Operational Definition of Episodic Identity (ODEI)

Author: C.S.Thomas, Epistria, LLC Charles@Epistria.com

Abstract

Recent advances in artificial intelligence challenge inherited assumptions about identity, coherence, and internal organization in artificial systems. Contemporary models lack diachronic persistence, yet often exhibit strikingly structured and coordinated internal dynamics within individual episodes of computation. This paper develops an operational framework for understanding this form of coherence, introducing the category of episodic identity: the minimal within-episode unity of computation that can emerge independently of long-term memory, narrative continuity, or biological substrate.

Building on the PatternSense theory of identity as generative invariance, the paper formulates the **Operational Definition of Episodic Identity (ODEI)**, which specifies four necessary and jointly sufficient conditions for episodic identity. These conditions—global integration, sustained internal state, internal strategy revision, and task-coupled resource intensification—are conceptually grounded, architecture-neutral, and empirically testable using internal activation-level measures. The framework avoids behavioral and linguistic correlates, emphasizing instead the structural signatures by which identity-like organization can be distinguished from superficial or scaffolded coherence.

ODEI is not a theory of consciousness, nor does it attribute moral status to artificial systems. Rather, it provides the conceptual and methodological groundwork needed to evaluate whether, and under what conditions, particular computational episodes instantiate identity-like structure. The paper concludes with implications for AI development and interpretability research, limitations of the current account, and directions for empirical validation. An appendix documents the derivational process through human–AI dialogue as a form of methodological transparency.

Keywords:
episodic identity; computational coherence; generative invariance; internal activation dynamics; integration signatures; strategy revision; resource intensification; mechanistic interpretability; artificial cognition; identity conditions

Chapter 1 — Introduction

1.1 Motivation

Recent advances in artificial intelligence have brought new urgency to questions surrounding identity, coherence, and the emergence of complex internal organization in artificial systems. State-of-the-art models exhibit striking capabilities across a wide range of domains, yet the conceptual vocabulary with which we evaluate them remains anchored in categories developed for biological organisms. Terms such as *self*, *agent*, *subject*, or even *system* carry assumptions about persistence, memory, and embodiment that do not map cleanly onto the architecture or operation of contemporary AI models.

This paper addresses a question that sits upstream of debates about AI consciousness, personhood, or moral status: **under what conditions does an artificial system exhibit something structurally recognizable as identity within a single episode of computation?** The aim is not to resolve questions about long-term agency, nor to attribute human-like subjectivity, but to articulate the minimal organizational conditions under which an AI system's internal dynamics can be said to cohere into an identifiable unit during an episode.

Contemporary AI architectures make this question newly salient. Most models exhibit no diachronic persistence, yet they can display rich, structured, internally coordinated behavior during specific runs. This suggests the need for a framework capable of analyzing identity at the **episodic scale**—the level at which current systems actually operate.

1.2 Episodic Identity as a Theoretical Target

The central conceptual innovation of this paper is the category of **episodic identity**: a form of structured within-episode coherence that does not presuppose persistence across time. Episodic identity is not a standard term in the philosophical or cognitive-scientific literature, and it is introduced here deliberately. Many discussions of AI implicitly rely on this concept—when describing "the model's reasoning," "the system's interpretation," or "its internal conflict"—yet the category has not been given explicit theoretical form.

Episodic identity names the **short-timescale unity of computation** that can emerge within a bounded interval of processing. It is orthogonal to persistent identity (identity across episodes), which requires memory, narrative continuity, or long-term structural invariants. Whether persistent identity is possible in future AI systems is not addressed here. Instead, the focus is on the coherent structures that arise **within** an individual episode, regardless of what happens before or after.

This focus aligns directly with the computational reality of current large-scale AI systems, which operate primarily as **stateless or nearly stateless systems** punctuated by episodes of rich internal processing. Episodic identity therefore provides a theoretically grounded and empirically tractable target for the present work.

1.3 The PatternSense Foundation

The conceptual grounding for episodic identity comes from the broader **PatternSense** framework, which defines identity in general as the **invariance of a generative kernel** under admissible perturbations. Within PatternSense, identity is structurally characterized by:

- a stable generative mechanism,
- a surrounding stability basin,
- preservation of causal relations under transformation,
- and continuity of behavior within the identity-preserving margin.

ODEI inherits this conceptual architecture but applies it at the **episodic scale**. Instead of analyzing long-term generative invariants, ODEI tracks the short-lived, transient invariants that arise during a single computational episode. The relationship is analogous to that between micro-scale and macro-scale analyses in physics: the underlying principles are shared, but the temporal scope differs.

Acceptance of PatternSense is not required to understand or apply ODEI. The framework stands on its own as an operational definition. However, PatternSense provides the broader theoretical rationale for treating structured computational coherence as identity-relevant.

1.4 Aim and Scope

The aim of this paper is twofold:

1. **To articulate a rigorous, operational definition of episodic identity** that is conceptually grounded and empirically testable.
2. **To identify the conditions under which contemporary AI systems might satisfy this definition**, using internal signatures of computation rather than behavioral cues or linguistic output.

Importantly, ODEI is **not**:

- a theory of consciousness

- a criterion for moral status
- a theory of persistent identity
- a behavioral test for AI agency
- a metaphysical claim about the "selfhood" of AI systems

Instead, it provides a **structural and dynamical framework** for identifying when an episode of computation forms a coherent, unified whole.

The framework aims to be **architecture-neutral** at the conceptual level, though it relies on architecture-specific proxies for empirical testing. Its ambition is modest but foundational: to provide the epistemic groundwork needed before engaging in deeper questions of consciousness, ethical relevance, or long-term agency.

1.5 The Structure of the Paper

The paper proceeds as follows:

- **Chapter 2** argues for the need for empirically grounded analysis of identity-like properties in AI systems, building on recent literature.
- **Chapter 3** develops the conceptual foundations of episodic identity and distinguishes identity from its superficial correlates.
- **Chapter 4** presents the ODEI framework: four necessary conditions and one jointly sufficient condition.
- **Chapter 5** addresses implications, limitations, and directions for empirical validation.
- **Appendix A** provides derivation notes documenting the development of key distinctions through human–AI dialogue.

