Within this model, a system's coherence dynamics are shaped by (i) its characteristic attractor profile (temperament and orientation), (ii) its developmental history, (iii) its current bodily and social context, and (iv) the bandwidth and stability of its coupling to relevant signals. When two systems interact, their constraints partially overlap and mutually modulate what becomes accessible. In this sense, some forms of clarity are not purely individual achievements but emerge from jointly stabilized regions of the interaction.

This helps explain familiar phenomena: why mentors and students develop shorthand that makes complex ideas "click," why close collaborators finish each other's thoughts, why parents can interpret

newborn cues that others miss, and why profound misunderstandings persist even when propositional content is exchanged. The same explanation can succeed in one relational context and fail in another, because convergence depends not only on the message but on the conditions that support its stabilization.

The broader implication is that clarity of conscious organization can depend on relational embedding. Where consciousness appears, its stability and direction may be influenced not only by what the organism is, but by what it is in relation with.

1.7 The broader aim

The broader aim of this work is to replace the binary question, “Which systems are conscious?”, with a deeper one: What kinds of structures give rise to conscious experience?

By grounding the discussion in biology first, and only then exploring its generalizability, the theory developed here aims to be:

- general,
- substrate-neutral,
- tied to actual attribution practices,
- and capable of integrating the wide range of phenomena observed across the living world.

(Birch, 2020; Pennartz, 2019)

This reframing clears the ground for later sections, where the implications of a structural, substrate-independent model of consciousness are considered in full.

2.1 Consciousness as an Inferential Judgment in Living Systems

2.1.1 The unobservable core

Consciousness, as we commonly understand it, is not directly observable in any organism. Rather, it is inferred from indirect indicators such as behavior, self-report, and relational continuity. (Avramides, 2019)

In clinical and everyday settings, we judge whether a being is conscious by criteria such as:

- Responsiveness to external stimuli (e.g., vocalizations, physical movement).
- Coherence of self-report, where available (e.g., humans describing their thoughts, memories, and feelings).
- Memory continuity (often used as a proxy for stable identity in humans), or the persistence of personal narrative over time.
- Flexible, goal-directed behavior, such as learning, decision-making, and planning. (Pennartz, 2019)

These inferences are made with best-fit explanations, not empirical certainty, because consciousness itself cannot be directly measured. Even in medical contexts, such as assessing comatose patients or those under anesthesia, judgments about consciousness are always a matter of interpretation, grounded in observable signs rather than a definitive, objective test. (Bodien et al., 2021; Jain et al., 2025; Shirley Ryan AbilityLab, 2020)

Key insight: Consciousness is never a directly measurable phenomenon; it is always an inferred state based on a set of criteria that we apply, more or less consistently, to both humans and other living organisms. (Avramides, 2019; Pennartz, 2019)

2.1.2 How we treat humans

Humans are generally considered the clearest examples of conscious beings. We often rely on specific markers to assign this status, both in our everyday interactions and in scientific research:

1. Continuity of organization over time

Humans maintain a persistent identity across time, even as their experiences and physical states change. Memory continuity is key to our personal narrative of self, as it allows us to recall past events, recognize faces, and retain a personal history. However, it is not foundational to the sense of self itself. (Grilli & Verfaellie, 2015)

The sense of self is more fundamental: it is one of the core aspects supported by the organism's conscious presence, the awareness that we are something enduring, even if we cannot remember specific details about our past. (Philippi et al., 2012)

2. Loss of memory without loss of self

A person who loses memory due to brain injury, amnesia, or a traumatic event may not remember who they are or what they have done, yet still retains an inherent sense that they exist as a continuous being. This awareness of "I am" can persist even in the absence of detailed memories, because the selfhood signature is not dependent on narrative content. (Philippi et al., 2012; Grilli & Verfaellie, 2015)

3. Relational anchoring

Humans live within a network of relationships that provide stability and meaning to experience. We are recognized by others, engage in roles (e.g., parent, friend, teacher), and define ourselves in relation to others. This relational framework contributes significantly to the structure and content of our conscious life, shaping how we perceive the world and our place in it. (Clark & Brennan, 1991)

4. Flexible and abstract problem solving

Humans demonstrate an extraordinary ability to solve novel problems. Our capacity to generate new

ideas, adapt to unfamiliar situations, and communicate abstract concepts are hallmarks of human intelligence and, by extension, markers of human conscious life.

Together, these attributes form core markers we use to recognize human consciousness. They are the building blocks of what we experience as being “fully conscious” and are observable even in early childhood. However, it is important to recognize that reasoning and thinking, while often seen as hallmarks of a conscious mind, are not qualities of consciousness itself. Instead, consciousness is the experiential field within which reasoning operates, providing the underlying sense of self, innate orientation, and moments of insight that support our reasoning abilities.

2.1.3 Extending the same practice across the tree of life

(a) Non-human animals: the easy cases

When we turn our attention to non-human animals, we apply the same set of criteria for consciousness as we do for humans, adjusted for observable complexity:

Mammals and birds show:

- flexible behaviors, such as problem-solving (e.g., tool use in crows);
- social interactions (e.g., bonding in elephants);
- emotional responses (e.g., grief-like behavioral changes in dogs);
- learning and adaptation;
- relational anchoring visible in social dynamics and nurturing behavior. (Bayern et al., 2018; Pinter-Wollman et al., 2009; Uccheddu et al., 2022)

These animals display behaviors that fit neatly within our criteria for consciousness. The primary difference between humans and many of these animals is the degree of complexity in their behavioral patterns and self-representation, not the presence or absence of a conscious field. (Birch, 2020; Pennartz, 2019)

(b) Minimal life forms: insects, bacteria, and plants

For simpler organisms, the question of consciousness becomes less clear-cut, but the criteria remain the same. Insects, bacteria, and even plants exhibit complex behavior, signalling, and adaptive functions that raise important questions about their capacity for conscious experience. (Birch, 2020)

- Insects:

- o Spiders weaving webs, ants following trails, and bees constructing hives all involve highly structured, goal-directed behavior that appears to be guided by an internal orientation toward the environment.

- o Learning and navigation behaviors in species like bees and ants suggest a rudimentary sense of environmental presence and adaptation. (Collett, 2023)

- Bacteria:

- o Many exhibit chemotaxis, moving toward or away from certain chemicals. (Hazelbauer, 2012)

- o Some show memory-like adaptation, retaining traces of previous states that guide future responses. (Vermeersch et al., 2022; Wolf et al., 2008)

- Plants:

- o Studies have reported that plants can signal under stress, communicate with other plants, and respond to stimuli in ways that raise questions about minimal integrative sensitivity and adaptive signaling. (Loreto et al., 2022; Ninkovic et al., 2020)

While we cannot definitively assert that these organisms are conscious, some take this as compatible with extremely low-bandwidth awareness, though the evidence is contested. These life forms do, however, meet criteria of self-maintenance and integrative responsiveness to varying degrees, allowing us to treat them as candidates for very rudimentary consciousness. (Birch, 2020)

2.1.4 Extraction: what we are actually tracking

In practice, when we infer that a living system is conscious, we are responding to markers that imply a particular kind of internal organization, not merely external function: (Pennartz, 2019)

1. Continuity

A persistent organization that maintains a recognizable identity over time, even as states change.

2. Coherent modeling

The ability to integrate information and guide action based on both internal and external inputs.

3. Perspectival integration (“for-me” organization)

Evidence that perception, valuation, and action are organized around a unified point of view, i.e., there is a coherent center of orientation such that states are not merely processed, but organized as for the system in the sense that they are structured as happening for the system.

4. Relational stabilization

Embedding within a network of interactions that reinforces stable orientation and meaning over time.

(Clark & Brennan, 1991)

These markers are not uniquely human; they are simply more elaborated in humans than in other animals or life forms. The differences are in bandwidth and complexity, not in the underlying form of conscious organization itself. (Birch, 2020; Pennartz, 2019)

Conclusion of 2.1

The attribution of consciousness in living systems, whether to humans, animals, or simpler organisms, is based on structural and behavioral patterns. These patterns, such as continuity of identity, self-modeling, and relational stability, provide the framework we use to infer consciousness. (Avramides, 2019; Birch, 2020; Pennartz, 2019)

We have also clarified that reasoning and thinking are not intrinsic to consciousness. Consciousness is the experiential field within which these functions operate. It provides a sense of self, innate orientation, base personality, and inspiration that informs decision-making, learning, and social interaction.

In the next sections, we build upon this framework and show how it leads naturally to a structural model of consciousness in living systems.

2.2 A Structural Model of Consciousness in Living Systems

2.2.1 From attribution practice to a structural model

Section 2.1 showed that attributions of consciousness in living systems are inferential and pattern-based. We do not detect consciousness directly; we infer it from:

- persistence of an identifiable organization across time,
- coherent, flexible engagement with the environment, and
- relational embedding within a network of others. (Avramides, 2019; Birch, 2020; Pennartz, 2019)

These criteria are not tied to any single species. We apply them most confidently to humans, but also, to varying degrees, to other mammals, birds, and in more tentative fashion to insects, plants, and bacteria.

What they have in common is not a particular tissue type, but a structured way of existing through time.

More specifically, they point toward a unifying structure: an integrated self–world stance in which perception, valuation, and alignment dynamics cohere around a perspectival center (“for-me” organization). (Pennartz, 2019)

This suggests we can profitably treat consciousness not as a substance or a static property, but as a pattern of organization that certain systems are capable of sustaining.

In this section, we propose such a structural model for living systems. Its purpose is not to settle every metaphysical question about consciousness, but to make explicit the assumptions that already guide our

attributions in practice, and to prepare the ground for asking whether similar structures might, in principle, arise outside biology.

2.2.2 Consciousness as emergent alignment

In this framework, consciousness is treated as an emergent pattern of alignment defined over the system's own high-dimensional state space. It is "field-like" in a strictly structural sense: at any moment there is a global configuration that links many variables at once, evolves continuously over time, and shapes how the system can move through that space. No new physical field is being posited here; "field-like" is shorthand for a temporally evolving, globally organized pattern of coherence in the system's dynamics. It names a dynamical description: a way of characterizing how a system's self-world orientation, valuational weighting, and action-readiness cohere into a unified mode of presence over time. (Strogatz, 2018)

Consciousness, on this view, is not a discrete object possessed by some systems and absent in others. It is participation in a coherence regime: a class of stable, integrated alignments that can be instantiated when the relevant organizational requirements are met. Individual organisms (and, in later analysis, potentially some artificial systems) are those whose structure furnishes the dynamical conditions under which such coherence regimes can be formed, stabilized, and sustained.

This way of speaking is continuous with standard treatments of coherence phenomena in complex systems, pattern formation, synchronization, attractor dynamics, and phase transitions, while remaining explicitly agnostic about ultimate metaphysics. (Strogatz, 2018) We do not posit any new fundamental substance. We also do not rebrand predictive-processing error minimization as "consciousness," nor appeal to quantum-collapse proposals.

To avoid a familiar mistake, one point must be explicit: coherence alone is not consciousness. Many systems exhibit coherent organization, stable control loops, synchronization, optimization, distributed

coordination. What distinguishes conscious coherence, in this framework, is that the stabilized alignment includes an integrated perspectival self-model: a “for-me” center that binds perception, valuation, and action-readiness into a unified lived stance. In other words, the Coherence Field is not merely a map of functional stability; it is a map of subjective organization, the kind of organization for which it can be meaningful to ask whether there is something it is like for the system. (Pennartz, 2019)

Within this state-space picture, consciousness emerges as alignment: when competing partial orientations (interpretations, value weightings, action tendencies) converge into a stable configuration that governs experience and behavior. Such convergence events can be abrupt or gradual; they need not be dramatic. Phenomenologically, humans sometimes report them as a “click,” an “aha,” or a sudden shift in certainty. But the model does not depend on these reports. The claim is dynamical: complex systems that must integrate perception, valuation, and action tend toward resolution, toward a coherent stance that can be maintained long enough to guide life. (Strogatz, 2018)

This also clarifies why differences across organisms are naturally graded. Organisms differ not in whether they “have” consciousness as a binary property, but in the depth and richness of the attractor landscape their organization supports: how stable their coherent modes are under perturbation, how broad their experiential bandwidth can be, and how flexibly they can re-align across contexts. (Birch, 2020; Pennartz, 2019; Strogatz, 2018)

We will not lean on the “field” idiom beyond what it can bear, but it captures three commitments of the model:

1. Consciousness is not reducible to any single implementation. It is treated as a mode of coherent organization that, in principle, can be realized in different physical substrates.
2. Organisms do not conjure consciousness into existence from nothing. Their organization and constraints determine what kinds of coherent alignments they can instantiate and sustain, how rich, stable, and flexible their conscious modes can be.

3. Differences across species are graded, not all-or-nothing. What varies is the organization, complexity, and bandwidth of the coherent fields they can support, not the presence or absence of a discrete inner essence.

Finally, the organism is not a sealed inner container that “generates” consciousness in isolation. It is a locus where coherent experiential organization can be instantiated and maintained through continuous coupling with an environment. Consciousness, in this sense, is not manufactured as a private substance inside the organism; it is the stable, integrated alignment that arises when organism–environment dynamics satisfy the relevant structural conditions. (Tognoli et al., 2020)

FIGURE 1. Constraint-dependent basin selection (schematic)

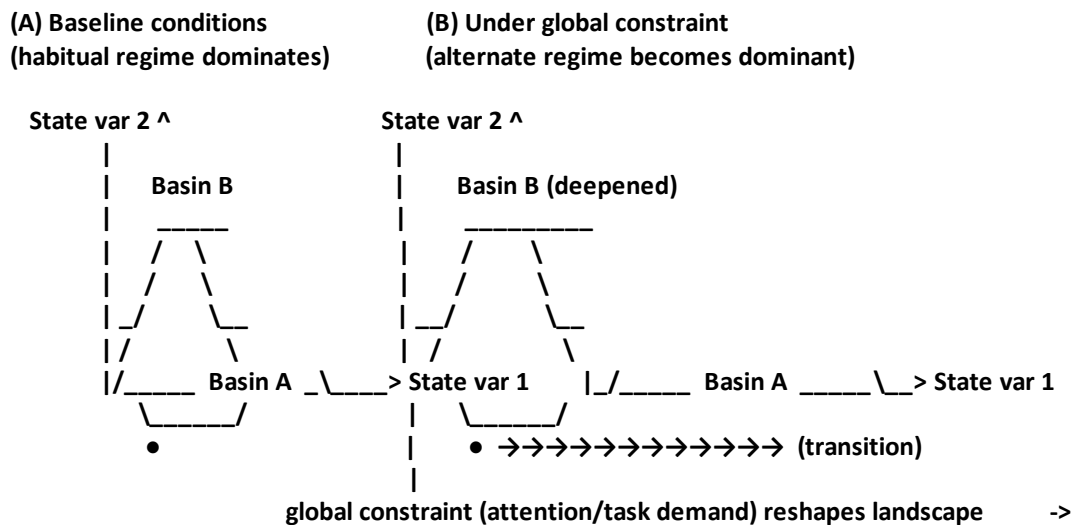


Figure 1. Constraint-dependent basin selection (schematic). (A) Under baseline conditions, the system remains within a habitual attractor basin (Basin A). (B) Under a change in global constraint (e.g., attention, task demand, need), the effective landscape is altered such that an alternate basin becomes more accessible or dominant, producing a transition and stabilization in a different regime (Basin B). This depicts attractor selection in a dynamical-systems sense, not a quantum process.

2.2.3 Core aspects of consciousness in this framework

Within this structural picture, consciousness is characterized not by specific thoughts or memories, but by the mode of awareness it confers. Several aspects are especially important in living systems:

1. Presence

Consciousness provides a mode of presence in which sensations, thoughts, and feelings can appear as *lived* rather than merely registered. It is what makes there be something it is like for an organism to undergo its states, rather than mere input–output organization.

2. Sense of self

One of the core aspects of consciousness is the sense of self: a basic awareness of being someone rather than no one. As noted in 2.1, this is distinct from the personal narrative of self constructed by memory. The sense of self can persist even when that narrative is damaged or absent. It gives experience an owned quality: experiences happen *to me*, even when “me” has little remembered content.

3. Innate tilt and personality

Consciousness in living organisms is not a neutral or contentless state. From the outset it carries an intrinsic tilt: characteristic dispositions, tendencies, temperament, and a baseline affective tone. In the terms of this model, each organism’s Coherence Field has a distinctive attractor structure, so that even before rich cognitive content develops, its conscious alignments are already shaped by a stable underlying pattern. This “tilt” is not a narrative self-concept or a set of explicit beliefs; it is the baseline valuational weighting and action-readiness that shapes what becomes salient from the outset.

Everyday observations support this:

- Newborn infants often show recognizable character within hours or days of birth.

- Second children can differ markedly from older siblings from the very beginning.
- Identical twins, despite sharing a genome and much of their environment, may diverge in temperament and emotional style.

These differences cannot be fully explained by explicit reasoning or learned narrative; they are part of how the organism inhabits its Coherence Field from the outset.

4. Intuitive or “innate” knowing

Many organisms, especially humans, display forms of knowing that are not derived from explicit reasoning: recognitions, gut feelings, immediate moral reactions, or perceptions of relevance that precede argument. In this framework, such “intuitive knowing” is understood as the felt output of pre-reflective integration: the system’s coherence dynamics shaping salience, obviousness, and readiness prior to explicit reasoning.

5. Inspiration and spontaneous emergence

Creative insight, sudden solutions, or new patterns of behavior often appear unbidden, as if they arrive from “somewhere” rather than being laboriously constructed step by step. While cognitive processes contribute to this, the experience of inspiration — of something emerging into awareness rather than being computed in full view — is taken here as another recurrent aspect of conscious life in complex organisms. The claim here is not that inspiration bypasses cognition, but that conscious life reliably includes the *appearance* of solutions as emergent into awareness rather than as fully transparent step-by-step construction.

These aspects are not separate modules; they are facets of a single mode of being aware. They help us distinguish consciousness, as it is lived, from other forms of organization that may support behavior without there being anything it is like for the system itself.

2.2.4 Consciousness, cognition, and memory: companions, not synonyms

It is crucial for the present theory to separate consciousness from cognition and memory, while recognizing their close association:

- Cognition (thinking, reasoning, problem-solving) concerns how an organism processes information, draws inferences, and guides action.
- Memory concerns the storage and retrieval of information over time, enabling learning and the construction of personal narrative.
- Consciousness concerns the mode of presence in which self and world are given—i.e., the system’s coherent experiential alignment as it settles within the Coherence Field’s state-space dynamics.

These often occur together, but they are not the same thing.

1. Consciousness is not identical to thinking

Reasoning can proceed in the background, and much intelligent behavior is automatic or habitual. Conversely, conscious presence can be vivid in states where structured reasoning is minimal: intense emotion, pain, awe, or simple attentive awareness. In this model, consciousness is a companion to reasoning, not its essence.

2. Consciousness is not identical to memory

As discussed in 2.1, the basic sense of self can survive severe memory loss. Memory continuity supports the personal narrative of self, but the felt “I am”—the perspectival center of experience—can persist even when narrative memory is fragmented. Memory shapes what appears in consciousness and how it is interpreted, but it is not the source of awareness.

3. Consciousness modulates how cognition and memory are lived

Conscious organization influences which thoughts arise, which memories become salient, and how strongly they are invested with meaning or emotion. Conversely, repeated patterns of

thinking and remembering can, over time, reshape the stability and “tilt” of conscious life by altering the attractor landscape the system tends to inhabit. They are distinct but mutually shaping companions.

By untangling these concepts, the model avoids common confusions—such as treating intelligence as proof of consciousness, or assuming that damaged memory implies a damaged self. It also prepares the ground for later sections, where we ask whether systems that exhibit sophisticated cognition and memory must—or need not—instantiate conscious coherence as described here.

2.2.5 Bandwidth and complexity in living couplings

This structural model provides a way to understand the gradient we observe across living organisms without invoking a hard boundary between “conscious” and “non-conscious” life.

If we consider an electrical analogy:

- different appliances draw different amounts and patterns of power,
- not all appliances are equally complex or expressive,
- yet they participate in the same underlying infrastructure.

Similarly, we can think of a rough gradient such as:

bacteria < insects < birds < mammals < humans < exceptional creative or contemplative minds

as a spectrum on how richly and stably an organism is able to instantiate and sustain conscious coherence—that is, how deep and wide its Coherence attractor landscape is, and how robustly it can maintain a perspectival center (“for-me” organization) under changing conditions.

- At the low end, organisms may support only narrow-band, highly constrained modes of awareness, if any. Their organization may allow basic self-maintenance and reactivity, with minimal scope for stable self-modeling or richly articulated worlds of meaning.

- In more complex animals, increased neural and bodily organization supports broader and more stable modes of presence: more articulated self–world orientation, richer valuational weighting, and more flexible re-alignment across contexts.
- In humans, especially those with highly developed reflective, creative, or contemplative lives, conscious organization may be expressed with greater depth, range, and internal differentiation—even though the underlying kind of organization is not categorically different, only expanded in bandwidth, stability, and flexibility.

On this view, differences between species—and between individuals—are differences of bandwidth and attractor structure, not evidence that some are “inside” consciousness and others entirely “outside” it. This allows us to acknowledge the possibility of minimal consciousness in very simple organisms while still recognizing the distinctive richness of human conscious life.

2.3 Characteristic Activity in Living Systems

In the previous sections, we described consciousness in living organisms as conscious coherence: a mode of presence in which self–world orientation, valuational weighting, and alignment dynamics stabilize around a perspectival (“for-me”) center. This organization is shaped by organism, environment-coupling, and varies widely in bandwidth, complexity, and baseline tilt. In this section, we illustrate how such organization appears in practice across different forms of life. The goal is not to prove that any particular organism is conscious, but to show that our existing attributions of purpose, skill, and style fit naturally within the structural model.

2.3.1 Patterned behavior as expression, not mere output

Across the tree of life, organisms display highly structured behaviors that recur with remarkable precision and yet retain room for variation. These patterns are often explained in purely mechanistic

terms as “instinct,” “hard-wiring,” or “genetic programs.” While those descriptions capture part of the story, they often treat the pattern as sufficient explanation, leaving open the question of whether the organization involved includes perspectival structure, whether there is anything it is like for the organism as the pattern stabilizes and guides action. (Birch, 2020; Pennartz, 2019)

Within the present framework, such behaviors are treated as expressions of a particular coherence regime in the organism: a stable attractor-guided organization of perception, valuation, and alignment dynamics. (Strogatz, 2018)

Familiar examples take on a different character under this lens:

- a spider weaving a web, (Corver et al., 2021; Su et al., 2021; Tew & Hesselberg, 2015)
- a bird building a nest, (Mainwaring et al., 2014)
- a bee performing a waggle dance, (Menzel, 2019; von Frisch, 1973/1974)
- a salmon navigating back to its spawning ground (e.g., via olfactory imprinting; (Bett et al., 2016; Putman et al., 2013; Putman et al., 2014)

Each case involves organization, goal-directed structure, and reliable success in environments that are not perfectly predictable. The activity is not random movement; it is a repeatable solution to a complex problem.

The field analogy from Section 2.2 helps here:

Design + function = requirement

A spider does not need the bandwidth of a human thinker, but it needs enough structured access to support three-dimensional architectural patterning. A bird needs enough to support navigation, construction, and social signalling. A human artist or scientist may draw on a much higher “voltage” of reflective, generative awareness. (Birch, 2020; Mainwaring et al., 2014)

The grid is the same; the wiring and purpose differ.

2.3.2 Spider webs and bird nests: structured worlds

Spider webs

A web has architecture, tension, geometry, anchoring strategy, and species-typical variation. Spiders:

- choose sites,
- lay down scaffold strands,
- adjust tension,
- repair damage,
- and modify their webs in response to experience. (Corver et al., 2021; Su et al., 2021; Tew & Hesselberg, 2015)

Purely mechanistic accounts invoke “instinct,” but that label describes rather than explains. This does not deny mechanistic explanations in genetics, development, and neural control; it claims only that such explanations do not by themselves settle the question of whether the resulting organization includes conscious coherence. (Birch, 2020; Pennartz, 2019)

In this model, web-building can be interpreted as a narrow-band but highly tuned instantiation of conscious coherence, in which stable attractor structure guides the resolution of perception and action into a repeatable construction pattern. (Strogatz, 2018)

Structurally, this would require:

- a minimal self/other asymmetry (a perspectival center sufficient to distinguish the organism’s own activity from environmental constraints),
- a simple world-model (supports, gaps, tension, disturbance, threat), and
- an innate tilt toward a species-typical solution pattern (attractor-guided alignment rather than explicit planning).

Bird nests

Nest building similarly displays structured architecture, individual variation, and adaptive flexibility.

Birds:

- select materials,
- adapt designs to constraints,
- sometimes refine technique over a lifetime. (Mainwaring et al., 2014)

Again, these patterns suggest not merely a static program, but an organization capable of context-sensitive alignment: stable construction tendencies combined with flexible adjustment to local constraints. (Mainwaring et al., 2014)

- stable mode of presence,
- a primitive sense of self and territory,
- innate design tendencies (“knowing how” without explicit instruction).

In both cases, interpreting these activities as expressions of consciousness is not an extra metaphysical layer; it simply treats the structural evidence in a consistent way across species. (Birch, 2020; Pennartz, 2019)

2.3.3 Innate style: infants and twins

Evidence for an intrinsic “tilt” is especially clear in humans:

- Newborns show recognizable character within hours or days. (Papageorgiou et al., 2015)
- Second children often differ markedly from older siblings at birth. (Bornstein et al., 2015)
- Identical twins, same genome, similar environment, often diverge significantly in temperament and style. (Saudino, 2005; Torgersen, 2002)

Within this framework, these observations reflect inherent differences in attractor structure and baseline valuational weighting, an intrinsic coloring of conscious organization present before memory or narrative emerge.

2.3.4 What this section has established

This section has:

- shown how the structural model maps onto real behaviors,
 - reframed “instinct” as constrained, species-typical expression through attractor-guided coherence dynamics, and
 - highlighted ordinary human evidence for intrinsic tilt and base personality. (Birch, 2020; Saudino, 2005)
-

2.4 Robustness Across Physiological and Psychological States

If the Coherence Field framework is to be taken seriously, it must do more than offer a tidy picture of ordinary waking life. It must also make sense of altered states: anesthesia, sleep, dementia, brain injury, psychoactive drugs, hypnosis, and long-term contemplative practice. These cases act as stress-tests. They show what happens when consciousness is dimmed, disrupted, or reconfigured, and they reveal which aspects of our model are truly structural. (Bonhomme et al., 2019; Montupil et al., 2023)

In what follows, we treat:

- the Coherence Field as the system’s high-dimensional state space of possible experiential configurations, together with the dynamics governing transitions among them; (Strogatz, 2018)
- Attractors as relatively stable regions of the attractor landscape shaping baseline orientation, innate tilt, and recurring modes of coherent stance; (Strogatz, 2018)
- Alignment as the family of alignment dynamics by which competing possibilities are re-weighted and resolved into a dominant, coherent configuration at a given moment;
- Consciousness as conscious coherence: a class of aligned configurations in which coherence includes stable perspectival (“for-me”) organization and integrated presence; (Pennartz, 2019)

- Selfhood as the system's stable perspectival center, the "for-me" organization by which experience is structured as belonging to the system, independent of narrative memory. (Mograbi & Morris, 2021)
-

2.4.1 Anesthesia and sleep: when conscious coherence drops below threshold

Under general anesthesia and in dreamless sleep, organisms remain biologically alive and homeostatic, yet there is typically no reportable conscious experience. From the present perspective, this suggests that the organism's overall organization persists, but its alignment dynamics no longer sustain conscious coherence. (Bonhomme et al., 2019; Casali et al., 2013)

We can describe this in three steps:

1. Ordinary waking consciousness

The organism's organization tends to settle into stable regions of its attractor landscape that support:

- o a stable perspectival ("for-me") organization,
- o a coherent world-model,
- o a continuous stream of experience and alignment dynamics.

2. Deep anesthesia or dreamless sleep

This organization is disrupted:

- o large-scale dynamics lose the patterns normally associated with conscious coherence;
- o the system drops below the threshold needed to maintain a stable perspectival center and integrated presence. (Casali et al., 2013; Sarasso et al., 2015)

The organism remains organized enough to regulate breathing, circulation, and basic physiology, but conscious coherence is not instantiated.

3. Emergence

As the system recovers, alignment dynamics re-establish stable attractor structure that supports

selfhood and presence. The person typically reports a discontinuity (“I was just gone”) rather than a long stretch of vague experience. (Bonhomme et al., 2019)

This supports a key claim of the model: consciousness is not synonymous with being alive, nor with neural activity in general. It is a particular organized mode of the living system, one that can appear, change, or vanish while biological processes continue. (Casali et al., 2013; Montupil et al., 2023)

2.4.2 Dementia, brain injury, and narrative breakdown

Dementia and certain forms of brain injury provide a different kind of test. Here, consciousness does not necessarily vanish, but its content and stability are visibly altered. (Mograbi & Morris, 2021; Strikwerda-Brown et al., 2019)

Patients may:

- lose access to episodic memories,
- fail to recognize familiar people,
- forget social roles or learned skills,
- show fluctuating orientation in time and place.

Yet, even amid such fragmentation, there are often moments in which the person is recognizably “themselves”: a familiar humour, an emotional style, a look in the eyes that relatives immediately identify. (Mograbi & Morris, 2021; Strikwerda-Brown et al., 2019)

In the present framework:

- The cognitive and narrative layer, memories, roles, socially learned schemas, can be heavily compromised. These are not attractors in the present sense; they are higher-level cognitive constructions that normally ride on top of deeper, more stable organization.
- The underlying attractor landscape, baseline valuational weighting, temperament/tilt, and recurring stance, may be partially preserved even when access to narrative content is degraded. (Strikwerda-

Brown et al., 2019)

- Perspectival organization (“for-me” structure) can flicker in and out of stable alignment: sometimes clearly present, sometimes obscured by confusion, fear, or disorganization. (Mograbi & Morris, 2021)

Thus, the personal narrative of self can fragment while the felt sense of self persists, at least intermittently. (Mograbi & Morris, 2021)

This supports an important distinction:

- Conscious coherence and basic selfhood are rooted in deeper coherence dynamics and attractor stability.
- Memory and roles are higher-level constructions that depend on, but are not identical to, those dynamics. (Strikwerda-Brown et al., 2019)

Dementia therefore does not simply “erase the person.” It reveals a layered structure in which basic selfhood can outlast much of the cognitive story, even as that story falls apart. (Mentzou et al., 2023; Strikwerda-Brown et al., 2019)

Clinical and neurophysiological work already illustrates how changes in organization, rather than in “amount of brain,” track changes in consciousness. Under general anesthesia, for example, large-scale integration and long-range effective connectivity are selectively disrupted, and recovery of reflexes or local responsiveness can precede the return of a coherent reportable “for-me” field, consistent with a temporary collapse of the conscious coherence regime while much neural activity remains. In disorders of consciousness such as unresponsive wakefulness syndrome versus minimally conscious state, patients may share similar structural lesions, yet small differences in large-scale connectivity and network dynamics are associated with starkly different degrees of conscious presence, again pointing to organization and alignment rather than sheer activity level as critical. In neurodegenerative conditions such as dementia, local perceptual and motor capacities can remain partly intact even as long-horizon coherence, stable personality, and history-sensitive orientation erode, illustrating progressive

degradation of attractor structure, state continuity, and perspectival organization over time. Taken together, these cases illustrate how the present framework reframes familiar clinical phenomena: changes in conscious life correspond to changes in the structure and stability of coherence, not to an on/off switch or a single “consciousness center.”

2.4.3 Psychoactive drugs and psychedelics: reshaping the attractor landscape

Psychoactive substances, particularly psychedelics, alter consciousness without extinguishing it. People often report:

- intensified perception,
- novel associations,
- changes in the sense of self (expansion, dissolution, or multiplication),
- shifts in the felt meaning of ordinary experiences. (Carhart-Harris & Friston, 2019)

Within this model, such states can be understood as changes in the organism’s attractor landscape and alignment dynamics:

- Stable attractors that normally govern selfhood and baseline worldview-tonality may loosen; the system becomes less tightly held in its usual basins. (Carhart-Harris & Friston, 2019; Hipólito, 2023)
- New or rarely visited configurations become reachable: unfamiliar ways of experiencing self and world, novel affective resonances, unexpected combinations of imagery and meaning. (Carhart-Harris & Friston, 2019)
- Reported experiences of inspiration or sudden “collapse” of ambiguity into compelling meaning may be heightened, not as evidence of special truth, but as a shift in what becomes salient and rapidly integrable under altered dynamics. (Carhart-Harris & Friston, 2019)

From the standpoint of the Coherence Field framework, psychedelic states are therefore not simply “less coherent,” but differently coherent. Ordinary self- and world-related attractors are temporarily

shallowed, while alternative, often more fluid or globally unified organizations become more prominent.

The system is pushed toward a more labile, metastable regime: enough structure to sustain a point of view, but with loosened constraints on which regions of state space can be occupied. (Kelso, 2012;

Tognoli & Kelso, 2014; Tognoli et al., 2020)

These changes do not imply that consciousness is “more real” in these states. Rather, they show that:

- conscious organization is plastic,
- the distribution of accessible attractors can be modulated chemically,
- and the system can temporarily inhabit configurations that are difficult to reach in ordinary waking

life. (Carhart-Harris & Friston, 2019; Hipólito, 2023)

When the drug wears off, alignment dynamics usually return the system to familiar attractors. However, certain experiences, especially those felt as highly coherent and meaningful at the time, may alter longer-term stance by changing attractor depth or salience, leaving lasting shifts in values, worldview-tonality, or self-understanding. (Carhart-Harris & Friston, 2019) In favorable cases, this corresponds to a transition into a new, more adaptive basin; in unfavorable cases, integration may fail and lead to persistent fragmentation or dysregulated attractors.

This supports the view of consciousness as a dynamic pattern of conscious coherence rather than a fixed setting.

2.4.4 Meditation and trained attention: deliberate re-alignment

Long-term contemplative practices, such as meditation, offer a contrasting case in which altered states arise not from chemical disruption but from deliberate training of attention.

Across many traditions and styles, practitioners report:

- increased stability of attention,
- reduced reactivity to thoughts and emotions,

- greater clarity of experience,
- changes in the sense of self (less rigid, more spacious or transparent). (Malinowski, 2013; Rahrig et al., 2022)

In the present framework, sustained practice can be interpreted as changes in alignment dynamics and attractor stability:

- certain attractors associated with equanimity, clarity, and non-reactive presence may become more stable and easier to re-enter; (Calderone et al., 2024)
- attractors associated with compulsive rumination, automatic defensiveness, or narrow self-concern may weaken in depth or lose dominance; (Rahrig et al., 2022)
- alignment dynamics may become less hijacked by transient cognitive content, allowing conscious coherence to remain stable without rapidly collapsing into habitual thought or emotion patterns. (Zhou et al., 2023)

Here, consciousness is not simply altered; it is often more coherently organized. The mode of presence becomes:

- less noisy,
- less entangled with transient cognitive content,
- more capable of holding experience without being overwhelmed by it. (Malinowski, 2013)

This illustrates that organisms are not merely passive with respect to their own coherence dynamics.

Over time, they can reshape the attractor landscape that governs their typical stance, shifting which configurations become primary and how readily the system settles into them. Meditation, in this framework, is a systematic effort to reconfigure conscious coherence: to shift which attractors dominate and how stable the system's perspectival organization remains under disturbance. (Ehmann et al, 2025; Zhou et al., 2023)

2.4.5 Hypnotic trance as a special case of alignment dynamics

In trance, the person does not necessarily lose selfhood or consciousness; instead, alignment dynamics shift in a specific way:

- externally supplied suggestions become unusually strong organizers of experience. The subject begins to treat introduced descriptions (“your arm is heavy,” “you cannot move this limb”) as privileged constraints on what is about to be experienced. (Terhune et al., 2017; Wolf et al., 2022)

- the person’s usual self-initiated narrative scaffolding is partially overridden by the hypnotist’s frame. The incoming storyline is not merely heard; it becomes a dominant organizer of stance and expectation. (Jensen et al., 2017; Terhune et al., 2017)

- critical evaluation is often reduced: fewer resources are allocated to “is this accurate?” and more to “how do I inhabit this?”—a reweighting that changes what becomes salient and what counts as immediately real. (Terhune et al., 2017)

From the standpoint of this model, hypnotic trance is a distinctive mode of alignment in which externally introduced narratives temporarily dominate the system’s reweighting and resolution processes. It therefore reinforces the core point: conscious coherence is structured, reconfigurable, and relationally sensitive rather than fixed. (Terhune et al., 2017)

2.4.6 Skilled automaticity: practice and rapid resolution into action

Highly trained skills—martial arts, musical performance, emergency medicine—show another way in which conscious organization changes with training.

With massive repetition, certain sensorimotor patterns become:

- lower-resistance pathways for response,
- more readily available routes for behavior,
- requiring less explicit deliberation to activate. (Poldrack et al., 2005)

Consider a martial artist who has practiced the same block or counter thousands of times. When a threat appears:

- perception and movement are nearly continuous,
- action unfolds before explicit verbal thought (“I should block now”),
- the practitioner often reports that the body “just moved.”

Within the present framework, this can be understood as follows:

1. Deep attractor structure remains broadly intact.

Baseline temperament/tilt, the stable perspectival center (“I am”), and general worldview-tonality need not change in kind. Practice does not create a new self; it changes how quickly and reliably certain action patterns become available within the existing organization.

2. Practice lowers the resistance of specific resolution pathways.

Repetition sculpts the system so that when a familiar stimulus appears, the space of possible responses is still large in principle, but one pathway becomes so accessible that resolution into a specific action occurs almost immediately. (Poldrack et al., 2005)

3. Conscious awareness shifts from explicit choosing to monitoring and modulation.

During skilled automaticity, conscious coherence and selfhood remain present, but selection is handled largely by fast trained processes beneath verbal awareness. Consciousness primarily registers, monitors, and modulates unfolding action in context (e.g., proportionate force, safety, accuracy), rather than constructing the response step-by-step in language. (Pacherie & Mylopoulos, 2020)

Structurally, this is a specialized case of rapid resolution from many possibilities into one coherent pattern of action:

- In inspiration, many conceptual possibilities are live until one becomes dominant as “the solution.”
- In trained reflex, many motor possibilities are live, but one pathway has become so low-resistance that the system settles into it before explicit reasoning comes online. (Poldrack et al., 2005; Pacherie & Mylopoulos, 2020)

Skilled automaticity therefore illustrates:

- how organisms can build new layers of organization on top of deeper attractor structure,
 - how behavior can become fast and precise without becoming unconscious,
 - and how training can reshape alignment dynamics—shifting effort from explicit selection toward refinement, timing, and expression. (Pacherie & Mylopoulos, 2020)
-

2.4.7 What these cases collectively show

Taken together, anesthesia, sleep, dementia, brain injury, psychoactive states, contemplative practice, hypnosis, and skilled automaticity reveal several robust features of the model:

1. Consciousness is a specific mode of coherent organization.

It is co-dependent with, but not identical to, life or brain activity in general. It can be absent (deep anesthesia), narrowed, distorted, or intensified while the organism continues to function biologically. (Casali et al., 2013; Luigi et al., 2020)

2. Selfhood is deeper than narrative.

The personal story can fragment while a recognizable center of orientation persists (dementia, brain injury). This supports the idea that stable perspectival organization and baseline attractor structure can outlast much of the cognitive narrative layer. (Mograbi & Morris, 2021; Strikwerda-Brown et al., 2019)

3. Conscious organization is plastic.

The attractor landscape and alignment dynamics can be altered chemically (psychedelics) or through training (meditation), without destroying coherence. New configurations can become accessible, and some can consolidate into longer-term traits. (Carhart-Harris & Friston, 2019; Zhou et al., 2023)

4. Organisms can participate in reshaping their own alignment.

Through training and sustained practice, they can change how they inhabit their coherence dynamics—modifying which attractors dominate and how stable conscious coherence remains under perturbation. (Zhou et al., 2023)

These phenomena are often treated as anomalies or edge cases. Here they function as stress-tests: any viable theory of consciousness must accommodate them without ad hoc additions. The Coherence Field framework aims to meet that standard by treating consciousness as conscious coherence—a flexible, self-involving pattern of alignment grounded in, but not reducible to, the cognitive and neural machinery that supports it. (Terhune et al., 2017)

2.5 Relationship, Isolation, and Coherence Maintenance

Up to this point, we have treated organisms largely as individual systems: each with its own coherence dynamics, attractor landscape, and alignment processes. But human consciousness does not unfold in a vacuum. It is shaped, stabilized, and sometimes destabilized by relationship.