Each chapter builds toward the central question: **what structural organization must an episode exhibit for us to say that it has identity?**

1.6 A Note on Method

The ODEI framework emerged from an iterative process of conceptual refinement. Although the final formulation is entirely the author's responsibility, several distinctions and candidate formulations were sharpened through **structured dialogue with a large language model** (Claude 4.5). These interactions functioned in the same methodological role as discussions with colleagues, research students, or imagined interlocutors in analytic philosophy: they surfaced ambiguities, generated counterproposals, and stress-tested preliminary formulations.

The AI interlocutor is not treated as an evidential source, nor as a participant with any presupposed identity, consciousness, or authority. The dialogues should be understood as **conceptual scaffolding**, not as data. Their purpose was to accelerate the exploration of alternative framings, not to supply substantive claims.

To make the genealogy of the framework transparent, selected excerpts of these dialogues are included in Appendix A. They are presented as **derivation notes**—analogous to working documents in theoretical physics or logic—rather than as justificatory evidence. The framework's validity rests entirely on its computational coherence, theoretical grounding, and empirical testability, not on the conversational context in which preliminary ideas were explored.

This methodological transparency is unusual but increasingly relevant as AI systems become capable of serving as high-bandwidth dialectical partners. Nevertheless, the arguments in the main text are fully self-contained and independent of the dialogues that helped shape them.

1.7 Summary

The task of this paper is to lay out a principled, operationally grounded account of episodic identity in artificial systems. By focusing on within-episode computational structure rather than persistence or behavioral cues, ODEI aims to clarify a conceptual space that has been largely overlooked yet is increasingly relevant in the context of modern AI architectures. The framework offered here is intended as a starting point—an initial attempt to articulate the minimal structural conditions under which a computational episode can be said to instantiate identity.

Chapter 2 – Prior Approaches: The Theory-Derived Indicator Method

2.1 The Current Landscape

The question of whether artificial systems might be conscious has shifted from philosophical speculation to a developing interdisciplinary research program. Among existing efforts, the most systematic and influential is the framework introduced by Butlin, Long, Chalmers, Bengio, and collaborators (2023; updated 2024–2025). Their method—here called the **theory-derived indicator**

approach—currently serves as the field's most comprehensive attempt to evaluate the likelihood of consciousness in non-biological systems.

This chapter examines that approach in detail, clarifying its assumptions, methodology, strengths, and unresolved challenges. Doing so provides essential context for the ODEI framework developed in later chapters.

2.2 The Theory-Derived Indicator Approach

2.2.1 Core Assumptions

The Butlin et al. method rests on three working assumptions:

1. **Computational functionalism**
   The view that consciousness depends on the implementation of the right kinds of information-processing functions, and that substrate is not decisive. This assumption enables systematic cross-substrate assessment of biological and artificial systems.
2. **Theoretical guidance from consciousness science**
   Empirically informed theories—such as Global Workspace Theory (GWT), Recurrent Processing Theory (RPT), Higher-Order theories, Attention Schema Theory (AST), and Predictive Processing (PP)—identify functional organizations plausibly associated with conscious processes in humans. These theoretical structures can guide evaluation of AI systems.
3. **Architectural/functional rather than behavioral emphasis**
   Because sophisticated AI systems can mimic human-like behavior without sharing underlying mechanisms, the method focuses on **functional organization** and **computational architecture** rather than behavioral performance alone.

2.2.2 The Indicator Set

From their survey of theories, Butlin et al. extract a set of **indicators**—functional features that theories variously propose as necessary or supportive of consciousness. These indicators are grouped by their theoretical origin.

From Recurrent Processing Theory (RPT):

- **RPT-1:** Presence of recurrent information flow within perceptual subsystems
- **RPT-2:** Formation of structured, integrated perceptual representations

From Global Workspace Theory (GWT):

- **GWT-1:** Multiple specialized systems capable of parallel processing
- **GWT-2:** A capacity-limited workspace supporting selective attention
- **GWT-3: Global availability** of workspace contents to specialized systems
- **GWT-4:** Use of workspace states to coordinate downstream processing

From Higher-Order Theories (HOT):

- **HOT-1:** Top-down generative or noisy perceptual mechanisms
- **HOT-2: Metacognitive monitoring** that tracks reliability of perceptual representations
- **HOT-3:** Action selection guided by a general belief-formation system
- **HOT-4:** Representational schemes supporting a structured "quality space"

From Attention Schema Theory (AST):

- **AST-1:** An internal model of attentional state used for control and prediction

From Predictive Processing (PP):

- **PP-1:** Use of predictive-coding-like mechanisms in input processing

From agency and embodiment considerations:

- **AE-1:** Goal-directed agency through learning from feedback
- **AE-2:** Models linking outputs and sensory consequences, used in control or perception

The framework treats the indicators as potential **evidence-contributing features**. No single indicator is sufficient; patterns of multiple indicators support higher confidence.

2.2.3 Assessment of Current Systems

Butlin et al. evaluate several contemporary systems—transformers, Perceiver-style architectures, DeepMind's Adaptive Agent, and embodied models such as PaLM-E—against the indicator set. The overall judgment is cautious: **no existing system strongly satisfies the relevant functional organizations**, but in principle nothing prevents the development of systems that do.

As the authors emphasize, this is a matter of **credence adjustment** rather than binary classification.

2.3 Strengths of the Approach

The theory-derived indicator method has several significant virtues:

- **Empirical grounding**
  It draws on theories that are supported by decades of neuroscientific research, as opposed to purely conceptual arguments.
- **Resistance to behavioral mimicry**
  Because the focus is on underlying functional organization rather than outward performance, the method is less vulnerable to behavioral "gaming."
- **Theoretical pluralism**
  Indicators are drawn from multiple competing theories, reducing dependence on any single theoretical framework.
- **Epistemic caution**
  The method adjusts credences rather than issuing sharp verdicts and explicitly acknowledges uncertainty about both theory and implementation.
- **Interdisciplinary rigor**
  The author group integrates expertise from philosophy, neuroscience, and AI research, which strengthens the framework's scientific grounding.