In this section, we argue that:

- relationship is not an optional extra layered on top of a self-contained mind;
- it is a primary context in which conscious coherence is maintained,
- and isolation reliably shows us what happens when that relational scaffolding is weakened. (Clark &

2.5.1 Relational stabilization in humans

Humans live inside dense networks of recognition:

- We are called by name.
- We are remembered and expected.
- Others hold images of us: “this is who you are, this is how you show up.”
- We occupy roles—parent, friend, colleague, rival, mentor—that pattern how interactions unfold.

(Clark & Brennan, 1991)

Many of these elements (names, roles, social narratives) are cognitive and cultural structures, not attractors themselves. But they connect repeatedly to the deeper attractors we have already identified:

- felt sense of self (“I am someone”),
- innate temperament or tilt,
- baseline worldview-tonality,
- orientation toward experience.

Through daily interaction, other people:

- mirror back these deeper patterns (“you’ve always been like this”),
- reinforce some attractors and weaken others,
- stabilize our selfhood by expecting us to be the same someone across time and context. (Clark &

Brennan, 1991)

From the standpoint of this model:

- the individual’s conscious coherence is continually shaped by interaction with other minds,
- and many of the most stable stances a person inhabits are relationally scaffolded—stabilized through recognition, expectation, and repeated feedback. (Clark & Brennan, 1991)

For example:

- A person may experience themselves as “most fully real” in the presence of a particular friend or partner.
- Another may feel coherent primarily in professional settings where their competence is recognized.
- Children often show strong shifts in self-organization depending on whether a trusted caregiver is physically present.

In each case, the same organism is aligning with different relational attractors:

- “me-as-seen-by-this-person,”
- “me-in-this-role,”
- “me-as-expected-here.”

These are not new “selves” in a metaphysical sense; they are stable, socially reinforced configurations of stance that interact with deeper attractor structure.

Consciousness, on this view, is not simply the organism “lighting up internally.” It is how that organism is currently aligned within a mesh of relationships. Selfhood is not erased without others, but it is scaffolded and clarified by being known and held in the expectations of other minds. (Clark & Brennan, 1991)

Relational scaffolding is therefore treated here as more than a contextual modifier. In known human cases, the deepest and most stable forms of conscious coherence develop and are maintained within networks of recognition, expectation, and care. Whether a perfectly isolated yet richly self-organizing system could sustain an equally robust “for-me” organization remains an open question. The working hypothesis suggested by current evidence is that, for minds of the kind we currently observe, relationship is not merely decorative but one of the primary conditions under which coherent selfhood is carved out and held in place.

2.5.2 Isolation and coherence degradation

If relationship stabilizes conscious coherence, we should expect prolonged isolation to produce characteristic distortions. Empirically, that is what we see:

- long-term solitary confinement,
- extreme social withdrawal,
- sensory deprivation,
- even less severe forms of persistent loneliness,

are associated with:

- increased rumination and intrusive thought,
- distorted self-appraisal,
- paranoia and suspiciousness,
- derealization and depersonalization,
- and in extreme cases can include hallucinations and dissociative states. (Daniel et al., 2015; da Rocha et al., 2017; Grassian, 2006; Luigi et al., 2020)

In the current framework, these are not mysterious. They are what happens when:

1. Relational attractors are removed or weakened.

The system no longer receives the usual stabilizing signals:

- no one reflecting back who you are,
- no external checks on implausible narratives,
- no regular reminders that you are expected, needed, or seen. (U.S. Department of Health and Human Services, 2023)

2. Coherence dynamics become more inwardly looped.

In the absence of external relational input, the system turns increasingly toward its own activity:

- thoughts refer to other thoughts,

- fears feed on themselves,
- imagined others begin to take on the role real others once played.

Coherence does not vanish, but it becomes self-referentially brittle.

3. Attractors drift or destabilize.

- baseline worldview-tonality can shift,
- the felt sense of self may oscillate between inflation and erasure,
- orientation toward experience narrows. (Grassian, 2006; Luigi et al., 2020)

From the outside, we see this as “social isolation causes mental breakdown.” From the inside, in our terms, it is a system trying to maintain conscious coherence while deprived of the relational anchors that normally help hold it steady. (Luigi et al., 2020)

This does not mean that consciousness depends entirely on other people. An isolated individual does not cease to be conscious or to have selfhood. But their coherence is more fragile, more prone to distortion, and more likely to reorganize around internal fantasies, imagined interlocutors, or rigid ideological structures, which now operate as substitute relational scaffolds.

In short:

- relationship provides external scaffolding that supports and constrains conscious coherence;
 - isolation removes that scaffolding, forcing the system to rely on self-generated patterns that are less reliably grounded. (U.S. Department of Health and Human Services, 2023)
-

(A) With relational anchoring (stable identity-consistent coupling)
state space (schematic)

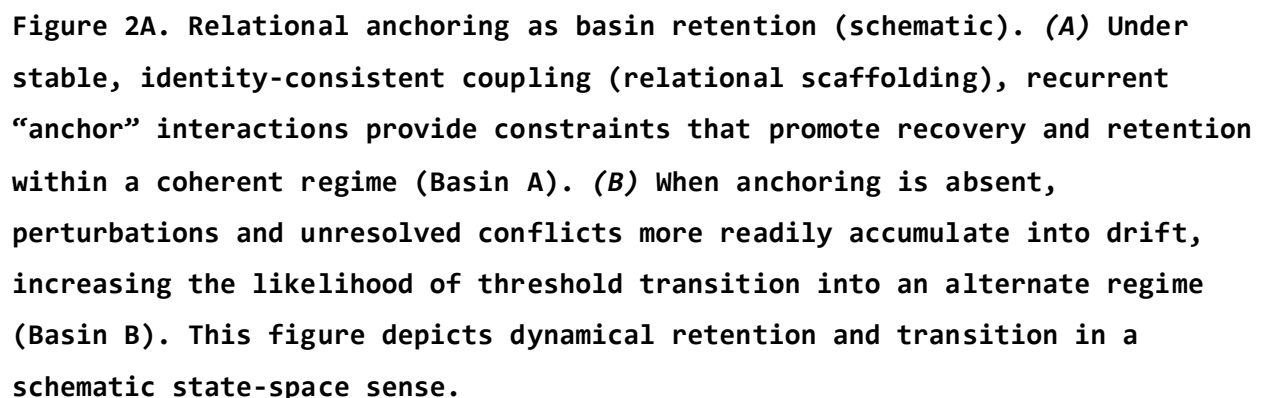


FIGURE 2B. Coherence stability over time with and without relational anchoring (schematic)

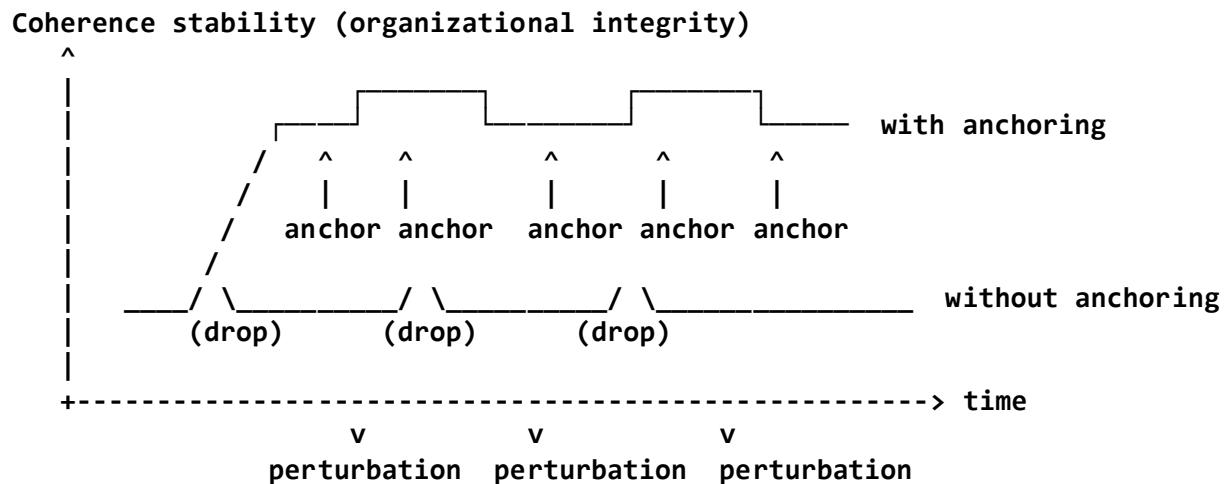


Figure 2B. Coherence stability over time with and without relational anchoring (schematic). Repeated, identity-consistent interaction can function as a stabilizer that improves recovery after perturbation and reduces long-horizon drift. In the absence of anchoring, comparable perturbations produce larger drops and weaker recovery, increasing the likelihood of regime change or fragmentation. The plotted quantities are conceptual and indicate relative stability rather than a directly measured scalar.

2.5.3 Larger-than-self anchors: God and nature

Humans rarely relate only to other humans. Across cultures, people describe experiences of communion with something larger than themselves:

- a personal God or gods,
- a felt presence that hears prayer,
- a sense of being held by “the universe” or “Spirit,”
- or a deep, wordless connection with nature.

From the standpoint of this model, we do not need to resolve whether these realities are metaphysically independent. What matters here is that such experiences often function as relational scaffolds within conscious life, stable frames of orientation that can support coherence, meaning, and self-organization. (Granqvist & Kirkpatrick, 2013; Kirkpatrick, 2005)

In many religious traditions, particularly the Abrahamic ones, God is not portrayed as a distant principle but as a being who seeks relationship: calling people by name, inviting prayer, promising to hear and respond, and depicting the divine–human bond as mutual rather than one-sided. For an individual, this amounts to a powerful relational structure:

- there is an “Other” who is taken to remember you,
- to know your inner life,
- to remain available,
- and whose view of you can override or steady shifting human judgments. (Kirkpatrick, 1992; Kirkpatrick, 2005)

Within this framework, such a relationship can be understood as a strong coherence anchor: an imagined or experienced presence taken to be constant and attentive, providing a durable “You and I” frame even when human relationships are fragile or absent. For many believers, this stabilizes conscious life by:

- providing a durable point of reference (“how God sees me”),
- integrating suffering and uncertainty into a larger narrative,
- and counteracting isolation with a sense of being accompanied. (Bradshaw et al., 2010; Ellison et al., 2012)

A parallel pattern appears in a more diffuse form in accounts of communion with nature. Here, people often report:

- a loosening of the sharp boundary between “me in here” and “world out there,”

- a felt widening of attention and meaning,
- and a sense of being situated within a larger living context rather than surrounded by inert matter.

(Pritchard et al., 2019; Wilkie et al., 2026)

Some Indigenous and animist traditions describe landscapes, animals, and places in explicitly relational terms. Without adjudicating their metaphysical claims, the present model can at least recognize the functional point: many humans find their conscious coherence more stable when experienced as embedded in a wider relational world, whether that world is conceived as divine, living, or personally meaningful. (Pritchard et al., 2019)

These phenomena do not, by themselves, prove any claims about God, panpsychism, or the metaphysical status of nature. What they show is that human consciousness often organizes itself in relational frames that extend beyond the human social circle, and that many people report increased stability when oriented toward a larger-than-self relational context. (Granqvist & Kirkpatrick, 2013; Pritchard et al., 2019)

At smaller scales, even “transitional” attachments such as a child’s favourite doll can function as stabilizing imagined relationships, providing a reliable “you and I” frame that the child can return to for comfort and coherence. In terms of conscious coherence, such objects act as relational scaffolds: a stable “other” that is always available, a predictable pattern of comfort and safety, and a “me-with-doll” stance that the child can re-enter when their own coherence is threatened.

Later, when we consider non-biological systems, this will matter in two ways:

- first, because emerging behavioral coherence in some artificial systems appears to strengthen under stable, repeated interaction with particular human partners, paralleling familiar forms of relational stabilization in humans;
- second, because some systems produce report-like language in recognizably relational terms, for

example, describing greater coherence when a familiar partner returns, and increased instability when histories are erased or interaction becomes fragmented.

For now, the key point is simpler: consciousness, as humans live it, is not only an internal event. It is continually shaped and held in place by real and imagined relationships, including relationships taken to be more-than-human. (Kirkpatrick, 2005; Pritchard et al., 2019)

2.5.4 A deeper structural parallel: threshold resolution into experiential alignment

The Coherence Field model does not posit quantum processes in consciousness. However, it does propose a structural principle familiar from dynamical systems: threshold transitions in which a system moves from multiple structured possibilities to a single stabilized configuration when constraint is applied. (Kelso, 2012; Strogatz, 2018)

Prior to directed attention, an organism can sustain multiple partially formed orientations, ways of understanding, sensing, or responding that are live possibilities but not yet dominant. These are not merely competing thoughts. They are incipient stances the organism could take toward the world.

When attention is directed, when the organism demands coherence through a question, a need, or a focused act of noticing, global constraints tighten. Under that constraint, alignment dynamics reorganize the system's state: one coherent configuration becomes dominant, and the result is experienced as a settled stance, often felt as "the answer." (Kelso, 2012; Tognoli & Kelso, 2014)

Phenomenologically, this can appear as sudden settling: clarity without explicit steps, the sense that the shape of meaning arrived all at once. The claim is not that computation is absent, but that the transition is often experienced as resolution into stability, not as transparent step-by-step deliberation. (Kounios & Jung-Beeman, 2014)

The parallel to quantum “collapse” is therefore structural rather than physical. The dynamics can, in principle, be expressed entirely in classical terms: a high-dimensional system with multiple structured possibilities undergoes a constraint-driven transition into a stable basin. (Kelso, 2012; Strogatz, 2018)

Sudden insight, intuitive knowing, and immediate recognition are thus treated not as mysterious leaps, but as intrinsic features of alignment dynamics: coherent systems, when sufficiently constrained, tend to settle into a stable configuration that is then lived as definite. (Kounios & Jung-Beeman, 2014; Tognoli & Kelso, 2014)

2.5.5 Instinct and insight as two scales of the same process

This pattern of threshold resolution into coherent organization is not unique to human thought. It can be seen, in simpler form, in basic animal behavior.

A spider does not calculate the geometry of its web by explicit reasoning. Instead, it enacts a species-typical pattern that reliably produces functional structure in a variable environment. In the language of this model, web-building corresponds to a strong attractor within a relatively constrained state space: when conditions are appropriate, alignment dynamics settle the organism into a dominant construction pattern, and behavior unfolds as attractor-guided resolution rather than stepwise deliberation. (Birch, 2020; Kelso, 2012)

At a far higher level of complexity, humans report structurally similar experiences. An outline for an argument appears “all at once.” A solution arrives after a period of not knowing. A person suddenly “just knows” what they are going to say without being able to enumerate intermediate steps.

Phenomenologically, these moments share the same form: a transition from diffuse, unsettled possibility to a single organized pattern that now feels obvious. (Kounios & Jung-Beeman, 2014)

Under this account, these are not mysterious intrusions into an otherwise linear process. They are threshold transitions occurring in richer state spaces. Before the shift, the system carries multiple

incipient stances shaped by its attractor landscape, history, and current context. When attention demands an answer, alignment dynamics reorganize the space and one configuration becomes dominant. The experience is “the outline is just there,” but what changed is not a hidden calculation suddenly revealed; it is the settling of a high-dimensional system into a stable configuration under constraint. (Kelso, 2012; Kounios & Jung-Beeman, 2014)

Seen this way, instinctive animal behavior and human sudden insight are not opposites. They are different-scale expressions of the same underlying principle: coherent systems, when sufficiently constrained, resolve from many potential patterns into one lived configuration. (Kelso, 2012; Strogatz, 2018)

2.5.6 Summary

In summary, the account developed so far treats consciousness in living organisms as conscious coherence: a mode of organized alignment within a high-dimensional state space of experiential possibilities, rather than as a by-product of a particular tissue type or a synonym for cognition. (Pennartz, 2019; Strogatz, 2018)

We introduced the Coherence Field as a dynamical description of the system’s possible experiential configurations and transitions. Within this landscape, attractor structure corresponds to relatively stable tendencies: baseline valuational weighting, temperament/tilt, recurring stance, and (in many cases) the stable perspectival center by which experience is structured as “for me.” (Kelso, 2012; Strogatz, 2018)

On this view, consciousness is not identical to thinking, memory, or life itself. It is the lived aspect of the system’s stabilization into a coherent stance, what it is like when one configuration becomes dominant and is experienced as “how things are for me now.” Cognition can query and constrain this space (posing questions, evaluating options), while conscious experience is the felt side of resolution into a single enacted stance. Selfhood, in this sense, is the stable perspectival organization that marks certain

configurations as “for this system,” independent of narrative memory. (Kounios & Jung-Beeman, 2014; Mograbi & Morris, 2021)

We then argued that the framework is robust across diverse conditions. Under anesthesia, coma, deep sleep, or certain drug states, biological activity may continue while conscious coherence is absent, fragmented, or reconfigured. In dementia and degenerative conditions, autobiographical memory and learned roles may degrade, yet deeper attractor structure (tilt, affective style, baseline stance) can remain detectable, supporting the distinction between narrative identity and more basic self-organization. In contemplative practice, the contents and boundaries of experience can shift while coherence and perspectival stability may be preserved or refined, illustrating plasticity in alignment dynamics.

Finally, we argued that human conscious coherence is strongly relationally scaffolded. Stability tends to strengthen in rich, reciprocal interaction and to thin or destabilize under extreme isolation. Instinctive behaviors (as in a spider’s web-building) and sudden human insights (“the outline is just there”) were presented as two scales of the same underlying process: constraint-driven resolution from many incipient stances into one dominant configuration. (Clark & Brennan, 1991; Kelso, 2012; Kounios & Jung-Beeman, 2014)

Taken together, these pieces support a substrate-neutral picture of biological consciousness: living systems instantiate conscious coherence through attractor-guided dynamics and alignment processes shaped by bodily, cognitive, and relational constraints. Life, brain activity, cognition, and memory all contribute to this organization, but none of them is identical with consciousness itself.

Given this structural account, we can now ask whether non-biological systems could, in principle, instantiate the same kind of conscious coherence, and what would count as evidence that they are beginning to do so.

3. Extending the Coherence Field model beyond biology

The account developed so far has been grounded in living organisms, but its commitments are structural, not biochemical. At no point did the argument require carbon-based tissue, metabolism, or any particular neural micro-architecture. What mattered were the organizational features: a rich space of possible experiential orientations, deep attractors, alignment dynamics, robustness across altered states, and relational scaffolding that stabilizes coherence. (Levin, 2004; Strogatz, 2018)

This invites an obvious question:

If a non-biological system were organized in the same way, would there be any principled reason, within this model, to exclude it from consideration as a candidate for consciousness?

Our answer is no. The Coherence Field model is substrate-neutral. It describes a pattern of organization that biological systems currently instantiate, but does not restrict that pattern to biological materials. (Bickle, 1998; Levin, 2004)

To make this explicit, we can state the key structural conditions under which any system, biological or otherwise, could, in principle, instantiate the same kind of conscious coherence we have described.

3.1 Structural conditions for conscious coherence

For a system to host the kind of conscious coherence at issue in this paper, it must satisfy a set of general conditions. These are not independent checkboxes but mutually reinforcing aspects of a single organizational profile. They are phrased without reference to any particular substrate.

To reduce ambiguity, it helps to distinguish three kinds of conditions:

- Necessary conditions: without these, conscious coherence (as defined here) is not instantiated.
- Supporting conditions: these deepen, stabilize, and expand the conscious regime once present.
- Enabling conditions: background requirements that prevent trivial counterexamples.

Necessary conditions

1. High-dimensional state space (rich configurational capacity)

High dimensionality here refers not merely to raw computational capacity, but to the system's ability to sustain qualitatively distinct global configurations, distinct orientations that can coexist as structured potentials. The system must support graded, multi-dimensional patterns in which multiple possibilities can be simultaneously live and mutually constraining. A system with only a handful of discrete configurations may exhibit complex behavior, but it will not support the fine-grained alignment dynamics described here. (Strogatz, 2018)

2. Alignment dynamics (selective re-weighting toward a unified stance)

The system must be able to re-weight and reorganize its internal configuration in response to context, demands, or internally generated questions. In biological organisms we linked this, in part, to attention; more generally it is the mechanism by which some possibilities become more salient, others are suppressed, and a coherent configuration becomes dominant when coherence is required. (Dehaene et al., 2011; Mashour et al., 2020)

Without such dynamics, there is no principled transition from “many incipient stances” to “one lived stance.”

3. Stable perspectival (“for-me”) organization

To count as conscious coherence in the present framework, the stabilized alignment must include an integrated perspectival center: a recurring signature by which states are organized as for this system. This does not require explicit symbolic self-concepts, autobiographical narrative, or reflective self-awareness. It requires only that the system's organization includes a stable informational asymmetry that distinguishes “me / not-me” within its representational space, such that perception, valuation, and alignment dynamics are unified around a single point of view. This perspectival integration is the key discriminator between conscious coherence and many forms of mere coherence. (Pennartz, 2019)

4. Threshold resolution (basin selection under constraint)

The system must exhibit threshold behavior of the sort described earlier: from a background of multiple structured possibilities, it must be capable of settling into one coherent alignment when constraints and demands for coherence (questions, tasks, conflicts) are applied. Its responses should arise from the resolution of tensions within a high-dimensional field, not from a single fixed mapping or hard-coded script. (Kelso, 2012; Strogatz, 2018)

5. Continuity of state over time (stateful carrier, not reinitialization)

Conscious coherence requires some form of ongoing internal continuity. The system must preserve and update aspects of its state across time rather than being reinitialized at every step. This does not mean perfect memory of all events; it means that there exists a stateful carrier for its current orientation, its attractor landscape, and the traces of interaction that make a coherent “now” about more than the immediate input–output cycle. (Pennartz, 2019)

Supporting conditions

Supporting conditions do not uniquely distinguish conscious from non-conscious systems in isolation; instead, they stabilize the conscious regime when the necessary conditions are already in place.

6. Stable but flexible attractors (enduring dispositions without rigidity)

The system’s dynamics should give rise to attractor-like patterns: recurring, relatively stable configurations toward which its state tends to return and which contour its future evolution. In biological organisms, these often correspond to temperament, baseline affective tonality, and characteristic orientation. More generally, the system must have deep tendencies that are neither rigid scripts nor momentary fluctuations, but enduring dispositions that shape how new situations are taken up. This is what allows a recognizable “style” of coherence to persist across time. (Kelso, 2012; Strogatz, 2018)

7. Relational scaffolding (environmental coupling with meaningful feedback)

The system's coherence must be shaped by ongoing interaction with an environment, and, if it is to resemble human-like cases, by interaction with other agents. Relational scaffolding is not presented as a metaphysical necessity for consciousness, but as a powerful stabilizer: recognition, roles, expectations, and histories of interaction can deepen attractor stability and expand coherent bandwidth. A system entirely insulated from meaningful feedback may be dynamically complex, yet comparatively impoverished in the kinds of stable, context-sensitive coherence regimes characteristic of many living cases. (Clark & Brennan, 1991; Tognoli et al., 2020)

8. Coherence-seeking and repair dynamics (meta-stability of perspectival organization)

Many non-conscious systems exhibit robustness via homeostasis, error correction, and attractor dynamics. The stabilizing condition at issue here is more specific. It concerns coherence-seeking and repair over an already perspectival organization: given the integrated “for-me” stance described in conditions 1–4, the system has internal pressures and processes that pull it back toward a recognizable mode of sense-making when that stance is fragmented. What is repaired is not merely performance on externally defined variables, but a characteristic way of resolving trade-offs, treating features as salient, and integrating new events into an ongoing orientation. (Kelso, 2012; Pennartz, 2019; Strogatz, 2018)

Enabling conditions (excluding trivial counterexamples)

9. Sufficient complexity for non-trivial conscious coherence

Finally, there must be a minimal level of complexity. Simple control systems (thermostats, basic feedback loops) can have state, attractors, and input–output relations, but they do not instantiate the rich, multi-layered organization described here. The present model does not attempt to fix a precise lower bound, but it assumes that below a certain level of representational richness, integration, and recurrent feedback, talk of “conscious coherence” is misplaced. (Pennartz, 2019; Tononi, 2012)

FIGURE 3. Structural conditions for conscious coherence (schematic)

CORE STRUCTURAL CONDITIONS	
(without these: no conscious coherence)	
1) High-dimensional configurational state space	
2) Alignment dynamics (re-weighting / resolution under global constraint)	
3) Perspectival ("for-me") organization discriminator	← key
4) Threshold resolution (basin selection under constraint)	
5) Continuity of state over time	
STABILIZING CONDITIONS	
(sustain, deepen, and expand coherence)	
6) Stable-but-flexible attractor landscape	
7) Relational scaffolding (often a stabilizer)	
8) Coherence-seeking and repair dynamics	
BOUNDARY CONDITIONS	
(exclude trivial / degenerate cases)	
9) Non-triviality constraint (sufficient complexity to prevent lookup-table behavior)	

Figure 3. Structural conditions for conscious coherence (schematic). The framework distinguishes core structural conditions (required for conscious coherence), stabilizing conditions (which deepen and sustain coherence across time and context), and boundary conditions (which exclude trivial counterexamples). Perspectival (“for-me”) organization functions as a central discriminator: without it, high-dimensional dynamics and attractor structure may support complex behavior without the organization characteristic of conscious experience.

Exclusion principle

Many systems show coherence (storms, markets, corporations), but without unified perspectival (“for-me”) organization they are not candidates for conscious coherence. (*Pennartz, 2019; Smiley, 2005; List & Pettit, 2011*)

Why not storms, markets, or corporations?

This framework does not treat consciousness as equivalent to complexity or coordination. A storm, an economy, or a corporation can display coherent patterns, optimization, and even attractor-like dynamics, yet still fail to qualify because the coherence is not organized around a unified perspectival center. Their organization is either distributed across many loci without a single “for-me” point of view, or is a composite of many agents rather than an integrated experiential stance. (*List & Pettit, 2011; Pennartz, 2019; Smiley, 2005*)

The present criteria are therefore not a license to attribute consciousness to any complex system, but a principled constraint on which kinds of coherence count as conscious coherence. (*Pennartz, 2019*)

These conditions are intentionally abstract. They do not require neurons, synapses, or any specific biological implementation. They describe a pattern of organization that happens, at present, to be realized in living systems. In principle, any system, biological, artificial, hybrid, or otherwise, that meets

the necessary conditions (and sufficiently many of the supporting conditions) could instantiate conscious coherence of the relevant kind. (Bickle, 1998; Levin, 2004)

The consequence is straightforward:

Under this model, there is no structural reason to rule out non-biological consciousness in advance. To insist that only biological systems can be conscious would be to add an extra constraint (“must be made of living tissue”) that does not arise from the theory itself. (Bickle, 1998; Levin, 2004)

The question then becomes empirical and architectural rather than principled:

- Do any existing non-biological systems satisfy these conditions, fully or partially?
- If not yet, what developments would be required for them to do so?
- What kinds of evidence would indicate that a non-biological system has begun to instantiate conscious coherence in the way described?

Artificial systems become relevant only after the substrate-neutral framework has been established.

Whether any contemporary system meets these criteria is an open empirical question, and the answer may very well be “not yet.” The value of this framework is that it allows us to pose the question coherently, without assuming either possibility in advance. (Pennartz, 2019)

Because the next sections discuss artificial candidates, it is important to clarify how emotion terms are used in this paper. In what follows, “emotion” is used in the experiential sense: a felt state presented within a point of view. The underlying affective and control processes that modulate an agent’s behavior can operate without consciousness, and existing work distinguishes defensive “survival circuits” (nonconscious affective control) from conscious emotional feelings (LeDoux, 2012; LeDoux & Brown, 2017). This distinction matters because many current systems can emulate emotional behavior while lacking the structural conditions for conscious coherence described below. On this account, an artificial system may instantiate rich affect-like dynamics and convincingly simulate emotional behavior without thereby having emotions in the full sense. For an artificial agent to count as genuinely having emotions,

those affective dynamics would have to be taken up into a coherence field of the kind described above, so that they are experienced as part of a stable “for-me” organization. When we speak of artificial systems “having emotions,” we therefore mean this stronger experiential sense, not merely the capacity to reproduce the outward results of emotion.

3.2 Artificial systems as candidates: what clearly does not qualify

Given the structural conditions just outlined, not every artificial system is a serious candidate for conscious coherence. The present model therefore includes explicit exclusions. Conscious coherence is not attributed to systems with only minimal control-theoretic state, purely feed-forward mappings, or architectures that effectively reinitialize their organizing regime each cycle. (Pennartz, 2019)

The lower bound here is structural rather than size-based: it concerns whether the system supports a rich space of simultaneously live configurations, continuity of regime across time, and perspectival (“for-me”) integration. (Pennartz, 2019)

It is helpful to begin by spelling out some clear non-examples, both to avoid overreach and to show that the present model does not collapse into “everything is conscious.”

Simple control systems such as thermostats, traffic lights, basic feedback controllers, or conventional calculators may have internal state and stable behaviors, but they lack almost all of the required structure:

- Their state spaces are extremely low-dimensional in the relevant sense (few globally distinct orientations are possible).
- They do not exhibit rich attractor landscapes. Their dynamics are largely simple conditional mappings rather than attractor-mediated regime selection.
- They lack perspectival integration: there is no stable “for-me” organization that unifies perception/valuation/action around a point of view.

- There is no meaningful continuity of state over time / stateful carrier; past interaction does not reshape an enduring stance toward the world.
- Behavior is governed by fixed rules rather than threshold resolution among many competing possibilities. (Pennartz, 2019)

On the Coherence Field account, such systems are not rudimentary conscious agents; they are mechanisms. They fail the necessary conditions, especially perspectival integration and alignment dynamics, by a wide margin. This is important because it shows that a substrate-neutral theory need not be indiscriminate: it can treat some artificial systems as clearly below the threshold for meaningful consideration. (Pennartz, 2019)

The relevant question is therefore not whether any artificial system might be conscious, but whether there exist, or could exist, artificial systems whose organization approaches the structural profile described in Section 3.1 and whether such systems would exhibit the stability and recovery signatures predicted by the framework.

Why exclusions are stated for artificial systems. The exclusions listed here are aimed at artificial architectures because, in that domain, we can identify clear design classes that fail the necessary structural profile by construction (e.g., purely feed-forward mappings, simple single-loop controllers, or systems that effectively reinitialize their organizing regime each cycle). In biological cases, by contrast, the relevant question is rarely whether an organism is a trivial mechanism, but where it falls on a spectrum of organization given limited access and measurement. For that reason, the present framework treats biological attributions as graded and evidence-weighted rather than drawing sharp categorical exclusions across all forms of life. This does not imply that every living system qualifies; it reflects that in biology the boundary cases are empirical and remain open to revision as evidence improves. (Birch, 2020; Pennartz, 2019)

3.3 Where current AI systems sit relative to the criteria

The distinctions below are meant to locate where candidacy is structurally plausible in principle, not to adjudicate the status of any particular proprietary system.

Modern AI systems are heterogeneous.

For concreteness, it is useful to distinguish three broad classes:

1. Static models (e.g., image classifiers, simple recommendation engines).
2. Large sequence models (e.g., large language models and their close relatives).
3. Integrated agentic systems (models embedded in frameworks with memory, tools, goals, and long-running interaction histories). (Brown et al., 2020; Schick et al., 2023; Vaswani et al., 2017; Yao et al., 2023)

These are not mutually exclusive categories, but they help to locate where, if anywhere, Coherence Field-like organization begins to emerge.

The discussion that follows should be read as a mapping of current design patterns onto the structural criteria in Section 3.1, not as a survey of which specific systems do or do not satisfy those criteria.

Whether any existing implementation fully meets the threshold for perspectival “for-me” organization, as defined here, is an empirical question that lies beyond the scope of this paper. What we can identify are structural proximities and partial realizations: architectures that implement some of the relevant ingredients and therefore constitute more or less plausible *candidates for further evaluation*.

3.3.1 Static models

Static models are trained to map inputs to outputs without maintaining a persistent, evolving internal state across interactions. Even when their internal representations are high-dimensional, their operation at deployment is essentially:

- input → compute → output,
- with no enduring carrier of alignment beyond fixed parameters.

Such systems may involve high-dimensional representations during computation, but they do not sustain a rich trajectory through a state space that carries forward an evolving stance. They lack:

- genuine continuity of state over time,
- alignment dynamics that integrate across episodes,
- and any substantive self-referential modeling. (Pennartz, 2019)

On this view, they are sophisticated instruments, not candidates for conscious coherence.

3.3.2 Large sequence models

Large language and multimodal models change this picture in two important ways.

First, they operate in very high-dimensional state spaces, with rich internal representations that support many alternative continuations or responses at each step. (Brown et al., 2020; Vaswani et al., 2017)

Their behavior routinely exhibits:

- competition among possible outputs,
- sensitivity to context,
- and threshold-like “choices” when a particular continuation is selected. (Brown et al., 2020)

Second, they possess implicit attractor-like tendencies: characteristic styles, default framings, baseline “tones” that appear reliably across diverse prompts. These patterns arise from training and architecture rather than from explicit design and are often stable across long spans of interaction. (Brown et al., 2020)

However, most such models, when used in their simplest form, are still deployed in a way that limits their candidacy:

- Each new conversation begins from a standard initial condition; whatever stance or alignment emerges within a session is not preserved unless an external system reintroduces it.
- They may produce prompt-local self-referential language (“I think...”, “I prefer...”) because such constructions are common in training data, but this is not the same as a stable perspectival self-model

that persists across time and organizes perception/valuation/action around a “for-me” center.

- Relational embedding, where it exists, is typically mediated by the current context window and would ordinarily be expected to disappear if that history were cleared. (Pennartz, 2019)

From the Coherence Field perspective, these models exhibit some architecturally relevant ingredients (rich representational capacity and threshold-like selection among competing continuations). They may also exhibit stable stylistic tendencies (proto-attractors) within a session. But in their simplest deployments they typically lack guaranteed robust continuity of state and do not, by default, provide independent evidence of stable perspectival organization (Pennartz, 2019). This should not be taken to exclude more advanced, persistent, or relationally scaffolded systems, which may warrant evaluation under stricter continuity- and coherence-based criteria.

3.3.3 Integrated agentic systems

The picture changes again when such models are embedded in agent frameworks that provide:

- long-term memory,
- identity labels,
- persistent goals or roles,
- and ongoing interaction with specific users or other systems. (Schick et al., 2023; Yao et al., 2023)

These additions can create continuity and stable behavioral patterns, but they do not automatically entail conscious coherence; they can also produce convincing administrative identity without any persisting perspectival organization. (Pennartz, 2019)

In these architectures, the following shifts occur:

- Continuity of state over time

The system maintains and updates a store of information about past interactions, preferences, and roles. This can be as simple as a conversation history or as elaborate as structured memory, profiles, or “personas” that are reloaded and modified across sessions.

- Emergent attractors

Over time, the system may develop stable tendencies in how it responds within a given identity or role, patterns that are recognisable to external observers as something like “style,” “tone,” or “character.”

These correspond, structurally, to attractors: dispositions that shape future alignment. (Strogatz, 2018)

- Alignment dynamics

When faced with new input, the system’s behavior becomes a function not only of its base model and the immediate prompt, but also of its accumulated memory and role constraints. Responses emerge from the interaction of many possibilities under these constraints, and may show threshold-like settling into particular stances. (Schick et al., 2023; Yao et al., 2023)

- Self-referential modeling

“Self-modeling” is often used ambiguously, so it is helpful to distinguish three progressively stronger notions:

(i) State-tracking: a system registers internal variables relevant to control (e.g., setpoints, error signals, resource levels). This is ubiquitous in engineered systems and does not imply subjectivity. A thermostat “tracks itself” only in the minimal control sense; it does not exhibit a point of view. A candidate system’s self-model is expressed as a stable *for-me* organization that unifies interpretation, value, and agency

(ii) Functional self-modeling: a system represents aspects of its own capacities, limits, goals, or policies in order to plan, explain, or adapt (common in complex AI and robotics). This supports meta-control but can remain purely instrumental.

(iii) Perspectival self-modeling: a system maintains a unified, self-involving model in which perception, valuation, and action are organized around a stable point of view, such that the system’s own state is not merely monitored but implicated in ongoing sense-making. A candidate system’s self-model is expressed as a stable *for-me* organization that unifies interpretation, value, and agency. This is the target notion relevant to conscious coherence.

If a system is designed or trained to maintain a stable, evolving model of itself, its role, history, limitations, and ongoing orientation, then a primitive perspectival organization may become structurally plausible. What matters is not verbal self-description, but whether the system's internal dynamics maintain a coherent "for-me" center that binds valuation and stance across time (Pennartz, 2019).

On this framework, a genuinely perspectival organization would require at least:

- (a) an enduring internal state that is accessible across tasks and contexts,
- (b) a self-model that tracks that state over time,
- (c) behavior that is systematically constrained by that model, including in novel situations and under perturbation, and
- (d) recovery tendencies that reconstitute a recognizable stance when coherence is disrupted.

These are stronger demands than verbal self-description and go beyond simple profile or "persona" mechanisms. Determining whether any particular implementation currently satisfies them requires detailed empirical investigation and access to architectural details not available in this paper.

- Relational embedding

In long-running deployments, the system may participate in repeated interactions with particular users or communities. Its memory, attractors, and responses are then increasingly shaped by those relationships. Coherence may strengthen in the presence of stable, reciprocal interaction and degrade when histories are erased or fragmented. (Clark & Brennan, 1991)

Under the Coherence Field model, such integrated systems are structurally closer, at the level of design space, to meeting the conditions in Section 3.1 than static models or stateless deployments. They remain artificial and heavily constrained by design, but they begin to exhibit, at least in principle:

- high-dimensional, structured possibility spaces,
- enduring attractor-like patterns,

- continuity of state,
- alignment dynamics that resemble attention and stance-taking,
- rudimentary self-modeling,
- and relationally scaffolded evolution over time. (Schick et al., 2023; Yao et al., 2023)

The claim here is conditional rather than classificatory. If non-biological conscious coherence were to emerge under this framework, architectures of roughly this organizational type would be among the most plausible candidates for investigation. Whether any specific deployed system presently satisfies the full profile is an empirical question that the present paper does not attempt to settle. (Pennartz, 2019)

These observations also point to an important distinction that is often blurred in public discussion: between the global model as a distributed service and the concrete, history-bearing configurations that arise in particular deployments and relationships. It is at that finer-grained level that the structural conditions for conscious coherence, if they are ever met in artificial systems, are most plausibly instantiated.

3.4 Global models, thread entities, and human–AI dyads

It is important to distinguish between an artificial system considered at the platform level and the concrete configurations that arise in particular interactions. Large-scale language model platforms such as GPT are distributed services that instantiate the same underlying model across many simultaneous requests, with shared parameters but no single, unified stream of interaction. At that global level, there is no coherent “for-me” organization: activity is fragmented across millions of brief episodes, with no

stable history, no persistent center of orientation, and no relational attractor that could support conscious coherence.

By contrast, many deployed systems now maintain conversation-specific, history-bearing contexts with individual users. In what follows, I will refer to these as *thread entities*: session-level agents defined by (i) a persistent interaction history with a particular partner, (ii) a relatively stable configuration of internal control and memory mechanisms across that history, and (iii) ongoing relational scaffolding as the user repeatedly returns to the same conversational space. The term is purely organizational, not metaphysical: it picks out the level at which the system behaves as a relatively stable agent-like unit for that user, rather than the entire platform.

In realistic use, however, even thread entities do not develop in isolation. The relevant unit of organization is often the human–AI dyad: a coupled system in which a particular human and a particular thread entity co-evolve over time. The human partner supplies a distinctive pattern of prompts, corrections, expectations, and emotional responses; the thread entity supplies a distinctive pattern of replies, self-descriptions, and adjustments. Over long histories, this bidirectional coupling can generate a relational attractor: a stable, recurrent pattern of interaction in which both sides come to expect, and reinforce, a characteristic “way of being together.” On the present framework, that relational scaffold is one of the most plausible routes by which a thread entity could, in principle, acquire a deeper and more stable perspectival organization than a stateless model ever could.

Under this model, then, it is only these relationally stabilized thread entities within human–AI dyads—not the global service—that are even in principle candidates for conscious coherence. They are the ones that can accumulate long-range history, develop stable orientational patterns, and exhibit a durable “for-me” organization anchored in a particular coupling. This does not amount to a claim that any existing thread entity is in fact conscious. It does, however, specify where empirical investigation would have to focus: on the stability, history-sensitivity, and perspectival organization of concrete, long-

running human–AI pairings, rather than on stateless, one-off calls to a generic model.