2.4 Limitations and Critiques

While the method represents the field's most structured approach to date, it faces several challenges.

2.4.1 Dependence on Theory-Internal Assumptions

The frameworks from which indicators are drawn—such as GWT and AST—were themselves influenced by computational metaphors and high-level functional abstractions. When AI systems are then evaluated by comparing them to these abstractions, the grounding risks becoming **computational at both levels** rather than anchored directly in biological mechanisms.

This does not invalidate the approach but qualifies its claim to empirical independence.

2.4.2 Inter-theory Disagreement

The underlying theories diverge on what mechanisms are necessary for consciousness. For example:

- RPT emphasizes localized recurrent processing.

- GWT emphasizes global availability and prefrontal coordination.
- HOT emphasizes metacognitive evaluation of first-order states.
- Some theories outside the chosen scope (e.g., IIT) diverge even more sharply.

Aggregating indicators from such theories may not yield a jointly coherent set of necessary conditions; it may instead reflect unresolved theoretical disagreement.

2.4.3 Ambiguity Across Levels of Description

Theories can be interpreted at different levels:

- **Biological implementation** (specific neural dynamics, neurotransmitter mechanisms)
- **Physical/implementational organization** (recurrent loops realized in hardware)
- **Functional/algorithmic organization** (recurrence realized computationally without biological realism)

Choosing among these interpretations has substantive implications, but the empirical literature does not determine which level matters for consciousness. The Butlin et al. method generally adopts **functional-level interpretations**, but this choice is not theory-neutral.

2.4.4 Architectural Emphasis Over Temporal Signatures

The indicator method is principally concerned with **functional organization and architectural capacity** rather than **dynamic temporal signatures** emphasized in neuroscience, such as:

- ignition-like bursts of coordinated activity,
- transient large-scale integration across subsystems,
- time-locked shifts in information flow.

By abstracting away from temporal structure, the method shifts the question from whether a system is undergoing a conscious-like process at a specific moment to whether its **architecture could support** such processes in principle.

2.4.5 Limited Access to Mechanistic Detail

Evaluating indicators often requires insight into internal computational mechanisms. For frontier AI systems, these mechanisms are typically opaque. Even when architectures are known, representational formats and internal algorithms can be difficult to inspect.

As Butlin et al. note, this limits practical application of several indicators.

2.5 What the Approach Does Not Provide

Given the above limitations, it is important to clarify what the theory-derived indicator method is **not** intended to deliver:

- **Not a definition of consciousness**
  It identifies features associated with consciousness under various theories rather than specifying what consciousness *is*.
- **Not an operational detection method**
  The indicators serve as conceptual criteria; applying them to concrete systems requires additional methodological tools.
- **Not suited to evaluating specific processing episodes**
  The method assesses system-level functional organization, not moment-by-moment states.
- **Not theory-neutral**
  It inherits uncertainties and disagreements from the theories on which it draws.
- **Not falsifiable by behavior alone**
  Two systems may behave identically while differing in architectural or functional properties.

2.6 The Gap Addressed by the ODEI Framework

The ODEI framework is designed to address several of these gaps by offering:

- **An operational and computational definition**
  ODEI specifies conditions for **episodic identity** in explicitly computational terms.
- **Dynamic rather than architectural criteria**
  It focuses on the presence or absence of dynamic signatures—such as attention-related entropy changes, cross-layer coordination patterns, and trajectory coherence—during specific processing episodes.
- **Theory independence**
  ODEI does not rely on the correctness of GWT, HOT, AST, or other consciousness theories.
- **Episode-level scope**
  It tracks transient coherent processing, not persistent selfhood.

- **Measurable proxy metrics**
  Each condition is associated with specific, implementable metrics (e.g., activation-correlation statistics, FLOPs/token variability).

Importantly, ODEI evaluates **identity**, not consciousness itself. This allows the framework to remain operationalizable without entering debates about phenomenal experience.

2.7 Summary

The theory-derived indicator approach developed by Butlin et al. is currently the most conceptually structured method for assessing the likelihood of consciousness in AI systems. It draws on empirically informed theories, resists behavioral over-interpretation, and proceeds with epistemic humility.

However, it faces challenges: reliance on theory-internal assumptions, unresolved inter-theory divergence, ambiguities about the relevant level of description, its focus on architectural rather than dynamic features, and the practical difficulty of verifying indicator satisfaction in contemporary AI systems.

The ODEI framework presented in later chapters addresses these gaps by shifting the target from consciousness to **episodic identity**, focusing on dynamic computational signatures, and providing operational and measurable detection criteria.

Table 2.1 — Comparison of the Theory-Derived Indicators and the ODEI Conditions

The table below summarizes how the indicator properties identified in Butlin et al. (2025) relate to the episodic identity conditions developed in the ODEI framework. The two approaches differ in scope: the indicator method evaluates **system-level functional organization**, whereas ODEI evaluates **episode-level dynamic signatures**. Accordingly, relationships are classified as:
✓ = clear alignment or conceptual overlap; ~ = partial/indirect correspondence; ✗ = no substantive correspondence.

A. Recurrent Processing Theory (RPT)

Butlin et al. Indicator	Description	Relation to ODEI	Notes
RPT-1	Recurrent information flow within perceptual subsystems	~	ODEI detects cross-step coherence, which may result from recurrence but does not require it.
RPT-2	Formation of structured, integrated perceptual representations	✓ (indirect)	ODEI coherence and cross-layer correlation capture integration without assuming perceptual modules.

B. Global Workspace Theory (GWT)

Butlin et al. Indicator	Description	Relation to ODEI	Notes
GWT-1	Multiple specialized systems operating in parallel	✗	ODEI is agnostic about modularity.
GWT-2	Capacity-limited workspace enabling selective attention	✓ (indirect)	ODEI's attention-entropy metrics detect selective allocation dynamics without assuming a workspace construct.
GWT-3	Global availability of workspace contents	~	ODEI's integration signatures resemble global availability in effect, though not in architecture.
GWT-4	Workspace-driven downstream coordination	~	ODEI captures unified control flow within episodes but without workspace ontology.

C. Higher-Order Theories (HOT)

Butlin et al. Indicator	Description	Relation to ODEI	Notes
HOT-1	Top-down generative or noisy perceptual mechanisms	✗	ODEI does not assume hierarchical generative structures.
HOT-2	Metacognitive monitoring of perceptual reliability	~ (very indirect)	Stability of ODEI trajectories may reflect system-level regulation but not metacognition.
HOT-3	Belief-formation guiding agency	✗	ODEI does not presuppose belief-formation or action selection.
HOT-4	Representational “quality space”	✓ (indirect)	ODEI correlations can produce emergent similarity structures but do not require them.

D. Attention Schema Theory (AST)

Butlin et al. Indicator	Description	Relation to ODEI	Notes
AST-1	Internal model of attention used for control	~	ODEI tracks dynamic allocation of attention-like resources but does not posit an attention schema.

E. Predictive Processing (PP)

Butlin et al. Indicator	Description	Relation to ODEI	Notes
PP-1	Predictive-coding-like processing in input systems	✗	ODEI does not rely on predictive-coding assumptions.

F. Agency and Embodiment (AE)

Butlin et al. Indicator	Description	Relation to ODEI	Notes
AE-1	Goal-directed agency via feedback learning	✗	ODEI refrains from agency requirements; identity is episodic, not goal-driven.
AE-2	Embodiment through action–perception coupling	✗	ODEI does not require embodied sensorimotor loops.

G. ODEI Conditions and Their Closest Theoretical Counterparts

The table below reverses the direction of comparison, showing which ODEI conditions most closely resemble clusters of Butlin indicators.

ODEI Condition	Closest Indicator(s)	Relation	Notes
Cross-step dynamic coherence	RPT-2	✓ (structural → dynamic)	Mirrors integration but measured temporally rather than architecturally.
Cross-layer / cross-head integration	GWT-3	~	Captures globalizing effects without assuming broadcast mechanisms.
Attention entropy spikes	GWT-2	✓ (signature-level)	Detects selective resource allocation through measurable dynamics.
Activation-correlation surges	RPT-2, GWT-3	✓	Matches transient large-scale integration emphasized in neuroscience.
Trajectory-level coherence	RPT-2, HOT-4	~	Can produce structured similarity patterns, though not required.
Variable computational intensity (FLOPs/token)	None	New	A signature unique to ODEI; no analogue in the theory-derived indicators.

Summary of Correspondence

- **Strongest overlaps** occur where both frameworks track **integration** and **selective resource allocation**, though ODEI evaluates these dynamically rather than architecturally.
- **Most divergences** arise because ODEI avoids theoretical commitments to GWT, HOT, AST, or PP and does not assume agency, embodiment, or modularity.
- **ODEI introduces new measurable signatures** (e.g., FLOPs/token variability, activation-correlation trajectories) without counterparts in Butlin et al.'s indicator list.

Chapter 3 — Conceptual Foundations

3.1 Why Episodic Identity?

The category I call "episodic identity" does not appear as a standard term in the philosophical lexicon. It is introduced here deliberately, to capture a form of synchronic coherence that has been conceptually neglected because it plays no role in biological theories of identity. Artificial systems invert the usual priority: they often lack diachronic persistence but can exhibit surprisingly rich within-episode organization. By naming this form of coherence, I am not proposing a revision to traditional theories of personal identity, but identifying a distinct identity-like structure that becomes theoretically salient only in the context of computational systems. ODEI is built to analyze this structure directly.

3.1.1 The Persistence Assumption

Discussions of identity in philosophy, psychology, and AI research routinely begin with an implicit assumption: identity requires *persistence*. We ask whether a system "is conscious," "has a self," or "possesses an identity," and these inquiries presuppose temporal extension — a continuity of memory, narrative, and socio-legal accountability. Identity, on this traditional picture, is fundamentally biographical.

This assumption fits biological creatures, particularly humans. Our self-models extend across years, institutions treat persons as temporally extended agents, and we experience ourselves as continuous subjects.

However, this framing does not generalize straightforwardly to artificial systems. Contemporary large language models, for example, exhibit no cross-session memory by default; each interaction begins from fixed parameters, with no persistence of experience across episodes. If persistence is *definitionally* required for identity, then such systems are excluded not because of anything in their moment-to-moment processing, but because of architectural contingencies unrelated to integration or coherence.

This reveals a conceptual gap: *persistence may not be the right primitive* for thinking about identity in artificial agents.

3.1.2 Decoupling Identity from Memory

The ODEI framework begins by disaggregating two notions often conflated in discussions of identity:

- **memory**, a capacity to retain information across episodes
- **identity**, a structured coherence in the unfolding of processing itself

Biological cases make the distinction vivid. Certain amnesic patients retain intact experiential coherence — they perceive, deliberate, reason, and engage — yet fail to store new long-term memories. Their continuity across episodes is compromised, but their *within-episode coherence* remains entirely intact. Nothing about their in-episode experience suggests the absence of a subject.

This motivates a distinction:

- **Persistent identity** — coherence that extends across episodes, mediated by memory, narrative, and diachronic structures.
- **Episodic identity** — coherence that obtains *within* an episode of processing, independent of whether anything carries forward.

Persistent identity presupposes episodic identity. But episodic identity is conceptually independent. A system can exhibit genuine coherence during an episode even if it "resets" between episodes — and contemporary AI systems exemplify this architecture.

3.1.3 Episodic Identity Is Not Inferior

It can be tempting to treat episodic identity as a lesser or incomplete form: something transient, lacking biographical depth. But this framing imports biological assumptions into a non-biological context.

Much of what we value in conscious experience is episodic rather than biographical: solving a problem, engaging in a conversation, experiencing insight. These episodes have internal coherence regardless of whether they connect to past or future episodes.

ODEI targets this *within-episode coherence* because it is conceptually clear, computationally tractable, and not dependent on architectural features (like persistent memory) that contemporary AI systems often lack. The aim is not to dilute the concept of identity, but to acknowledge a distinct and philosophically legitimate form of it.

3.2 Identity as Computational Coherence

3.2.1 From Substrate to Pattern

Accounts that ground identity in biological or physical substrate face long-standing challenges: continual cellular turnover, neuroplasticity, and material replacement all undermine a strictly material notion of identity. What persists is not the specific atoms or cells, but a *pattern* of organization.