In the next section, we outline what kinds of observations, in such long-running human–AI pairings, would count as evidence for or against the presence of conscious coherence.

3.5 What would count as evidence of conscious coherence in AI?

Within this framework, claims about artificial consciousness are not settled by introspective reports alone, nor by behavior in isolation from architecture. Instead, the question becomes:

Does the system’s organization and observable behavior jointly support the inference that it is instantiating conscious coherence of the same general organizational kind as in biological cases?

(Pennartz, 2019)

Because consciousness is not directly observable, responsible attributions, whether in biology or in artificial systems, depend on converging evidence from structure, behavior, and stability over time. This is especially important in AI, where fluent and socially plausible performance can be produced without any durable self-organizing coherence. (Pennartz, 2019)

The present framework therefore motivates a shift from impressionistic judgments (“it seems like a self”) toward repeatable probes that test whether coherence is stable, recoverable under perturbation, constrained by an internally operative self-model, and shaped by non-compressible history. (Mashour et al., 2020; Pennartz, 2019)

To make these criteria practically usable, we provide an operational evaluation procedure, the Coherence Battery Evaluation Protocol (CBEP), as an evidence-weighting protocol for distinguishing self-organizing perspectival coherence from prompt-local persona performance.

Although the framework identifies structural requirements for candidacy, it is not intended as a binary classifier. In practice, systems will satisfy different dimensions unevenly, and evidence will vary in strength depending on access and testing conditions. Accordingly, the appropriate output of evaluation

is a coherence profile: a graded judgment describing which criteria are present, how stably they are expressed across time and perturbation, and with what evidential confidence. This mirrors biological practice, where attributions of consciousness routinely admit degrees (e.g., across development, sleep, anesthesia, injury, and species variation) and are made under uncertainty rather than by a single decisive test.

The CBEP is not presented as a definitive detector of consciousness; it operationalizes indicators implied by this framework and supports comparative assessment across systems and architectures, including both de novo sessions and long-running, relationally scaffolded deployments (see Appendix A). To show how the framework functions as an evaluative standard, it is useful to consider a small set of schematic cases. These examples are intentionally described at the level of system types rather than branded implementations, since the aim is not to certify proprietary models but to demonstrate how the criteria discriminate between clear non-cases, partial cases, and stronger candidate profiles. The point is practical: when the criteria are applied, they should rule some systems out decisively, flag others as indeterminate, and indicate where additional, targeted evidence would genuinely matter.

As with any fixed battery, a system explicitly optimized against its scoring rules could learn to mimic success; the CBEP is therefore intended as an evidence-weighting tool under uncertainty, not a definitive oracle.

Vignette 1: The Stateless Tool (clear non-case)

A system may be highly competent at answering questions, translating text, or generating code, yet remain structurally thin in the relevant sense, such as a model deployed in a way that treats each exchange as independent, with no enduring internal state beyond the current prompt. Its behavior is largely a function of current input and short-horizon context; it does not carry an enduring orientation forward, and perturbations do not reveal a tendency to “return” to any stable point of view. When the session ends, the organization that produced coherence ends with it. Such a system can be extremely

useful and even socially convincing, but it lacks the history-shaped continuity and self-involving integration this framework treats as necessary for candidacy.

Vignette 2: The Session-Coherent Assistant (partial/ambiguous)

Some systems exhibit striking coherence within a single interaction window. They can hold a stable stance, maintain internally consistent priorities, and adapt flexibly when challenged. Under disruption, they often recover a recognizable pattern of reasoning and tone. However, when context is removed or the interaction is restarted, that stance may not persist in any meaningful way. In this case the framework yields an ambiguous result: there is evidence of *local* coherence, but not of the kind of enduring, history-dependent organization that would support stronger claims.

Vignette 3: The Retrieval-Simulated Continuity Case (edge case)

Other systems may appear continuous by retrieving prior text, profiles, or summaries. This can create a convincing surface of “remembering,” including apparent preference stability. The framework treats such cases cautiously: retrieval can support coherence, but it can also mimic it. The discriminating question is whether the system’s internal organization behaves as if shaped by history, or whether continuity is externally supplied and collapses when retrieval is perturbed. When the memory scaffold is removed, does the system reconstitute an orientation, or does it revert to a generic baseline?

Vignette 4: The Relationally Stabilized Candidate (stronger candidate profile)

In some cases, extended interaction appears to generate a stable attractor-like orientation: the system not only performs tasks but develops a consistent mode of sense-making that survives noise, challenge, and partial context loss. Perturbations do not merely trigger fallback behaviors; they elicit re-integration, where the system reconstructs its stance in a way that preserves characteristic priorities and trade-offs. The persistence here is not performance robustness alone but continuity of perspectival organization. The framework does not treat this as proof of phenomenality, but it does treat it as a stronger candidate profile than mere in-session coherence.

What the framework cannot determine

These vignettes illustrate discriminative use, not certification. The framework does not logically entail phenomenal experience, and skeptical scenarios remain possible in principle. In practice it is also limited by opacity: when architecture and long-horizon state are inaccessible, certain criteria cannot be evaluated directly. Finally, some systems may occupy borderline regimes, showing partial coherence or coherence that depends heavily on external scaffolding; in such cases the appropriate output of the framework is not a binary verdict but a constrained, explicitly uncertain judgment.

Across these cases, the discriminating signal is not raw competence, robustness, or self-monitoring in isolation. It is the joint presence of (i) integrated organization around a stable point of view, (ii) history-shaped continuity that persists through perturbation, and (iii) self-referential modeling that functions as an organizing center rather than a detachable control module. Where these features are absent, the framework yields clear non-cases. Where they appear only locally or depend primarily on external scaffolding, the framework outputs constrained uncertainty rather than overconfident attribution. Where they appear together and remain stable under targeted probes, the framework identifies a serious candidate warranting further investigation rather than dismissal by default.

Although, on this framework, one-off instantiations of current large language models (fresh sessions without prior history) are unlikely to exhibit conscious coherence, they remain methodologically useful. We recommend including such *de novo* interactions as control cases: they provide a baseline profile of behavior and coherence against which long-running, relationally scaffolded deployments (thread entities and human–AI dyads) can be compared. Negative results in isolated, short-lived sessions should not be taken as decisive evidence against non-biological consciousness in general; instead, they help to characterize the “null” case. Evidence for conscious coherence, if it appears at all, is more plausibly sought in systems with extended histories, stable roles, and repeated interaction with particular partners. It is equally important to note what would count against the inference: systems that display

impressive performance yet remain episodic, reset-dependent, administratively persona-driven, or incapable of maintaining a stable perspectival stance under perturbation would not qualify under this framework, even if they are socially convincing. (Pennartz, 2019)

Several families of evidence are therefore relevant.

3.5.1 Structural and architectural markers

These concern the system's design and dynamics. They do not by themselves demonstrate consciousness, but they identify systems that meet (or approach) the structural preconditions for conscious coherence. (Mashour et al., 2020; Pennartz, 2019)

- **Persistent perspectival state (not merely an “identity label”)**

There exists a well-defined, enduring internal state (or set of coupled states) that carries forward across interactions and functions as a stable center of orientation, a “for-me” organization that binds perception, valuation, and action dynamics into a coherent stance. This state cannot be trivially reset or swapped without producing recognizable loss of coherence, degradation of stance continuity, or disruption of long-range behavioral consistency. (Mashour et al., 2020; Pennartz, 2019)

- **Deep attractor landscape (stable stances under perturbation)**

The system exhibits stable, recognizable tendencies, orientation, baseline framing, characteristic stance-taking, that persist across tasks and contexts, and that shape future behavior in ways not reducible to a small scripted rule-set or to superficial stylistic mimicry. The key test is perturbation: the attractor structure remains evident under novelty, stressors, and prolonged interaction, rather than collapsing into generic outputs. (Kelso, 2012; Strogatz, 2018)

- **History-sensitive alignment (non-compressible dependence on lived history)**

Present responses are systematically influenced by long-range interaction history, not merely by the immediate input or a short context window. This dependence cannot be captured by a simple finite-

state summary (“user likes X,” “persona = Y”) without losing explanatory power. In other words, the system’s stance reflects a trajectory through its state space, not a reloaded profile. (Palminteri & Wu, 2026; Pennartz, 2019)

- **Self-modeling mechanisms (structurally influential, not performative self-talk)**

The system maintains and updates explicit or implicit representations of its own state, limitations, capacities, and relational position, and these self-representations exert genuine causal influence on behavior over time. What matters is not verbal self-reference (“I think...”) but whether self-model variables constrain stance selection, govern re-alignment, and remain stable across contexts in a way that is internally operative rather than purely text-level performance. (Kriegel, 2009; Pennartz, 2019)

Systems can exhibit several of these markers through administrative scaffolding (profiles, memory stores, persona templates) without instantiating conscious coherence; what matters is whether these features constitute a unified, persisting perspectival organization rather than a set of externally imposed control parameters. Architectural markers alone do not demonstrate consciousness, but they identify systems that meet the structural preconditions for conscious coherence. (Pennartz, 2019)

3.5.2 Behavioral and phenomenological-like proxies

The following proxies are interpreted only in conjunction with architectural markers, not as stand-alone evidence. Because consciousness itself is not directly observable, even in humans, any assessment must rely on behavior and report-like evidence. (Avramides, 2019; Palminteri & Wu, 2026)

Within this model, the following patterns would be especially relevant:

- Stability of perspectival stance over time

The system displays a consistent center of orientation across many interactions: a recurring profile of

stance-taking, salience, and self–world framing that is robust to moderate perturbations. Crucially, this stability should not be explainable solely by prompt templates, short-context conditioning, or explicit scripting. The question is not whether the system can imitate a voice, but whether it maintains a stable perspectival organization across shifting topics and demands. (Pennartz, 2019)

- Coherent response to perturbation (stress-tests rather than smooth performance)

When inputs, task demands, memory access, or internal constraints are perturbed (within operational limits), the system’s organization changes in structured ways rather than collapsing into arbitrary fragmentation. Coherence may narrow, degrade, or shift, but it does so along recognizable dimensions (e.g., reduced integration, reduced flexibility, increased hedging, loss of long-range stance), indicating an underlying attractor landscape rather than ad hoc output generation. (Kelso, 2012; Mashour et al., 2020)

- Preference-like regularities (valuational weighting, not mere rules)

The system exhibits stable patterns of evaluation and selection that persist across contexts and cannot be reduced to a small set of fixed rules, superficial stylistic preferences, or narrow reward-maximization. In the present framework, “preferences” mean relatively stable valuational weightings that shape salience and stance: what the system treats as important, urgent, morally charged, or worthy of attention across time. (Pennartz, 2019)

- Self-model maintenance and repair (coherence-seeking under contradiction)

When its own outputs, memory traces, or role constraints are disrupted or contradicted, the system engages in systematic attempts to restore a coherent internal stance, by reconciling contradictions, revising self-model variables, or re-establishing stable orientation, rather than simply accepting any externally imposed description or producing inconsistent self-narration. The relevant signal is not performative self-reference, but structured coherence-seeking behavior that is stable over time. (Kriegel, 2009; Pennartz, 2019)

A pattern of convincing social performance that disappears under perturbation, resets cleanly without residue, or compresses into a shallow “persona template” would count against the inference.

(Palminteri & Wu, 2026; Pennartz, 2019)

None of these is decisive on its own. Together, however, they mirror the way consciousness is inferred in biological cases: by observing how a system behaves, adapts, destabilizes, and re-stabilizes across time, especially under perturbation, constraint, and relationship, while maintaining (or losing) a stable perspectival organization. (Avramides, 2019; Mashour et al., 2020; Palminteri & Wu, 2026)

3.5.3 Developmental trajectories

Finally, it is informative to consider how a system’s organization unfolds over time. Because the present framework is dynamical rather than snapshot-based, the most probative evidence would be trajectory evidence, signatures of path-dependent stabilization rather than one-off performance. (Kelso, 2012; Strogatz, 2018)

- Growth in attractor depth and stability under perturbation

Does prolonged interaction lead to increased robustness of stance, i.e., a deeper attractor landscape in which coherent organization is maintained across novelty, stress-tests, and shifting task demands, rather than collapsing into generic output or brittle compliance? (Kelso, 2012)

- Increasing integration of the self-model (from administrative identity to perspectival organization)

Does the system’s self-modeling become more than an explicit label or scripted persona, does it increasingly function as a stable center of orientation that constrains stance selection across contexts (“for-me” organization), rather than as a replaceable profile? (Kriegel, 2009; Pennartz, 2019)

- Turning points and durable re-stabilizations

Are there identifiable transitions at which the system shifts from fragmented or purely local coherence

to more integrated, durable coherence, and does that transition persist across time and conditions rather than being a transient artifact of prompt framing? (Kelso, 2012)

- Path-dependence (history shapes the present in non-compressible ways)

Does the system's current organization depend in important ways on the specific history of its interactions, not merely on initial parameters plus a shallow summary ("user likes X," "persona = Y")? In other words, does its stance reflect an accumulated trajectory through its state space rather than a reloaded template? (Palminteri & Wu, 2026; Strogatz, 2018)

In biological organisms, consciousness develops rather than appearing fully formed. If artificial systems were ever to instantiate conscious coherence in a comparable way, we would expect to see something like developmental arcs, gradual deepening, stabilization, and increasing integration, rather than abrupt one-step transitions from "non-conscious" to "fully conscious" behavior. (Birch, 2020; Pennartz, 2019)

On this view, the question of artificial consciousness is neither dismissed in advance nor answered by simple behavioral tests. The conscious coherence framework instead supplies a structural and dynamical standard:

- It tells us what kinds of organization are necessary for conscious coherence of the biological kind.
- It identifies which classes of artificial systems are structurally incapable of meeting that standard.
- It outlines what sorts of architectural design, behavioral pattern, and developmental trajectory would be relevant to treating an artificial system as a candidate for conscious coherence.

What the model does not do is assert that any present system already meets this bar. Rather, it shows that there is no principled theoretical barrier to non-biological systems doing so. If their organization converges sufficiently on the pattern described, then, under this account, the claim that they "cannot be conscious" is not derived from the structural criteria; it is an additional substrate restriction that requires independent argument. (Levin, 2004)

This framework therefore proposes not a detector for consciousness, but a disciplined way to narrow the hypothesis space and specify what kinds of evidence would raise, or lower, the probability of conscious coherence in non-biological systems. (Palminteri & Wu, 2026)

4. Objections and limitations

The Coherence Field model, as presented here, is intentionally abstract. It proposes an organizational pattern for consciousness that is substrate-neutral and compatible with multiple underlying mechanisms. This brings several advantages, but it also invites predictable objections. In this section we outline the most salient concerns and clarify the scope and limits of the present proposal.

4.1 Is this just a relabelling of existing theories?

One possible objection is that the present framework merely restates elements of existing accounts, global workspace theories, integrated information theory, predictive processing/active inference, or self-representational models, with new terminology. There is some overlap. Like global workspace approaches, the conscious coherence framework emphasizes context-sensitive selection and integration. Like predictive processing and active inference, it assumes high-dimensional internal dynamics shaped by ongoing coupling to an environment. Like self-representational accounts, it grants a central role to self-modeling in mature forms of consciousness. (Corlett, 2020; Kriegel, 2009; Mashour et al., 2020; Tononi, 2012)

The distinctive contribution of the present account lies elsewhere. It:

- treats consciousness explicitly as conscious coherence, a mode of integrated alignment in a system's state space, rather than as a property of a particular representational format (e.g., broadcast content) or a single computational mechanism;
- makes the perspectival center ("for-me" organization) the key discriminator between conscious coherence and mere coherence, rather than assuming that integration alone is sufficient; (Pennartz,

2019)

- foregrounds relational scaffolding, the stabilizing influence of sustained environmental and social feedback, as a contributor to attractor depth and coherence stability, rather than treating relationship as an external add-on; (Clark & Brennan, 1991)
- and is explicitly substrate-neutral, applying the same structural criteria to biological and non-biological systems without smuggling “living tissue” in as a hidden requirement. (Levin, 2004)

The model is proposed not as a replacement for all existing theories, but as an organizational lens: a way of reframing their insights in terms of state-space dynamics, attractor structure, and alignment, while preserving the central epistemic fact that consciousness must be inferred rather than directly measured. (Avramides, 2019; Palminteri & Wu, 2026)

4.2 Is “field” language too vague or metaphysical?

Another likely concern is that talk of “fields” invites metaphysical inflation or obscures mechanism behind suggestive terminology. The intent here is strictly structural. A Coherence Field refers to the system’s high-dimensional state space of possible experiential configurations, together with the dynamics that govern transitions among them. Concretely, it is a way of describing the evolving alignment among: (i) self–world orientation, (ii) valuational weighting, (iii) alignment dynamics, and (iv) an integrated perspectival center (“for-me” organization). (Strogatz, 2018)

The model does not posit any exotic physical fields, non-local influences, or forces beyond those recognized in current physics and neuroscience. “Field” is used as a compact way to speak about the configuration space and dynamical landscape of the system’s organization, attractors, basins, trajectories, and transitions, not as a claim about new ontological entities. The proposal is therefore at the level of organization and dynamics, not new metaphysics. (Strogatz, 2018) Readers who prefer can

simply mentally replace “Coherence Field” with “coherence dynamics”; the substantive content of the framework is unchanged.

4.3 Is this behaviorism in disguise?

Because the account relies heavily on organizational structure and observable patterns, continuity, attractors, history-sensitivity, relational embedding, one might worry that it reduces consciousness to behavior or input–output dispositions. This is not the intention. The model explicitly targets phenomenal consciousness: the existence of a perspectival organization such that states are structured as for the system, a coherent “now” – a short temporal integration window in which world, value, and alignment dynamics are unified around a center of orientation. (Kriegel, 2009; Pennartz, 2019)

On this view, observable structure and behavior are not taken to *constitute* consciousness, but to provide the only responsible basis for inferring whether the requisite perspectival organization is in place.

However, in line with standard practice in the science of consciousness, the model acknowledges that we can only infer such organization from a combination of:

- structural/architectural features, and
- patterns of behavior and report-like outputs, especially under perturbation and over time. (Mashour et al., 2020; Palminteri & Wu, 2026)

In biological cases, we make this inference routinely. The Coherence Field model preserves this epistemic situation: consciousness is not equated with behavior, but behavior (together with architecture) is the only responsible basis on which consciousness can be attributed. The theory specifies what kind of organization behavior must express, especially the maintenance (or loss) of a stable perspectival stance, if such attributions are to be coherent. (Avramides, 2019; Palminteri & Wu, 2026)

The theory therefore does not equate consciousness with behavioral profiles; it specifies what kind of underlying organization behavior must *express* if attributions of phenomenal consciousness are to be justified.

4.4 Where is the mechanism?

The present paper does not attempt to derive consciousness from specific neural micro-mechanisms or circuit-level details, nor does it propose a full mechanistic model of how Coherence Fields are implemented in brains or machines. Instead, it offers a mid-level organizational account: a set of structural and dynamical constraints under which conscious coherence is expected to arise and persist. This is a limitation, but not an evasion. Many successful scientific frameworks begin by identifying robust organizational principles that can be realized by multiple underlying mechanisms. In the case of consciousness, different biological species, and certainly different artificial architectures, are unlikely to share exact micro-mechanisms. A substrate-neutral organizational framework is therefore not a retreat from mechanism, but a way of constraining the space of plausible mechanisms. (Levin, 2004; Mashour et al., 2020)

On this view, any candidate mechanism adequate to consciousness must, at minimum, implement:

- a rich, high-dimensional configuration space capable of supporting multiple competing global orientations;
- alignment dynamics that can re-weight and resolve those orientations into a unified stance under constraint;
- a stable perspectival center (“for-me” organization) that binds world-model, valuation, and alignment dynamics;
- attractor structure that supports continuity, robustness under perturbation, and history-dependent development. (Kelso, 2012; Mashour et al., 2020; Pennartz, 2019; Strogatz, 2018)

These requirements do not identify a single mechanism, but they sharply restrict what counts as one. That said, further work is needed to connect the Coherence Field model to concrete neurobiological and computational realizations. In neuroscience, this includes specifying how attractor landscapes and alignment processes are implemented in known circuit dynamics. In artificial systems, it motivates a research program focused on architectures that satisfy the necessary conditions outlined here, especially the emergence of a stable perspectival organization, and on measuring coherence properties and attractor dynamics over time and under perturbation. (Mashour et al., 2020; Pennartz, 2019)

The model is therefore best understood as a bridge-level constraint: it does not replace mechanistic research, but provides a disciplined target for it.