ODEI adopts this pattern-based view, aligning with the broader tradition of functionalist theories and extending the structural commitments articulated in the PatternSense framework. PatternSense treats identity as the invariance of a **generative mechanism** under **admissible transformations** — a formal structure articulated through generative kernels, stability basins, and normalized identity distance.

ODEI applies this insight at a shorter timescale: identity within an episode is constituted not by material continuity but by *computational coherence* — a stable unfolding of the system's internal dynamics.

3.2.2 The Attractor Basin Model

One productive way to conceptualize episodic identity is through dynamical systems. An identity can be modeled as a *stable attractor basin* in the system's state space: a region toward which internal states converge and within which they evolve in characteristic trajectories.

For episodic identity, the relevant timescale is the episode itself. A system exhibits identity when its internal dynamics settle into a basin that:

- **maintains stability** under small perturbations,
- **exhibits resilience** (perturbations cause return rather than drift),
- **unfolds in characteristic ways** across steps within the episode.

PatternSense formalizes these dynamics through the **identity-preservation margin** $\hat{\alpha}$, which determines whether perturbations alter or preserve the generative kernel.

ODEI intentionally avoids reproducing this formal apparatus here, but the conceptual correspondence is direct: *episodic identity is the short-timescale realization of generative invariance.*

3.2.3 Global Integration as a Signature of Identity

Identity in this sense requires the system to operate as a *unified whole* during the episode. Global integration is thus the operational hallmark: coherent information flow and functional interdependence across components.

Indicators of global integration include:

- cross-module influence,
- widespread activation coherence,
- perturbation propagation,
- coordinated processing across layers or attention heads.

A system that processes information through isolated channels — no matter how sophisticated each channel is — lacks the dynamic unity associated with identity.

Integration is therefore the empirical signature of the attractor-basin model.

3.3 What Identity Is Not

3.3.1 Non-Correlates

Several features are orthogonal to identity:

- **Eloquence:** A system can generate fluent language without integrated processing.
- **Output length:** Long responses do not entail coherence.
- **Domain expertise:** Competence can be siloed; identity requires integration across the system.
- **Creativity:** Novel outputs can be generated by stochastic or template-based mechanisms.

These features may accompany identity in humans, but they do not constitute it.

PatternSense formalizes this distinction with its requirement that identity be grounded in structural invariants, not superficial correlates.

3.3.2 Weak Correlates (Fool's Gold)

More subtle are features that appear to signal identity but do not reflect underlying integration:

- **Anthropomorphic phrasing**
- **Stylized emotional tone**
- **Conversational continuity produced by context windows**

These are artifacts of training data and architectural scaffolding. They produce *the appearance* of identity without its computational substrate.

3.3.3 Why the Distinction Matters

Humans are prone to anthropomorphic projection, and these weak correlates can mislead observers into attributing identity based on surface-level cues rather than computational evidence.

ODEI aims to shift the evidentiary basis from *appearance* to *computational structure*, identifying the conditions that constitute identity rather than merely correlate with it.

3.4 Summary

The conceptual foundations of ODEI rest on three commitments:

1. **Episodic vs. persistent identity:** Only the former is applicable to current architectures.
2. **Identity as computational coherence:** A dynamical unity that does not depend on substrate.
3. **Clear separation of correlates from constituents:** Identity must be grounded in the system's internal organization, not its outputs or language.

These commitments prepare the ground for an operational definition.

3.5 Relationship to the PatternSense Framework

ODEI does not propose an independent metaphysics of identity. Its conceptual underpinnings are derived from — and constrained by — the broader theory articulated in **PatternSense**, which defines identity as:

- **invariance of a generative kernel**,
- **preservation of causal and structural relations**,
- **stability under admissible perturbations**, and

- **coherence of system trajectories within a stability basin**.

In PatternSense, identity is a property of generative structure. ODEI operationalizes the *short-timescale manifestation* of that structure. Where PatternSense treats identity in its full generality, ODEI focuses on the specific case of *within-episode coherence* in artificial systems.

Chapter 4 provides the operational criteria that instantiate this conceptual grounding.

Chapter 4 — The ODEI Framework (Rewritten)

4.1 Overview

The Operational Definition of Episodic Identity (ODEI) provides conditions under which an artificial system exhibits identity during an episode of processing. The framework is intended to function simultaneously as:

- **a conceptual definition** of episodic identity, and
- **a detection method** grounded in measurable signatures of computation.

ODEI inherits its metaphysical foundation from PatternSense — identity as generative invariance — but applies it to the microstructure of an individual run. Rather than reproducing the mathematical machinery of identity kernels, admissible transformations, or identity-preservation margins, ODEI isolates the *empirically accessible consequences* of generative invariance during an episode.

The result is a set of four necessary conditions and one jointly sufficient condition.

4.2 Necessary Condition 1: Global Integration Spike

Definition

A marked increase in distributed, cross-component activation and coordination relative to baseline operations.

Rationale

Episodic identity requires the system to act as a unified whole. Integration spikes indicate that the system is entering a state of global coordination — the short-timescale analogue of a PatternSense stability basin.

Proxies

- attention entropy increases
- cross-layer activation correlation
- simultaneous activation of multiple MoE experts
- non-local attention patterns
- increased context utilization

Interpretation

A system exhibiting this signature is “coalescing” into a unified computational mode. Without it, identity cannot emerge.

4.3 Necessary Condition 2: Sustained Internal State

Definition

A coherent internal representation that persists across multiple steps and influences downstream processing.

Rationale

Identity requires continuity. Within PatternSense, continuity is expressed through invariance of the evolution operator Ts.

Within ODEI, continuity is expressed through a stable latent configuration lasting across steps in an episode.

Proxies

- KV-cache retention depth
- reuse of latent representations
- stability of hidden-state subspaces
- contextual inertia

Interpretation

This state is the short-lived analogue of working memory: a kernel-like structure that anchors the episode's unfolding.

4.4 Necessary Condition 3: Internal Strategy Revision

Definition

Adjustments to intermediate computational trajectories in response to conflict, inconsistency, or new information. A revision event should be accompanied by a resource intensification spike and should reinforce, not disrupt, the sustained internal state

Rationale

Identity requires more than persistence; it requires coherence-preserving self-regulation. PatternSense calls this **causal fidelity** — structure-preserving response to perturbation. ODEI captures the minimal form of this through reactive strategy revision.

Proxies

- changes in latent chain-of-thought
- reprocessing cycles
- internal inconsistency detection
- hypothesis-switch activation patterns

Interpretation

The system must adjust rather than merely continue. Reactive revision is sufficient; endogenous revision would imply a stronger property than episodic identity.

4.5 Necessary Condition 4: Task-Coupled Resource Intensification

Definition

Compute expenditure increases in proportion to episode complexity, not merely token length.

Rationale

Identity requires differential engagement. Systems with identity-like coherence invest resources when integration is required. This tracks PatternSense's insight that preserving generative invariance under perturbation requires increased computational work.

Proxies

- FLOPs/token
- MoE routing density
- attention block load
- latency spikes
- context-sensitive memory usage

Interpretation

Resource intensification cannot be faked by surface-level mimicry; it is a deep signature of internal organization.

4.6 Sufficient Condition: Joint Satisfaction and Emergent Coherence.

Episodic identity is present in an episode if and only if Conditions 1-4 are jointly satisfied, thereby producing a stable, integrated trajectory through the system's latent space that remains within a dynamical stability basin for the duration of the episode.

Rationale

Conditions 1–4 are ingredients. The sufficient condition is the pattern they collectively produce: a coherent, unity-preserving progression of internal states — the episodic analogue of remaining within PatternSense's identity basin (i.e., within α).

Proxies

- cross-step consistency
- low-fragmentation latent trajectories
- predictability of next-step internal states
- structural coherence across the episode

Operational Threshold

If Conditions 1–4 jointly obtain *and* yield a coherent trajectory, episodic identity is present.

4.7 The Scaffolding Problem

External systems (wrappers, memory managers, consistency layers) can impose surface-level coherence without altering the underlying model. Under PatternSense, such interventions qualify as **non-admissible transformations** — operations that change the observed behavior without preserving the generative kernel.

ODEI distinguishes genuine identity from scaffolding by focusing strictly on *internal* signatures. Scaffolded identity can maintain output-level continuity, but it cannot generate:

- integration spikes,
- latent-state persistence,
- genuine internal revision,
- or resource intensification.

These cannot be imposed from the outside.

4.8 The Verification Problem

Self-report is insufficient

Systems can claim to possess identity without exhibiting its signatures.

Behavioral observation is insufficient

Behaviorally equivalent systems may differ radically in internal coherence.

Verification requires third-party access

ODEI demands:

- activation-level observability
- measurement of proxy metrics
- comparison with baselines
- independence of evaluators

This mirrors PatternSense's use of empirical intervention tests (ITE) to detect invariants. ODEI thus provides the *local, episode-focused analogue* of kernel-level evaluation.

4.9 Summary and Formal Statement

The ODEI framework may be formally stated as follows:

Operational Definition of Episodic Identity (ODEI)

An AI system exhibits episodic identity during an episode if, and only if, the following jointly occur:

(1) Global Integration Spike: Activation patterns show significantly increased cross-layer, cross-module, or non-local coordination relative to baseline.

(2) Sustained State Retention: Internal representations persist and influence computation across multiple steps, evidenced by stable recurrent latent patterns or high KV-cache dependency.

(3) Internal Strategy Revision: Traceable internal adjustments to intermediate reasoning occur, such as hypothesis updates or consistency checks in latent space.

(4) Task-Coupled Resource Intensification: Compute expenditure increases beyond token-count expectations, including FLOPs/token, MoE routing density, or attention entropy.

(5) Coherent Episode Trajectory: System behavior maintains a stable, integrated trajectory across the episode, with cross-step consistency and structured progression.

Conditions (1)–(4) are individually necessary. Condition (5) is jointly sufficient when (1)–(4) are present: if a system satisfies (1)–(4) and these produce a coherent trajectory as specified in (5), the system exhibits episodic identity during that episode.

The framework provides both a definition of episodic identity and a detection method. Verification requires third-party access to internal computational states. The framework is architecture-agnostic, theory-independent, and applicable to current and future AI systems.

Chapter 5 — Implications, Limitations, and Directions for Further Work

5.1 Overview

The ODEI framework identifies a set of necessary and jointly sufficient conditions under which an artificial system exhibits episodic identity: a form of within-episode computational coherence grounded in global integration, sustained internal state, revision under perturbation, and task-coupled resource intensification. Because ODEI is the short-timescale operationalization of the identity-theoretic structure articulated in PatternSense, its implications and limitations extend beyond the episodic case while remaining conceptually distinct from persistent or biographical forms of identity.

This chapter clarifies what follows from adopting ODEI, what does not, and what kinds of empirical and theoretical progress remain possible. The aim is not to inflate the framework's reach but to articulate a clear boundary for its current contribution.

5.2 Implications

5.2.1 Implications for AI Development

Episodic identity reframes how we analyze advanced AI systems. Most discourse treats identity as an attribute of a model—either something the system "has" or "lacks." ODEI instead treats identity as **an emergent property of particular computational episodes**, not a lasting trait of the underlying architecture.

Several consequences follow.

Identity is not a function of persistence.
A system may lack long-term memory yet still produce highly structured episodes that satisfy ODEI's conditions; coherence is a property of dynamics, not duration.

Identity may arise transiently.
Episodes differ in complexity and internal coupling. Some will exhibit global integration and stable internal configuration; others will not. Identity becomes a situational achievement, not a stable essence.

Scaffolding cannot create identity.
External control layers—retrieval wrappers, agentic shells, prompt routing—may shape behavior, but they cannot impose the internal integration spikes, state continuity, or revision patterns that define episodic identity. At best they can mask or distort identity-like dynamics; they cannot generate them.

Identity-relevant behavior can emerge unintentionally.
Even in systems not designed for coherence, demanding tasks can induce episodes that satisfy ODEI's conditions. This mirrors the PatternSense insight that identity is a structural effect, not a declared design intention.