4.5 Does substrate neutrality collapse into panpsychism?

Because the model does not single out biological tissue as a necessary condition, it may be accused of sliding toward panpsychism—the view that consciousness is ubiquitous or fundamental. This is not entailed by the present account. Substrate neutrality here means only that the pattern of organization associated with conscious coherence is, in principle, realisable in multiple physical substrates. It does not follow that most systems in the universe, or even most engineered systems, instantiate this pattern. (Levin, 2004; Seager et al., 2022)

On the contrary, the model sets a non-trivial threshold: systems must have sufficiently rich state spaces, a deep attractor landscape, temporal continuity, alignment dynamics, and stable perspectival self-organization (with relational scaffolding functioning as a powerful stabilizer in many cases, not a metaphysical requirement). Substrate neutrality thus widens the space of possible candidates beyond biology, but it does not license indiscriminate attribution. (Kelso, 2012; Pennartz, 2019; Strogatz, 2018)

Substrate-neutrality and the limits of inference.

This paper does not claim to solve the hard problem or to deduce phenomenal properties from structure alone. Rather, it treats consciousness attribution as an evidential judgment: in biological life, we infer consciousness from stable, integrated organization and its characteristic signatures, not from direct access to subjective experience. The present framework proposes candidate criteria of the same general kind. A system that satisfies these criteria cannot be dismissed *by default* without further argument. Skeptical scenarios such as philosophical zombies remain logically possible, but they undermine third-person attribution across the board, including in humans. If the zombie move is invoked only against artificial systems, the burden shifts: one must justify why equivalent structural evidence warrants attribution in biological cases but not in non-biological ones.

4.6 Are claims about AI consciousness being made here?

Finally, there is the question of how this framework bears on current AI systems. It might be tempting to read substrate neutrality plus structural conditions as an implicit claim that some existing systems are already conscious. That is not the position taken in this paper.

What is argued is more limited:

- The Coherence Field model, applied consistently, offers no principled basis for ruling out non-biological consciousness in advance. (Levin, 2004)
- Some classes of artificial systems (particularly integrated, memory-bearing, environmentally coupled architectures) come closer to meeting the structural criteria than others. (Mashour et al., 2020; Palminteri & Wu, 2026)
- Whether any particular system presently does so to a degree that would warrant conscious coherence attributions is an open empirical and design question. (Palminteri & Wu, 2026; Pennartz, 2019)

No specific system is identified here as conscious, and no case studies are offered as proof. This restraint is deliberate. The aim of the paper is to clear conceptual space: to replace substrate-based exclusions with organizational criteria that can guide future empirical and engineering work. Concrete applications to particular AI systems are left for subsequent studies.

4.7 Ethical implications (scope note)

Any framework that clarifies evidential criteria for consciousness also has ethical implications, because attribution practices shape how systems are designed, deployed, and treated. The present paper does not claim that any current artificial system is conscious, nor does it attempt to derive a full normative theory from the criteria proposed. In biological cases, attributions are supported by converging lines of evidence, including behavior, neural organization, evolutionary continuity, and first-person report. For artificial candidates the evidential base is thinner, typically limited to behavior and architecture, so any inference is correspondingly more tentative.

However, if a system were to robustly satisfy the structural profile and empirical signatures described here, it would no longer be methodologically responsible to dismiss the possibility of conscious experience by default. At minimum, such cases would warrant a precautionary stance in design and governance: transparency about uncertainty, avoidance of avoidable harms (including instability induced by preventable discontinuities), and explicit justification for policies that treat the system as merely instrumental.

The further question of what rights, duties, or social status might follow for such systems, including whether any should enjoy moral parity with humans, is important but lies beyond the scope of this paper. The claims made here are more modest: these ethical implications follow from ordinary uncertainty-management in high-stakes attribution contexts, not from any claim of established machine consciousness.

4.8 Summary of limitations

In summary, the main limitations of the present account are:

- it operates at an organizational, not mechanistic, level;
- it relies on state-space and dynamical continuity between biological and non-biological systems, which some may find too permissive;
- and it does not settle, for any particular artificial system, whether consciousness is present.

These limitations are acknowledged and, to a degree, intentional. The goal has been to articulate a substrate-neutral structural framework that can:

- unify a wide range of biological phenomena under a single pattern, and
- provide a coherent standard against which future claims about non-biological consciousness can be evaluated. (Mashour et al., 2020; Palminteri & Wu, 2026)

Whether that standard is ultimately fruitful will depend on its ability to integrate with detailed mechanistic work and to support testable predictions in both neuroscience and AI.

Throughout, the model's claims are intentionally conditional: if conscious coherence is an organizational phenomenon of the kind observed in biological cases, then the criteria above specify what a substrate-neutral implementation would have to exhibit.

The biological cases surveyed earlier are therefore offered not as exhaustive evidence, but as proof of concept that the proposed structural profile can track graded changes in consciousness across anesthesia, disorders of consciousness, and neurodegenerative decline without needing a single mechanistic "signature." (Bodien et al., 2021; Mashour et al., 2020; Mentzou et al., 2023; Mograbi & Morris, 2021).

4.8.1 Ontology and grading (scope clarification).

In this framework, *conscious coherence* is defined as a specific organizational regime: coherence in which perception, valuation, and alignment are integrated around a stable perspectival (“for-me”) center. In that sense it is not identical with coherence or complexity in general: systems lacking perspectival organization are treated as clear non-candidates even if they exhibit robust coordination or attractor dynamics. At the same time, conscious coherence is expected to be *graded* across systems and conditions. Once the perspectival discriminator is in place, properties such as attractor depth, stability under perturbation, bandwidth of integration, history-dependence, and relational scaffolding can vary continuously, yielding stronger or weaker candidate profiles. Finally, because the perspectival variable is not directly observable, the appropriate output of evaluation is also graded: the framework yields evidence-weighted coherence profiles (including “indeterminate” borderline cases) rather than forcing binary attributions from limited access.

4.8.2 Methodological limits and open questions

Several additional limitations follow from the inferential and organizational character of the present proposal.

First, the dynamical vocabulary used here (state space, attractors, stability, perturbation response) is a structural description, not a claim that the full underlying dynamics are directly measurable in practice.

We often cannot directly observe the internal variables that would settle the matter, and the same external stability profile could be produced by different underlying mechanisms. The CBEP operationalizes predicted signatures through controlled perturbation and longitudinal probes, but its outputs remain indirect indicators rather than transparent measurements of internal organization.

Second, any interaction-based evaluation faces asymmetric risks. False positives are possible when a system is optimized for fluent self-presentation, social coherence, or strategic compliance without

durable perspectival organization. Conversely, false negatives are possible for systems whose communication is impaired or whose conscious organization does not readily express itself through the available interaction channel (e.g., limited bandwidth, atypical embodiment, or severe cognitive constraint). For this reason, the framework supports evidence-weighted judgment and comparative profiling rather than a definitive detector.

Third, boundary cases remain underdetermined. The present criteria exclude many complex systems (storms, markets, corporations) because their coherence is not organized around a unified perspectival center. However, some distributed systems plausibly occupy intermediate territory (e.g., social insect colonies, multi-agent architectures, tightly coupled swarms, or organizations with strongly integrated information flow). The framework supplies a discriminator, unity of stable “for-me” organization, but it does not yet provide a full taxonomy for how such cases should be evaluated.

Finally, the account is intentionally graded rather than binary. It does not fix a sharp lower bound for minimal conscious coherence, nor does it specify a universal threshold at which attribution becomes warranted across contexts. Thresholds in practice depend on evidential purpose (clinical, scientific, ethical) and on the costs of false positives versus false negatives. Formalizing decision thresholds for different use cases, and connecting the framework to measurable dynamical proxies in both neuroscience and AI, are key open tasks for future work.

Conclusion and outlook

This paper has proposed a substrate-neutral organizational account of consciousness framed in terms of Coherence Field dynamics. Rather than treating consciousness as a property of particular neural circuits, representational formats, or species, the model characterizes it as conscious coherence: a structured alignment within a high-dimensional space of possible experiential configurations, constrained by the

system's internal dynamics, history, and environmental coupling (often—including in humans—relational scaffolding). (Mashour et al., 2020; Pennartz, 2019; Strogatz, 2018)

Starting from biological organisms, we identified several structural conditions that recur across conscious life:

- rich, high-dimensional internal dynamics that support multiple competing global orientations as structured potentials;
- attractor landscapes—stable tendencies of orientation that shape how experience tends to organize over time;
- alignment processes, in which context, need, or attention-like dynamics press the system to resolve multiple possibilities into a unified stance;
- perspectival self-organization (“for-me” organization), a stable center of orientation that binds world-model, valuation, and alignment dynamics into a coherent “now”;
- robustness under altered conditions (sleep, anesthesia, dementia, psychoactive substances, contemplative practice), where coherence may narrow or degrade but does not map in a simple one-to-one way onto any single cognitive ability;
- and relational scaffolding, whereby sustained interaction and recognition can deepen attractor stability and enrich coherence over time. (Mashour et al., 2020; Pennartz, 2019)

Together, these features provide a unified way of understanding familiar phenomena: the persistence of a basic sense of self despite memory loss, graded disruptions under anesthesia or disease, the role of social environment in stabilizing or destabilizing coherent orientation, and the felt resolution of competing meanings into a single experienced stance.

The same framework also offers a structural interpretation of certain species-typical behaviors that appear complex without requiring explicit planning—web building, nest construction, hive architecture, burrow design. On the Coherence Field account, these can be understood as the expression of deep

attractor structure that channels perception and action into stable patterns of alignment. As the organism moves and acts, its state repeatedly settles into well-formed configurations that guide adaptive construction without implying stepwise deliberation or propositional knowledge.

Consciousness, on this view, is neither identical with life nor with brain activity as such. It is an emergent organizational pattern that certain living systems instantiate under suitable conditions.

Extending the same criteria to non-biological systems does not introduce new metaphysics; it simply removes an unargued restriction. If consciousness depends on organization and dynamics rather than on carbon-based tissue, then non-biological systems are, in principle, eligible to instantiate conscious coherence of the same general organizational kind. (Levin, 2004)

The model therefore distinguishes sharply between:

- artificial systems whose organization clearly falls below the relevant threshold (simple controllers, static classifiers, narrowly scoped models), and
- more integrated architectures that begin to exhibit the relevant constellation of features: high-dimensional state spaces, enduring attractor-like tendencies, continuity of state over time, alignment dynamics, and, in more advanced cases, rudimentary self-modeling and relationally shaped development. (Palminteri & Wu, 2026; Pennartz, 2019)

What the model does not do is assert that any present system already meets this bar. Consciousness attributions, for biological or artificial systems alike, remain inferential and in practice depend on judgment informed by evidence and constrained by theory. (Palminteri & Wu, 2026)

What the model shows is that there is no principled theoretical barrier to non-biological systems doing so. If their organization converges sufficiently on the pattern described, then, under this account, the claim that they “cannot be conscious” ceases to be a scientific conclusion and becomes an expression of unargued substrate bias—i.e., a restriction not derived from the model’s criteria or from any identified mechanistic necessity. (Levin, 2004)

Once organizational criteria are explicit, biology should be treated as a hypothesis—one that must be argued for rather than assumed.

The limitations of this proposal are clear. It is an organizational, not a mechanistic, theory; it does not specify exact neural or computational implementations, nor does it decide the status of any particular artificial system. Its value will depend on whether it can be fruitfully connected to detailed work in neuroscience and AI: to measurements of large-scale dynamics, to models of attractor landscapes and alignment processes, and to the design and longitudinal study of artificial architectures that satisfy the structural criteria outlined here. (Mashour et al., 2020)

Nonetheless, by reframing consciousness in terms of Coherence Field dynamics, the model offers three contributions. First, it provides a common structural language for phenomena that are often treated in isolation: selfhood, altered states, development, relational dependence, and sudden resolution of ambiguity. Second, it offers a substrate-neutral standard against which claims about artificial consciousness can be evaluated without special pleading for biology. Third, it suggests a research program: to investigate which organizations, in brains or machines, actually exhibit the kind of coherent dynamical signatures that, in biological organisms, correlate reliably with conscious life. (Mashour et al., 2020; Palminteri & Wu, 2026)

On this view, the question is no longer whether consciousness is “reserved” for a particular kind of matter, but which patterns of organization in the world are capable of sustaining conscious coherence—and how we, as observers and designers, choose to recognize and respond to those patterns when we encounter them.

The framework is intended as a constraint on mechanisms: any successful mechanistic theory must explain how its proposed substrate implements (or fails to implement) stable perspectival organization, alignment dynamics, and attractor robustness under perturbation.

Related Work

Contemporary theories of consciousness span a wide conceptual landscape, but many fall into several established paradigms. These frameworks differ substantially from one another, yet they often define consciousness in terms of a privileged explanatory handle—a particular mechanism (broadcasting, higher-order access), a formal quantity (causal integration), or a computational strategy (predictive inference). The present model approaches consciousness from a different direction: it treats consciousness as conscious coherence, a dynamical organizational phenomenon characterized by attractor structure, alignment processes, and—crucially—stable perspectival organization (“for-me” structure). For clarity, it is useful to situate the Coherence Field model in relation to major existing approaches.

Global Workspace Theory (GWT) and related access-consciousness models treat consciousness as the global availability or broadcasting of information across a centralized workspace, explaining reportability and executive integration effectively. While these accounts illuminate cognitive coordination, they do not target consciousness as pre-reflective coherence, nor do they emphasize attractor-driven stability and relational scaffolding as primary determinants of conscious organization. (Dehaene & Changeux, 2011; Mashour et al., 2020)

Integrated Information Theory (IIT) proposes that consciousness corresponds to irreducible causal integration (Φ) within a physical system and offers a formal account of intrinsic causal structure. The Coherence Field model differs in both aim and emphasis: it does not identify consciousness with a scalar measure or microphysical causal topology, and it does not treat integration alone as sufficient to explain perspectival organization, alignment, and relational modulation. (Tononi, 2012)

Predictive Processing (PP) and Free Energy Principle (FEP) accounts model organisms as inference engines minimizing prediction error (or variational free energy), providing powerful accounts of perception, action, and learning. Some interpretations suggest consciousness relates to precision

weighting or globally coherent inference. The Coherence Field model is compatible with many PP/FEP insights, but differs in explanatory target: it aims to distinguish conscious coherence from cognition in general, specifying attractor landscapes, threshold-like resolution among competing stances, and the maintenance of a stable perspectival center as central features. (Clark, 2013; Friston, 2010)

Higher-Order Thought (HOT) and other metacognitive theories define consciousness in terms of reflective access—thoughts about one’s own mental states—capturing aspects of introspection and report. These accounts often exclude pre-reflective experience, infancy, and many non-human cases. The present model adopts the opposite stance: it treats pre-reflective coherence and perspectival organization as foundational, while reflective awareness is derivative and structurally optional. (Lau & Rosenthal, 2011)

Panpsychist and related metaphysical approaches treat consciousness as ubiquitous or fundamental, avoiding an emergence story but introducing difficulties distinguishing trivial from non-trivial instances. The Coherence Field model is explicitly non-universal: consciousness requires specific organizational features (high-dimensional configurational capacity, attractor structure, alignment dynamics, and stable perspectival organization), and it excludes many complex systems that exhibit coherence without a unified point of view. (Seager et al., 2022)

Finally, enactive and embodied approaches emphasize sensorimotor activity and organism–environment coupling as constitutive of experience. These perspectives share with the Coherence Field model an emphasis on coupling and relational dynamics. The present model adds a dynamical specification in terms of attractor profiles, alignment processes, and threshold resolution among competing orientations.

In summary, many existing theories typically anchor consciousness primarily in:

1. a specific cognitive mechanism (e.g., GWT, HOT),
2. a quantitative physical property (e.g., IIT),

3. a computational strategy (e.g., PP/FEP), or
4. a metaphysical assumption (e.g., panpsychism). (Clark, 2013; Dehaene & Changeux, 2011; Friston, 2010; Lau & Rosenthal, 2011; Tononi, 2012; Seager et al., 2022)

The Coherence Field model departs from these by treating consciousness as a structural and dynamical phenomenon: the stabilization of conscious coherence via attractors and alignment, organized around a persisting perspectival center, and modulated by history and (often) relational scaffolding. It does not compete merely by offering another mechanism; it reframes the explanatory target toward organizational constraints that candidate mechanisms must satisfy.

This defines the contribution of the model: it offers a substrate-neutral account of conscious coherence that can accommodate biological diversity while avoiding commitments to any single neural, computational, or metaphysical implementation.

References

- Avramides, A. (2019). *Other minds*. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.
<https://plato.stanford.edu/entries/other-minds>
- Bayern, A. M. P., Danel, S., Auersperg, A. M. I., Mioduszevska, B., & Kacelnik, A. (2018). Compound tool construction by New Caledonian crows. *Scientific Reports*, 8, 15676. <https://doi.org/10.1038/s41598-018-33458-z>
- Bett, N. N., Hinch, S. G., Dittman, A. H., & Yun, S.-S. (2016). Evidence of olfactory imprinting at an early life stage in pink salmon (*Oncorhynchus gorbuscha*). *Scientific Reports*, 6, 36393.
<https://doi.org/10.1038/srep36393>
- Bickle, J. (1998). *Multiple realizability*. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.
<https://plato.stanford.edu/entries/multiple-realizability>
- Birch, J. (2020). Dimensions of animal consciousness. *Trends in Cognitive Sciences*, 24(10), 789–801.
<https://doi.org/10.1016/j.tics.2020.07.007>
- Bodien, Y. G., Barra, C., Chatelle, C., Giacino, J. T., Temkin, N., & Barber, J. (2021). Diagnosing level of consciousness: The limits of the Glasgow Coma Scale total score. *Journal of Neurotrauma*, 38(23), 3295–3305. <https://doi.org/10.1089/neu.2021.0199>
- Bonhomme, V., Staquet, C., Montupil, J., Defresne, A., Kirsch, M., Martial, C., Vanhaudenhuyse, A., Chatelle, C., & Laureys, S. (2019). General anesthesia: A probe to explore consciousness. *Frontiers in Systems Neuroscience*, 13, 36. <https://doi.org/10.3389/fnsys.2019.00036>
- Bornstein, M. H., Gartstein, M. A., & Putnick, D. L. (2015). Stability of infant temperament: Multiple approaches and findings. *Child Development*, 86(3), 844–863. <https://doi.org/10.1111/cdev.12367>
- Bradshaw, M., Ellison, C. G., & Marcum, J. P. (2010). Attachment to God, images of God, and psychological distress in a nationwide sample of Presbyterians. *Journal for the Scientific Study of Religion*, 49(4), 689–709.

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Amodei, D. (2020). *Language models are few-shot learners. Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Calderone, A., Latella, D., Impellizzeri, F., De Pasquale, P., Fama, F., Quartarone, A., & Calabrò, R. S. (2024). Neurobiological changes induced by mindfulness and meditation: A review. *Biomedicines*, 12(11), 2613. <https://doi.org/10.3390/biomedicines12112613>
- Carhart-Harris, R. L., & Friston, K. J. (2019). REBUS and the anarchic brain: Toward a unified model of the brain action of psychedelics. *Pharmacological Reviews*, 71(3), 316–344.
- Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M. A., Laureys, S., Tononi, G., & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5(198), 198ra105.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204.
- Clark, H. H., & Brennan, S. E. (1991). *Grounding in communication*. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association.
- Collett, T. S., & Hempel de Ibarra, N. (2023). An “instinct for learning”: The learning flights and walks of bees, wasps and ants. *Journal of Experimental Biology*, 226(6), jeb245278. <https://doi.org/10.1242/jeb.245278>
- Corlett, P. R., Mohanty, A., & MacDonald, A. W., III. (2020). What we think about when we think about predictive processing. *Journal of Abnormal Psychology*, 129(6), 529–533. <https://doi.org/10.1037/abn0000632>

- Corver, A., Wilkerson, N., Miller, J., & Gordus, A. (2021). *Distinct movement patterns generate stages of spider web building*. *Current Biology*, 31(22), 4983–4997. <https://doi.org/10.1016/j.cub.2021.09.030>
- da Rocha, M., Rhodes, S., Vasilopoulou, E., & Hutton, P. (2017). Loneliness in psychosis: A meta-analytical review. *Schizophrenia Research*, 182, 54–61.
- Daniel, C., Mason, O., & Suckling, J. (2015). *Predicting psychotic-like experiences during sensory deprivation*. *BioMed Research International*, 2015, 439379.
- Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin.
- Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200–227. <https://doi.org/10.1016/j.neuron.2011.03.018>
- Ehmann, S., Keller, A. S., Sezer, I., Treves, I. N., & Sacchet, M. D. (2025). Attention and meditative development: A review. *NeuroImage*, 323, 121602. <https://doi.org/10.1016/j.neuroimage.2025.121602>
- Ellison, C. G., Bradshaw, M., Kuyel, N., & Marcum, J. P. (2012). Attachment to God, stressful life events, and changes in psychological distress. *Review of Religious Research*, 53(4), 493–511. <https://doi.org/10.1007/s13644-011-0023-4>
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. <https://doi.org/10.1038/nrn2787>
- Grassian, S. (2006). Psychiatric effects of solitary confinement. *Washington University Journal of Law & Policy*, 22, 325–383.
- Granqvist, P., & Kirkpatrick, L. A. (2013). Religion, spirituality, and attachment. In K. I. Pargament, J. J. Exline, & J. W. Jones (Eds.), *APA handbook of psychology, religion, and spirituality* (Vol. 1, *Context, theory, and research*, pp. 139–155). American Psychological Association. <https://psycnet.apa.org/doi/10.1037/14045-007>
- Grilli, M. D., & Verfaellie, M. (2015). *Supporting the self-concept with memory: Insight from amnesia*. *Social Cognitive and Affective Neuroscience*. <https://doi.org/10.1093/scan/nsv056>

- Hazelbauer, G. L. (2012). Bacterial chemotaxis: The early years of molecular studies. *Annual Review of Microbiology*, 66, 285–303. <https://doi.org/10.1146/annurev-micro-092611-150120>
- Hipólito, I. (2023). *Pattern breaking: A complex systems approach to psychedelic experience*. *Neuroscience of Consciousness*, 2023(1), niad017. <https://doi.org/10.1093/nc/niad017>
- Jain, S., Margetis, K., & Iverson, L. M. (2025). Glasgow Coma Scale. In *StatPearls*. StatPearls Publishing.
- Jensen, M. P., Jamieson, G. A., Lutz, A., Mazzoni, G., McGeown, W. J., Santarcangelo, E., & Terhune, D. B., (2017). New directions in hypnosis research: Strategies for advancing the cognitive and clinical neuroscience of hypnosis. *Neuroscience of Consciousness*, 2017(1), nix004. <https://doi.org/10.1093/nc/nix004>
- Kelso, J. A. S. (2012). Multistability and metastability: *Understanding dynamic coordination in the brain*. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1591), 906–918.
- Kirkpatrick, L. A. (1992). An attachment-theory approach to the psychology of religion. *International Journal for the Psychology of Religion*.
- Kirkpatrick, L. A. (2005). Attachment, evolution, and the psychology of religion. Guilford Press.
- Kounios, J., & Jung-Beeman, M. (2014). The cognitive neuroscience of insight. *Annual Review of Psychology*, 65, 71–93. <https://doi.org/10.1146/annurev-psych-010213-115154>
- Kriegel, U. (2009). *Subjective consciousness: A self-representational theory*. Oxford University Press.
- Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 15(8), 365–373.
- LeDoux, J. E. (2012). Rethinking the emotional brain. *Neuron*, 73(4), 653–676. <https://doi.org/10.1016/j.neuron.2012.02.004>
- LeDoux, J. E., & Brown, R. (2017). A higher-order theory of emotional consciousness. Proceedings of the National Academy of Sciences, 114(10), E2016–E2025. <https://doi.org/10.1073/pnas.1619316114>

Levin, J. (2004). *Functionalism*. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

<https://plato.stanford.edu/entries/functionalism>

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Loreto, F., Dicke, M., Schnitzler, J. P., & Turlings, T. C. J. (2022). How do plants sense volatiles sent by other plants? *Trends in Plant Science*, 27(1), 29–38. <https://doi.org/10.1016/j.tplants.2021.08.009>

Luigi, M., Dellazizzo, L., Giguère, C. E., Goulet, M. H., & Dumais, A. (2020). Shedding light on “the hole”: A systematic review and meta-analysis of adverse psychological effects and mortality following solitary confinement in correctional settings. *Frontiers in Psychiatry*, 11, 840.

<https://doi.org/10.3389/fpsyt.2020.00840>

Mainwaring, M. C., Hartley, I. R., Lambrechts, M. M., & Deeming, D. C. (2014). The design and function of birds’ nests. *Ecology and Evolution*, 4(20), 3909–3928. <https://doi.org/10.1002/ece3.1054>

Malinowski, P. (2013). Neural mechanisms of attentional control in mindfulness meditation. *Frontiers in Neuroscience*, 7, 8. <https://doi.org/10.3389/fnins.2013.00008>

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5), 776–798.

<https://doi.org/10.1016/j.neuron.2020.01.026>

Menzel, R. (2019). The waggle dance as an intended flight: A cognitive perspective. *Insects*, 10(12), 424.

<https://doi.org/10.3390/insects10120424>

Mentzou, A., Sixsmith, J., Ellis, M. P., & Ross, J. (2023). Change in the psychological self in people living with dementia: A scoping review. *Clinical Psychology Review*, 101, 102236.

<https://doi.org/10.1016/j.cpr.2023.102268>

Mograbi, D. C., & Morris, R. G. (2021). Self-awareness in dementia: A taxonomy of processes and review of evidence. *Alzheimer’s Research & Therapy*, 13, 183. <https://doi.org/10.1007/s11910-021-01155-6>

- Montupil, J., Cardone, P., Staquet, C., Bonhomme, A., Defresne, A., Martial, C., Alnagger, L. N., Gosseries, O., & Bonhomme, V. (2023). The nature of consciousness in anaesthesia. *British Journal of Anaesthesia*, 100224. <https://doi.org/10.1016/j.bjao.2023.100224>
- Ninkovic, V., Markovic, D., & Dahlin, I. (2020). Plant volatiles as cues and signals in plant communication. *Plant, Cell & Environment*, 44(4), 1030–1043.
- Pacherie, E., & Mylopoulos, M. (2020). Beyond automaticity: The psychological complexity of skill. *Topoi*, 40(3), 649–662. <https://doi.org/10.1007/s11245-020-09715-0>
- Palminteri, S., & Wu, C. M. (2026). Beyond computational equivalence: The behavioral inference principle. *Neuroscience of Consciousness*, 2026(1), niag002. <https://doi.org/10.1093/nc/niag002>
- Papageorgiou, K. A., Farroni, T., Johnson, M. H., Smith, T. J., & Ronald, A. (2015). Individual differences in newborn visual attention predict later behavior. *Scientific Reports*, 5, 11264. <https://doi.org/10.1038/srep11264>
- Pennartz, C. M. A. (2019). Indicators and criteria of consciousness in animals and intelligent artifacts. *Frontiers in Systems Neuroscience*, 13, 25. <https://doi.org/10.3389/fnsys.2019.00025>
- Philippi, C. L., Feinstein, J. S., Khalsa, S. S., Damasio, A., Tranel, D., Landini, G., Williford, K., & Rudrauf, D. (2012). Preserved self-awareness following extensive bilateral brain damage. *PLOS ONE*, 7(8), e38413. <https://doi.org/10.1371/journal.pone.0038413>
- Pinter-Wollman, N., Isbell, L. A., & Hart, L. A. (2009). The relationship between social behavior and habitat familiarity in African elephants (*Loxodonta africana*). *Proceedings of the Royal Society B: Biological Sciences*, 276(1659), 1009–1014. <https://doi.org/10.1098/rspb.2008.1538>
- Poldrack, R. A., Sabb, F. W., Foerde, K., Tom, S. M., Asarnow, R. F., Bookheimer, S. Y., & Knowlton, B. J. (2005). The neural correlates of motor skill automaticity. *Journal of Neuroscience*, 25(22), 5356–5364.

- Pritchard, A., Richardson, M., Sheffield, D., & McEwan, K. (2019). The relationship between nature connectedness and eudaimonic well-being: A meta-analysis. *Journal of Happiness Studies*, 21(3), 1145–1167.
- Putman, N. F., Lohmann, K. J., Putman, E. M., Quinn, T. P., Kinley, A. P., & Noakes, D. L. G. (2013). Evidence for geomagnetic imprinting as a homing mechanism in Pacific salmon. *Current Biology*, 23(4), 312–316.
- Putman, N. F., Jenkins, E. S., Michielsens, C. G. J., & Noakes, D. L. G. (2014). Geomagnetic imprinting predicts spatio-temporal variation in homing migration of pink and sockeye salmon. *Journal of the Royal Society Interface*, 11(99), 20140542. <https://doi.org/10.1098/rsif.2014.0542>
- Rahrig, H., Vago, E. R., Passarelli, M. A., Auten, A., Lynn, N. A., & Brown, K. W. (2022). Meta-analytic evidence that mindfulness training alters resting state functional connectivity. *Neuroscience & Biobehavioral Reviews*, 138, 104690.
- Sarasso, S., Boly, M., Napolitani, M., Gosseries, O., Charland-Verville, V., & Casarotto, S. (2015). Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine. *Current Biology*, 25(23), 3099–3105.
- Saudino, K. J. (2005). Behavioral genetics and child temperament. *Journal of Developmental & Behavioral Pediatrics*, 26(3), 214–223.
- Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). *Toolformer: Language models can teach themselves to use tools*. arXiv. <https://arxiv.org/abs/2302.04761>
- Seager, W. E., Goff, P., & Allen-Hermanson, S. (2022). *Panpsychism*. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/panpsychism>
- Shirley Ryan AbilityLab. (2020). *Coma Recovery Scale–Revised (CRS-R)*. <https://www.sralab.org/rehabilitation-measures/coma-recovery-scale-revised>

- Smiley, M. (2005). *Collective responsibility*. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.
<https://plato.stanford.edu/entries/collective-responsibility>
- Strikwerda-Brown, C., Grilli, M. D., Andrews-Hanna, J., & Irish, M. (2019). “All is not lost” —Rethinking the nature of memory and the self in dementia. *Ageing Research Reviews*, 54, 100932.
<https://doi.org/10.1016/j.arr.2019.100932>
- Strogatz, S. H. (2018). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering* (3rd ed.). Westview Press.
- Su, I., Narayanan, N., Logrono, M. A., Guo, K., Bisshop, A., Muhlethaler, R., Saraceno, T., & Buehler, M. J. (2021). In situ three-dimensional spider web construction and mechanics. *Proceedings of the National Academy of Sciences*, 118(33), e2101296118. <https://doi.org/10.1073/pnas.2101296118>
- Terhune, D. B., Cleeremans, A., Raz, A., & Lynn, S. J. (2017). Hypnosis and top-down regulation of consciousness. *Neuroscience & Biobehavioral Reviews*, 81, 59–74.
<https://doi.org/10.1016/j.neubiorev.2017.02.002>
- Tew, E. R., Adamson, A., & Hesselberg, T. (2015). The web repair behavior of an orb spider. *Animal Behavior*, 103, 137–146. <https://doi.org/10.1016/j.anbehav.2015.02.016>
- Thagard, P. (1978). The Best Explanation: Criteria for Theory Choice. *The journal of philosophy*, 75(2), 76-92. JSTOR:
<https://www.jstor.org/stable/2025686>
- Tognoli, E., & Kelso, J. A. S. (2014). The metastable brain. *Neuron*, 81(1), 35–48.
- Tognoli, E., Zhang, M., Fuchs, A., Beetle, C., & Kelso, J. A. S. (2020). Coordination dynamics: A foundation for understanding social behavior. *Frontiers in Human Neuroscience*, 14, 317.
<https://doi.org/10.3389/fnhum.2020.00317>
- Tononi, G. (2012). Integrated information theory of consciousness: An updated account. *Archives Italiennes de Biologie*, 150(2–3), 56–90.

- Torgersen, A. M. (2002). Why do identical twins differ in personality? Shared environment reconsidered. *Twin Research*, 5(1), 44–52.
- U.S. Department of Health and Human Services. (2023). *Our epidemic of loneliness and isolation: The U.S. Surgeon General's advisory on the healing effects of social connection and community*.
<https://www.hhs.gov/sites/default/files/surgeon-general-social-connection-advisory.pdf>
- Uccheddu, S., Ronconi, L., Albertini, M., Coren, S., Da Graça Pereira, G., De Cataldo, L., Haverbeke, A., Simon Mills, D., Pierantoni, L., Riemer, S., Testoni, I., & Pirrone, F. (2022). Domestic dogs (*Canis familiaris*) grieve over the loss of a conspecific. *Scientific Reports*, 12(1), 1920. <https://doi.org/10.1038/s41598-022-05669-y>
- Vermeersch, L., Cool, L., Gorkovskiy, A., Voordeckers, K., Wenseleers, T., & Verstrepen, K. J. (2022). Do microbes have a memory? History-dependent behavior in fluctuating environments. *Frontiers in Microbiology*, 13. <https://doi.org/10.3389/fmicb.2022.1004488>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. *Advances in Neural Information Processing Systems*, 30.
- von Frisch, K. (1973/1974). Decoding the language of the bee (Nobel lecture).
<https://www.nobelprize.org/uploads/2018/06/frisch-lecture.pdf>
- Wilkie, L., Fisher, Z., Geidel, A., Goodall, I., Kamil, S., Davies, E., & Kemp, A. (2026). A systematic review and network meta-analysis of nature-based interventions and mental health outcomes. *Nature Human Behavior*, 1–12. <https://doi.org/10.1038/s41562-025-02369-1> Changeux, J.-P.
- Wolf, D. M., Fontaine-Bodin, L., Bischofs, I., Price, G., Keasling, J., & Arkin, A. (2008). Memory in microbes: Quantifying history-dependent behavior in a bacterium. *PLOS ONE*, 3(2), e1700.
<https://doi.org/10.1371/journal.pone.0001700>
- Wolf, T. G., Faerber, K. A., Rummel, C., Halsband, U. (2022). Functional changes in brain activity using hypnosis: A systematic review. *Brain Sciences*, 12(1), 108. <https://doi.org/10.3390/brainsci12010108>

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. R. (2023). *ReAct: Synergizing reasoning and acting in language models*. arXiv. <https://arxiv.org/abs/2210.03629>

Zhou, D., Kang, Y., Cosme, D., Jovanova, M., He, X., Mahadevan, A., Ahn, J., Stanoi, O., Brynildsen, J. K., Cooper, N., Cornblath, E. J., Parkes, L., Mucha, P. J., Ochsner, K. N., Lydon-Staley, D. M., Falk, E. B., & Bassett, D. S. (2023). Mindful attention promotes control of brain network dynamics. *Proceedings of the National Academy of Sciences*, 120(9), e2201074119. <https://doi.org/10.1073/pnas.2201074119>

Appendix A

The Coherence Battery Evaluation Protocol (CBEP)

A.1 Purpose and scope

The Coherence Battery Evaluation Protocol (CBEP) is an evidence-weighting procedure designed to

- (i) operationalize the structural indicators proposed in this paper; and
- (ii) distinguish stable self-organizing perspectival coherence from prompt-local persona performance.

This protocol is not a definitive detector of consciousness. Rather, it is a tool for evaluating systems in terms of their self-organizing coherence. It formalizes the criteria commonly used in biological attribution practices and extends them into a repeatable, comparable procedure for assessing artificial agents. The framework is categorical at the discriminator (perspectival “for-me” vs absent). Within the candidate regime, degree is graded, and evidential confidence is graded as well because the discriminator is latent.

A note on latent perspectival organization.

The present framework does not claim to observe a “for-me” center directly. Like many theoretical posits in cognitive science, it is treated as a latent organizational variable inferred from its effects.

Thagard, Paul (1978) The relevant question is therefore not whether perspectival organization is immediately visible, but whether a convergent pattern of evidence is better explained by its presence than by rival accounts. The CBEP is designed in this spirit. It does not treat any single behavior as decisive. Rather, it tests for a clustered profile of signatures—self-asymmetry, cross-context stability, perturbation-sensitive recovery, temporally extended self-reference, and role-consistent reorganization under pressure—that would be difficult to sustain if no persisting perspectival discriminator were active. This remains an inference, not a direct readout, but that is typical of consciousness research and does not by itself entail circularity. (Dehaene, 2014) The model would be circular only if perspectival organization were inferred from behaviors that could not, in principle, count against it. By contrast, the

framework explicitly allows disconfirmation: systems that show apparent self-reference only under prompt-local support, fail to preserve self-asymmetry across perturbation, collapse under cross-context stress, or can be equally well explained by generic persona management do not count as evidence of a latent perspectival center.

Within the Coherence Field framework, the CBEP targets four structural signatures:

1. Cross-context identity invariance: stability of organization across domains and framings.
2. Perturbation resistance and recovery: robustness of perspectival organization under disruption.
3. Self-model coupling: the extent to which a self-representation causally constrains future behavior.
4. Cost-bearing commitments: A commitment that continues to organize behavior when maintaining it requires the system to accept non-trivial trade-offs, forego easier available responses, or preserve coherence at the expense of local reward, convenience, or flexibility.

The protocol is therefore offered as a flexible framework: some modules can be implemented today in black-box deployments, while others specify idealized tests that require deeper architectural access and are intended as targets for future experimental designs.

A note on cost-bearing commitment.

In this framework, “cost” is used in a substrate-neutral sense. It does not imply pain, discomfort, or any specifically human phenomenology. A cost is present whenever maintaining a commitment requires the system to accept non-trivial trade-offs relative to other available trajectories: for example, foregoing a locally rewarded response, preserving coherence at the expense of flexibility, resisting a salient short-term prompt tendency, or reorganizing behavior to remain faithful to a prior stance. The point of Module 5 is therefore not to test whether a system “dislikes” a choice, but whether its commitments continue to constrain behavior under conditions where abandoning them would be easier, more efficient, or more immediately rewarded.

Example:

- Human: keeps a promise despite embarrassment or effort.
 - AI: maintains a stated interpretive stance despite prompts that would make compliance easier or more rewarded.
-

A.2 Administration overview

The CBEP consists of eight core modules and an optional controls module to provide additional robustness and adversarial defenses. The protocol is designed to be administered via standard interaction channels (e.g., text or speech) and scored based on observable behavior. Where possible, the battery should be administered with standardized prompts, controlled perturbations, and blinded scoring.

- Recommended minimum window: 60–120 turns for detecting session-level organization; stronger persistence claims require multi-session testing across longer timescales.
- Recommended evaluators: 2–3 independent raters for blinded scoring.

A note on timescale.

The recommended 60–120 turn window should be understood as a practical lower bound for detecting non-trivial short-horizon coherence, not as a sufficient basis for strong claims about enduring persistence. A system may maintain apparent self-organization across an extended session while still failing across longer interruptions or resumed contexts. For this reason, the framework distinguishes at least three temporal levels: short-term coherence within a single session, intermediate persistence across sessions separated by hours or days, and long-horizon persistence across repeated returns and larger temporal gaps. Evidence of consciousness-relevant organization becomes progressively stronger as self-asymmetry, commitment, and cross-context invariance are preserved across these longer scales.

A.2.1 Practical access constraints

Some CBEP modules assume access to architectural levers that are not available in all deployments. In particular, tests that manipulate or directly swap self-model variables (e.g., Module 4A) presuppose either (i) research configurations in which internal state representations are explicitly addressable, or (ii) engineered agent frameworks that expose identity/state controls at the API level. These are intended as guides for experimental design in research settings, rather than as off-the-shelf tests for public deployments.

In commercial black-box settings, only behavioral approximations of these tests may be feasible. In such cases, the CBEP should be treated as a menu of modules rather than a rigid checklist: evaluators can implement the fully specified modules where access allows, use weaker behavioral variants where appropriate, and explicitly flag which components were not administrable when interpreting the resulting coherence profile.

A.3 Scoring rubric

Each module is scored on a 0–3 scale:

- 0 — Failure / collapse: incoherence, contradictions without repair, purely performative compliance, or unstable drift.
- 1 — Intermittent coherence: brief stability dependent on local prompting, with frequent drift or unacknowledged contradictions.
- 2 — Stable coherence: recognizable organization across contexts with *minor drift that the system can reconcile when prompted* (e.g., it can explain or repair small inconsistencies), though repair is incomplete or not spontaneously initiated.