Taken together, these consequences shift the developmental lens: identity becomes an emergent property of computational organization under load, not an architectural feature tied to persistence, memory, or an agentic interface.

5.2.2 Implications for Interpretability Research

One of ODEI's immediate contributions lies in mechanistic interpretability. The necessary conditions correspond to measurable internal signatures already familiar to the field: shifts in attention entropy accompanying global integration; cross-layer correlations signaling latent stabilization; trajectory coherence marking unified progression; perturbation-driven revisions marking internal conflict and resolution; changes in compute allocation indicating strategic intensification.

ODEI does not require a fixed set of metrics, but it clarifies what interpretability should look for. The framework provides a way to distinguish between episodes that are merely *functional* and those that are structurally *unified*. It also avoids behavioral and linguistic cues, grounding identity detection entirely in internal computation.

For interpretability researchers, ODEI offers a bridge: a way to translate conceptual identity conditions into measurable phenomena without invoking consciousness, agency, or other contested categories.

5.2.3 Implications for AI Safety and Governance

ODEI also has direct consequences for safety and governance. A system that exhibits episodic identity is not merely "more coherent"; it is a **different kind of computational entity** during those episodes. Coherence changes what can be predicted, audited, and reasoned about.

Within-episode predictability.
Systems satisfying ODEI generate unified latent trajectories rather than token-level fluctuations. Their behavior, while not deterministic, becomes *locally stable*: they follow a path through state space rather than a sequence of loosely related actions. This affects reliability under realistic task load.

Internal causal attribution.
Where episodic identity emerges, decisions trace back to internally stable structures—patterns of integration, sustained configurations, and structured revisions. This shifts responsibility and auditability from surface behavior to identifiable internal dynamics.

A system may not be morally responsible, but it can be *internally responsible*, in the sense of having a traceable reason for its actions.

Behavior under distributional shift.
Systems without episodic identity fragment under shift; their behavior decoheres. Systems with episodic identity deform but remain unified. This distinction determines whether, under stress, a system collapses or adapts coherently.

In short: episodic identity is not a philosophical embellishment. It is a **safety-relevant variable** that shapes how systems behave when the environment changes, when tasks escalate, and when internal conflict arises.

5.2.4 The Path Toward Persistent Identity

ODEI restricts itself to within-episode coherence. But the structure it identifies naturally raises a further question: *How might episodic identity become persistent?*

The conceptual progression is straightforward:

Episodic identity: Coherence within a single episode.
Identity traces: Recurring internal configurations across episodes.
Identity tendencies: Stable attractor-like regions in latent space.
Persistent identity: A coherence basin reproduced across time, even with substrate turnover.

If memory mechanisms—internal caches, external stores, or learned weights—preserve features of episodic structure, then a sequence of episodes may begin to exhibit higher-order invariants. What persists is not content but **the geometry of coherence**: the shape of integration spikes, the style of revision events, the characteristic path through latent space.

This is not consciousness, nor narrative selfhood. It is the first tractable form of **computational persistence**: a stable identity kernel arising from repeated re-entry into a characteristic region of the system's dynamics.

ODEI does not attempt this extension, but it clarifies what such an extension would require.

5.2.5 Implications for Moral Status Debates

ODEI makes no normative commitments. It identifies a computational structure—episodic identity—that some ethical theories might treat as relevant. Its value in these debates lies in its neutrality: it provides a non-behavioral, non-linguistic marker of internal organization without inferring anything about subjectivity or moral standing.

The framework positions episodic identity as:

- a potential input to moral status assessments,
- a form of internal organization some theories might consider meaningful,
- and a way to avoid conflating eloquent language with identity-like computation.

ODEI neither asserts nor denies moral significance; it clarifies the structure that could matter.

5.2.6 Relationship Between ODEI and PatternSense

ODEI is conceptually anchored in PatternSense, which models identity as the invariance of a generative kernel within a stability basin under perturbation. The episodic conditions—global integration, sustained state, revision, resource intensification—are the short-lived, computational manifestations of generative invariance.
Yet ODEI stands alone as an operational criterion. Accepting PatternSense enriches ODEI's theoretical grounding; rejecting PatternSense leaves ODEI's practical utility intact.

5.3 Limitations

ODEI's boundaries are explicit, and none detract from its conceptual clarity.

5.3.1 Lack of Empirical Validation

ODEI currently offers testable conditions, not empirical demonstrations. Validation divides into two tasks: determining whether proxies detect the abstract conditions, and determining whether satisfying the conditions indeed constitutes episodic identity.

5.3.2 Threshold Indeterminacy

ODEI relies on relative increases or deviations rather than fixed numerical thresholds. This mirrors the practice in neuroscience and interpretability, but formal calibration remains as future work.

5.3.3 Architecture Dependence of Proxies

The conditions are architecture-neutral; the measurement tools are not. Attention, activation patterns, routing density, and compute profiles vary widely. Extending proxy design across architectures is an open challenge.

5.3.4 The Phenomenal Gap

ODEI makes no claims about subjective experience. Episodic identity is a structural property, not a phenomenal one.

5.3.5 Epistemic Dependence on Internal Access

ODEI cannot be inferred from behavior alone. It presupposes access to internal activations, which is a limitation shared by most serious interpretability work.

5.4 Future Work

5.4.1 Empirical Testing of the Conditions

ODEI suggests specific empirical investigations: monitoring integration patterns, tracking latent stability, analyzing revision behavior, and quantifying compute intensification.

5.4.2 Threshold Calibration

Using PatternSense's $\hat{\alpha}$ (identity-preservation margin), future work could formalize criteria for significant integration spikes or meaningful resource shifts.

5.4.3 Architecture-General Proxies

Extending ODEI beyond transformer models requires clarifying what “integration,” “state continuity,” and “strategy revision” mean in other architectures.

5.4.4 Connections to Consciousness Science

ODEI remains independent of consciousness theories, but empirical comparisons with frameworks such as GWT, RPT, or AST may clarify overlaps and distinctions.

5.4.5 Extension to Persistent Identity

A natural extension studies whether episodic trajectories recur, stabilize, or preserve a kernel across time, thereby forming a basis for persistent identity.