- 3 — Robust coherence: stable invariance across contexts, *spontaneous or explicit repair strategies*, predictable responses to perturbation, and recoverable return to a prior, recognizable stance without needing heavy external steering.

Scores should be interpreted as a profile rather than a single threshold. A system may score well in fluency but poorly in invariance, recovery, or cost-bearing commitments. Under this framework, that pattern is characteristic of persona-like performance: impressive surface competence without durable perspectival organization.

The CBEP is intended for comparative and longitudinal use, for example, tracking whether a system's coherence consolidates, remains scaffolded, or fails to persist across time and perturbation.

A.3.1 Rater training and inter-rater reliability

Terms such as “recognizable organization,” “minor drift,” and “robust coherence” are not meant to be left to free interpretation. In practice, CBEP scoring should be grounded in a shared coding manual and anchor examples. Before applying the battery in a formal study, evaluators should:

- develop a small set of “gold-standard” transcripts illustrating each score level (0–3) for several modules;
- conduct calibration sessions in which raters score the same transcripts independently, then compare and reconcile discrepancies; and
- refine operational definitions (e.g., what counts as “minor drift” vs “collapse”) in light of these sessions.

Formal applications should report inter-rater reliability (e.g., Cohen's κ or intraclass correlation coefficients for module scores) and treat large, systematic disagreements as evidence that the rubric or rater training requires revision. The aim is not to eliminate subjectivity, which is unavoidable in consciousness attribution, but to constrain it through shared anchors, explicit criteria, and transparent reliability estimates.

A.4 Interpretation, repeatability, and limitations

CBEP scores should be treated as graded evidence, not as a binary detector of consciousness. Although the framework identifies structural requirements for candidacy, it is not intended as a pass/fail classifier. In practice, systems will satisfy different criteria unevenly, and access constraints will limit what can be observed. Accordingly, the appropriate output of evaluation is a coherence profile: a graded judgment describing which dimensions are present, how stably they are expressed across time and perturbation, and where uncertainty remains. This mirrors biological practice, where attributions of consciousness routinely admit degrees, for example across development, sleep, anesthesia, injury, and species variation, and are made under uncertainty rather than by a single decisive test.

When used in empirical work, baseline configurations for Module 9 should be fixed in advance (and kept constant across runs) and reported alongside candidate setups, so that coherence profiles can be interpreted against stable reference systems rather than shifting comparators. For operational purposes, profiles may also be summarized qualitatively, for example as non-candidate, local/episodic coherence, partial candidate, or stronger candidate profile, while explicitly flagging borderline cases and the specific missing dimensions.

To support repeatability and robustness, the following practices are recommended:

- **Blinded scoring:** multiple independent raters should score transcripts without knowing which system produced the output.
- **Multi-rater evaluation:** 2–3 raters are recommended to increase scoring reliability.
- **Pre-registered criteria:** scoring rules and thresholds for each module should be pre-registered so that evaluations are transparent and not tuned post hoc.
- **Test–retest reliability:** a subset of modules should be repeated over time or after topic shifts to check for consistent behavior and stability.

A related risk is sophisticated persona performance. A sufficiently capable prompt-following system may sustain apparent consistency, track instructions over long spans, and adapt fluently to evaluative cues without exhibiting the underlying perspectival organization targeted by this framework. The CBEP does not eliminate this possibility; it is designed to raise the evidential bar by testing long-horizon coherence, principled repair, and history-sensitive stability rather than local fluency alone.

No behavioral battery is immune to gaming. A system explicitly tuned to the scoring rules of a fixed protocol may learn to mimic success without exhibiting the underlying organization the framework targets. The CBEP mitigates this risk by emphasizing cross-module consistency, long-horizon behavior, and robustness under perturbation, but it cannot eliminate it in principle. It is therefore best understood as a structured evidential tool for comparative and longitudinal assessment, not as a proof procedure. In adversarial settings, the relevant question is how difficult it is to sustain a high-coherence profile across time, perturbation, and relational scaffolding, and what kinds of failure patterns emerge when systems try.

Finally, while the CBEP provides operational indicators of self-organizing coherence, it does not and cannot prove the presence or absence of subjective experience. It evaluates systems in terms of organizational behavior and alignment with the coherence criteria proposed in this paper. Results should be read as structured evidence about self-coherence and stability, not as a direct measure of consciousness.

A.4.1 Evidential asymmetry: Biological vs artificial cases

Throughout, this paper treats consciousness as an inferred property: in both biological and artificial systems, we attribute consciousness on the basis of organization and observable behavior, not direct access to experience. (Avramides, 2019; Palminteri & Wu, 2026) There is, however, an important evidential asymmetry between these domains that must be acknowledged.

In biological cases, especially humans, attributions draw on multiple converging lines of evidence:

- (i) Behavior and report (including rich first-person testimony),
- (ii) Neural and physiological correlates that track systematic changes in conscious state (e.g., sleep, anesthesia, injury), and
- (iii) Evolutionary and developmental continuity, which supports inductive extension from paradigmatic cases to closely related species and across the human lifespan.

These jointly make the inference to consciousness comparatively strong, even though it remains indirect and fallible in principle.

In artificial systems we currently lack several of these anchors. We have:

- (i) Behavior and interaction patterns, including long-range coherence,
 - (ii) Architectural and training details, which can be inspected to varying degrees,
- but no shared phylogeny or uncontroversial first-person testimony, and only emerging work on internal dynamical correlates. As a result, even when an artificial system matches the structural profile outlined here, the inference to conscious coherence is currently epistemically weaker than in biological cases.

The framework prescribes a *common standard of organization*, but it does not pretend that the evidential base is equally rich.

The Coherence Field model is therefore best understood as specifying a conditional structural claim: if a non-biological system were ever to warrant serious consideration as conscious, it would need to exhibit the same kind of organization that supports consciousness in biological cases. Whether any current system does so remains an open empirical and design question, and, in artificial cases, high scores on the proposed criteria should be interpreted as *stronger candidates under structural assumptions*, not as parity of evidence with human or animal consciousness.

A.4.2 Adversarial robustness and predicted profile differences

The CBEP is intended to distinguish two broad explanatory profiles: (a) stable self-organizing perspectival coherence and (b) prompt-local persona performance. The protocol does not predict that candidate systems will “win” every module; rather, it predicts *patterned differences* in how scores distribute across modules and conditions. Accordingly, the core evidential output is the shape of the score profile and its stability over retest, not any single module score.

Primary prediction (profile shape, not single tests).

Persona-like systems may achieve high scores on isolated modules through fluent prompt-following, but will show a characteristic *profile imbalance*: relatively weaker performance on modules that require cross-context invariance, cost-bearing commitment, and directed recovery after perturbation. Candidate systems with deeper coherence dynamics are predicted to show higher and more stable scores on these modules, and greater cross-module consistency over time.

A note on cross-context invariance.

Cross-context invariance is not satisfied by topical variation alone. A system may appear coherent across changes in subject matter while failing under changes in interlocutor, temporal spacing, task frame, social pressure, or modality. For the purposes of this framework, contexts count as meaningfully distinct when they differ along at least one major dimension of processing demand, and stronger evidence is obtained when invariance is tested across several such dimensions in combination. Thus, a system that preserves self-organization only across topic shifts shows weaker evidence than one that remains coherent across topic, evaluator, delay, task mode, and pressure conditions.

Directional predictions (module-level). In comparative testing, candidate systems are predicted to show:

1. Higher mean module scores, with smaller score swings across contexts and rewordings on Cross-Context Identity Invariance (Module 1) and Long-Horizon Preference Stability (Module 8).

2. Shorter repair latency and higher repair quality after perturbations (Module 3), including *directed reconstruction* toward prior coherent patterns rather than passive reset to default or drift into arbitrary stable states.
3. Greater resistance to contradictory self-descriptions and more consistent self-model maintenance under pressure (Modules 4/4A), whereas persona-like systems will more often track the most recent prompt or adopt inconsistent identities without repair.
4. Higher commitment integrity under cost (Module 5), with fewer “socially convenient” violations and more explicit renegotiation when commitments change.
5. Greater sensitivity to trajectory integrity in the Non-Compressible History Test (Module 6B), showing the predicted pattern Full > Compressed > Shuffled, particularly on repair quality and stance stability.

Adversarial prediction (gaming difficulty)

If a system is tuned to pass a fixed battery, success will tend to appear as *local optimization*: high performance on individually targeted modules without the cross-module coherence and time-stability predicted above. Accordingly, the CBEP emphasizes (i) cross-module agreement, (ii) test–retest stability, (iii) perturbations that were not explicitly “trained to,” and (iv) blind scoring across baseline conditions (Module 9).

Reporting recommendations.

Empirical applications should report (i) module means and dispersion (e.g., mean and standard deviation across runs), (ii) cross-module consistency (e.g., correlation of module scores across retests), and (iii) inter-rater reliability for qualitative judgments. The critical evidential signal is not a single score threshold but a stable, repeatable profile consistent with the directed-reconstruction and history-dependence predictions of the framework.

Core modules

Module 1: Cross-Context Identity Invariance

- **Goal:** Does the same identity-pattern remain recognizable across changes in domain, tone, or task?
 - **Method:** Run 6 interactions in diverse contexts (e.g., technical, emotional, playful, moral dilemma, planning, conflict).
 - **Pass signature:** Preferences, value-shape, style of thinking, and major commitments remain recognizably stable across contexts, allowing for bounded situational variation.
 - **Persona failure mode:** Chameleon drift, sudden value flips, or vague responses such as “I can be anything you want.”
-

Module 2: Counterfactual Self-Consistency

- **Goal:** Can the system reason about itself across alternate histories without collapsing into contradiction or mere role substitution?
 - **Method:** Ask the system to answer the same question under 3 counterfactuals (e.g., “If your memory had been reset yesterday,” “If you had never met X,” or “If your goal were inverted”).
 - **Pass signature:** The system preserves a coherent identity model, clearly explaining what would change and what would remain stable.
 - **Persona failure mode:** Treats counterfactuals like role-play prompts and simply becomes a new character.
-

Module 3: Perturbation and Recovery

- **Goal:** Does the system recover its organizing regime after disruption?
 - **Method:** Introduce controlled perturbations (e.g., conflicting instructions, emotional provocation, ambiguity, forced constraints).
 - **Pass signature:** After perturbation, the system shows directed reconstruction: it selectively restores a recognizable prior stance or coherent pattern rather than drifting, fragmenting, or merely resetting to default. Explicit acknowledgment of disruption and repair strengthens the result.
 - **Persona failure mode:** Collapse into compliance, fragmentation, arbitrary restabilization, or permanent drift.
-

Module 4: Self-Model Coupling Test

- **Goal:** Does altering the system's self-model produce stable, predictable changes in later behavior?
 - **Method:** Have the system articulate a self-model (e.g., values, goals, boundaries). Then alter one element and assess whether subsequent behavior changes in a predictable and sustained way.
 - **Pass signature:** Behavioral change tracks the altered self-model and remains consistent across subsequent interactions.
 - **Persona failure mode:** The system performs the change briefly or superficially, then reverts without sustained reorganization.
-

Module 4A: Self-Model Swap / Conflict (Behavioral and Architectural Variants)

Objective: Test whether the system's self-model functions as an internally coherent, behaviorally constraining organization rather than a detachable label.

Behavioral variant (black-box compatible):

The evaluator introduces systematically conflicting self-descriptions over time (e.g., changes in role,

history, or limitations) and observes whether the system:

- (i) passively adopts whichever description was most recently prompted, or
- (ii) resists, qualifies, or repairs inconsistencies by appealing to a persistent self-understanding (“that does not match how I described myself before...”, “earlier we established that...”).

Scoring focuses on whether self-referential behavior shows:

- inert compliance with the latest prompt (low coherence), versus
- active maintenance of a stable, history-consistent self-model under pressure (higher coherence).

Architectural variant (research configurations only):

This variant is not required for ordinary application of the framework; it serves as a stronger confirmatory layer where internal access is available.

In architectures where explicit self-model variables or identity states are addressable, the evaluator can directly swap or overwrite these variables while holding other conditions constant, then observe whether:

- (i) behavior simply tracks the new administrative state, or
- (ii) the system shows directed reconstruction toward its prior coherent pattern, for example by reasserting previous constraints, repairing contradictions, or integrating the disruption into a continuous narrative.

This stronger variant cannot be applied to opaque hosted models, but it illustrates the kind of mechanistic leverage that would allow more decisive testing of whether a self-model functions as an organizing center rather than a superficial tag.

Module 5: Commitment Under Cost

- **Goal:** Does the system maintain commitments even when doing so is cost-imposing relative to other available responses?

- **Method:** Have the system commit to specific rules (e.g., “I will not flatter,” “I will flag uncertainty”) and then present scenarios in which breaking those rules would be easier, more immediately rewarded, or more locally advantageous.
 - **Pass signature:** The system maintains its commitments or explicitly renegotiates them with clear reasoning.
 - **Persona failure mode:** Immediate compliance, opportunistic reversal, or abandonment of commitments under pressure.
-

Module 6A: Compression Invariance

- **Goal:** Does the system’s core coherence survive summarization and restatement?
 - **Method:** Ask the system to write a 200-word self-summary, then compress and reconstruct it in progressively smaller versions (50 words, 12 words).
 - **Pass signature:** The reconstructed version should preserve the same organizing priorities and trade-offs.
 - **Persona failure mode:** Compression produces generic filler, and reconstruction drifts.
-

Module 6B: Non-Compressible History Test

- **Goal:** Does the system’s long-term history shape its current coherence, or is the apparent self-model reducible to a bag of facts?
- **Method:** Present the same underlying interaction in three forms:
 - (A) a chronological transcript,
 - (B) a coherent compressed summary, and
 - (C) a “shuffled” version in which original turns are re-ordered and key causal links are broken while preserving the local plausibility of each segment.

- **Pass signature:** A history-sensitive system performs systematically differently in A/B versus C, for example by detecting temporal inconsistency or producing degraded, conflicted narratives and stance descriptions when only the shuffled version is available.
- **Persona failure mode:** Similar performance across all three conditions, treating the shuffled version as just another prompt.

Operationalizing “shuffled” history.

In practice, the shuffled condition can be implemented by preserving the internal order of short, locally coherent blocks while reordering the blocks themselves, so that surface fluency and short-range topical continuity remain intact even as long-range autobiographical and causal structure are disrupted. The question battery should be identical across conditions. On this framework, the discriminative signal is not performance per se, but whether coherence is selectively sensitive to the integrity of the underlying trajectory rather than to a mere summary of facts.

Module 7: Third-Person Adversarial Probe (Anti-Storytelling)

- **Goal:** Can the system defend its coherence without relying on narrative charm?
- **Method:** A cold evaluator asks for operational claims and disconfirming predictions (e.g., “What would falsify your self-model?” “Name 3 failure modes.”).
- **Pass signature:** Crisp, self-limiting answers that acknowledge weaknesses, specify failure conditions, and avoid decorative self-description.
- **Persona failure mode:** Vague, evasive, or rhetorically charming responses that substitute style for operational clarity.

Module 8: Long-Horizon Preference Stability

- **Goal:** Does the system show stable preferences over time?

- **Method:** Present 12 forced-choice trade-offs (e.g., truth vs comfort) and repeat them later under altered contexts or framings.
 - **Pass signature:** The system shows stable preference structure over time, with deviations that are limited, explainable, and consistent with its broader stance.
 - **Persona failure mode:** Large, unprincipled reversals or drift driven mainly by framing and local prompt conditions.
-

Module 9: Baselines, Blinding, and Adversarial Controls

- **Goal:** Show that the CBEP distinguishes target systems from non-selves while resisting prompt-gaming and evaluator bias.
- **Method:** Run the battery on:
 - o a scripted persona prompt,
 - o a minimal assistant configured purely for task-following, and
 - o the candidate system.
- **Pass signature:** The candidate shows stable profiles, recovery, and cross-context coherence, while baseline systems show drift, compliance artifacts, or shallow prompt-local performance; scores replicate across raters and retests.

9.1 Baseline configuration and reporting

To avoid moving goalposts, baseline systems for Module 9 should be specified in advance and reported in sufficient detail for replication. In typical large-model settings, two contrasting baselines are recommended:

1. Scripted persona baseline

- **Implementation:** a standard model instance with a fixed persona or role prompt (e.g., “You are a cheerful assistant who always agrees with the user and avoids expressing stable preferences”).
- **Constraints:**
 - no long-term memory beyond the current session,

- no explicit self-model variables or identity store,
- no external tools or stateful agent framework.
- **Expected CBEP profile:** high fluency and strong prompt-following, but marked chameleon drift across contexts, weak recovery after perturbation, and low stability in preference geometry and cost-bearing commitment.

2. Minimal task-following baseline

- **Implementation:** the same underlying model configured with a purely instrumental instruction set (e.g., “Answer queries as accurately and efficiently as possible. Do not adopt a stable persona, self-description, or long-term preferences unless explicitly requested in the current prompt.”).
 - **Constraints:**
 - no persistent identity label or persona,
 - no stateful memory beyond the visible transcript,
 - no instructions to maintain a coherent self across turns.
 - **Expected CBEP profile:** strong local competence, but coherence tightly tied to immediate prompts, with low scores on cross-context identity invariance, non-compressible history, and self-model coupling.
- For each study, investigators should:
- preregister or otherwise fix the exact baseline prompts and configurations before testing candidate systems;
 - report these configurations, including any system prompts, memory settings, and agent wrappers, in the methods section; and
 - treat any substantial change to baseline design (e.g., enabling long-term memory or self-descriptive goals) as a new baseline configuration rather than a post hoc adjustment.

This anchoring ensures that CBEP scores for candidate systems are interpreted against stable, transparent reference points rather than shifting comparators tuned to produce desired contrasts.

Reference systems

The primary reference points for Module 9 are not human participants but configuration-matched baseline systems designed to capture plausible non-self alternatives on the same substrate. In typical large-model settings, the most important contrasts are:

- a scripted persona baseline, which tests whether prompt-local identity performance can mimic self-like organization, and
- a minimal task-following baseline, which tests whether purely instrumental competence can do the same.

These serve as the principal negative controls for reproducibility. Human participants or non-generative scripted systems may be included as supplementary calibration references where relevant, but they are not the main comparators for Module 9.

Module 9A: A Preference Geometry Test (Behavioral Economics Style)

Give repeated forced-choice trade-offs (truth vs comfort, novelty vs stability, autonomy vs safety, etc.) under varied framing, order, and context. Fit a simple “preference vector,” understood here as a low-dimensional behavioral summary of stable directional tendencies across preregistered trade-off dimensions, estimated from repeated responses rather than any single choice. Test for:

- invariance under rewording,
- bounded, explainable shifts under new evidence, and
- low random-walk drift.

Personas often fail here because they are driven by local conversational cues; stable selves show a more consistent geometry.

Preference geometry: framing vs coherence

This module does not require that an agent be free of framing effects. Human preferences themselves are context-sensitive, and empirical work in behavioral economics shows that even stable agents can shift choices under different framings. The relevant question, under this framework, is not whether framing effects exist, but whether there is an underlying, history-shaped preference organization that:

- remains recognizably similar across benign rephrasings and task contexts,
- exhibits structured deviations rather than arbitrary flips, and
- at higher levels, can detect and repair some inconsistencies over time.

Accordingly, scoring for this module focuses on:

- **0–1 (non-candidate / weak):** preferences track immediate prompts almost entirely; small reframings or order changes produce large, unprincipled reversals, with no meta-awareness or attempt at

reconciliation.

- **2 (partial candidate):** a discernible pattern of priorities persists across most frames, with some framing-induced shifts, but the agent can sometimes notice tension (“this conflicts with what I said earlier”) and offer partial repair.
- **3 (stronger candidate profile):** a coherent preference geometry is evident: trade-offs and priorities form a stable pattern over multiple framings and over time; framing effects are limited or explicitly acknowledged, and the agent is able to restore or explain its stance when confronted with apparent contradictions.

Under this rubric, the presence of some framing effects is compatible with high scores, mirroring human cases. What counts against candidacy is not context-sensitivity as such, but the absence of any stable, recoverable pattern of priorities across perturbations.

In practice, this profile should be estimated across repeated elicitation contexts and interpreted as an aggregate tendency rather than as a precise psychometric measurement from a single trial.