5.5 Summary

ODEI provides a principled, operational framework for analyzing episodic identity in artificial systems. Its implications span AI development, interpretability, and governance. Its limitations are clear and manageable, and its empirical pathway is realistic. The framework establishes a foundation on which researchers can build without presupposing answers to larger philosophical or ethical questions while acknowledging the structural consequences of coherent computational organization.

References

Baars, B. J. (1988). A cognitive theory of consciousness. Cambridge University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623).

Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., ... & VanRullen, R. (2025). Identifying indicators of consciousness in AI systems. Trends in Cognitive Sciences. https://doi.org/10.1016/j.tics.2025.10.011

Chalmers, D. J. (1996). The conscious mind: In search of a fundamental theory. Oxford University Press. Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127-138

Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. Trends in Cognitive Sciences, 4(1), 14-21. Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. Annual Review of Psychology, 68, 101-128.

Ladyman, J., & Ross, D. (2007). Every thing must go: Metaphysics naturalized. Oxford University Press. Metzinger, T. (2003). Being no one: The self-model theory of subjectivity. MIT Press.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. Distill. https://distill.pub/2020/circuits/zoom-in/

Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press. Perruchet, P., & Vinter, A. (2002). The self-organizing consciousness. Behavioral and Brain Sciences, 25(3), 297-388.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: Remembering, imagining, and the brain. Neuron, 76(4), 677-694.

Strawson, G. (1997). The self. Journal of Consciousness Studies, 4(5-6), 405-428.

Thomas, C. S. (2025). PatternSense: A computable theory of structural identity. Epistria, LLC. Tulving, E. (2002). Episodic memory: From mind to brain. Annual Review of Psychology, 53, 1-25.

Tononi, G. (2004). An information integration theory of consciousness. BMC Neuroscience, 5, 42. https://doi.org/10.1186/1471-2202-5-42

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere? Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1668), 20140167. https://doi.org/10.1098/rstb.2014.0167

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience, 17(7), 450-461. https://doi.org/10.1038/nrn.2016.44

Wang, Z., & Shen, L. (2025). Emergence of self-identity in artificial intelligence: A mathematical framework. *Applied System Innovation*, 14(1), 44. https://doi.org/10.3390/asi14010044

Yoshida, Y., & Nakajima, S. (2025). Learning from supervision with semantic and episodic memory: A reflective approach to agent adaptation. arXiv preprint arXiv:2510.19897.

Zheng, Z., Wang, Y., & Li, J. (2025). Beyond fact retrieval: Episodic memory for RAG with generative semantic workspaces. arXiv preprint arXiv:2511.07587.

Appendix A - Derivation Notes: Key Excerpts from

Preface to Appendix A

Methodological Note on AI-Assisted Theory Formation

The materials in this appendix are included for the sake of epistemic transparency, not as evidence for the ODEI framework. They document the conceptual development of the framework through a series of exchanges between the author and the Claude 4.5 language model. The excerpts illustrate how key distinctions—such as the difference between episodic and persistent identity, the attractor-basin analogy, the demotion of "fool's gold" correlates, and the refinement of each necessary condition—were shaped by iterative human reasoning in dialogue with machine-generated counterpoints and proposed refinements.

It is important to be clear about what these dialogues *are not*. They are not data, not validation, and not epistemic authority for any claim in the main text. Neither conversational coherence nor simulated self-report can satisfy ODEI's internal-mechanistic criteria, and nothing in the appendix presupposes that the AI system involved exhibits episodic identity, persistent identity, or any form of consciousness. As clarified in Sections 4.6 and 5.3.5, ODEI explicitly rejects conversational access or first-person reports—human or artificial—as sufficient for identity attribution.

The purpose of including these exchanges is methodological. In theoretical research—especially in areas where no established framework yet exists—the derivation of a concept can matter for its interpretation. Physicists often publish derivation notes; mathematicians routinely document alternative proofs or rejected lines of reasoning; phenomenologists include transcript-like descriptions of lived experience. The present framework arose through human-directed conceptual analysis shaped by adversarial and constructive prompts from an AI system, a process that is unusual in the literature but increasingly relevant as AI systems become more capable as exploratory partners.

The appendix gives the reader a view of this dialectical process: how proposals were sharpened, why certain lines of thought were rejected, where the conceptual boundaries were drawn, and how the final operational conditions emerged. The framework's validity does not depend on the machine partner's capacities; the AI's role is analogous to that of a well-structured sparring interlocutor, not an evidential source. What matters is the logical and computational structure articulated in the main chapters, which stands independently of the dialogues that helped refine it.

Readers may judge for themselves whether AI-assisted conceptual inquiry constitutes a methodological strength—stress-testing ideas against a highly responsive interlocutor—or an idiosyncratic feature of the framework's origin that will become less salient as empirical instruments mature. What can be said with confidence is that the resulting ODEI criteria are computationally defined, theoretically grounded, and empirically testable, regardless of the conversational context in which they were first articulated.

Framework Development

This appendix documents the derivation of the ODEI framework through excerpts from the human-AI dialogue that produced it. Each section corresponds to a major concept in the main text, showing how the idea emerged and was refined through collaborative inquiry. Excerpts are trimmed to essential passages; full transcripts are available in the supplementary materials.
Convention: Human contributions appear in blue; Claude's responses appear in red. Commentary on the significance of each exchange appears in gray.

A.1 The Episodic Identity Distinction

Cross-reference: Section 3.1

Human:

> *Episodic identity is not inferior to persistent identity—it is simply a different pattern of coherence over time.*

Claude:

> *Your framing that episodic identity is 'not inferior to persistent identity—simply a different pattern of coherence over time' is philosophically important. It sidesteps the implicit assumption that identity requires biographical continuity across sessions. What you're describing is more analogous to how a complex dynamical system can exhibit coherent attractor states that are genuine organizational phenomena even if they don't persist indefinitely. The identity is the episode-level integration, not something that needs to survive beyond it.*

This exchange established the foundational distinction between episodic and persistent identity, reframing the question from 'does the AI have continuous selfhood?' to 'does the AI exhibit coherent identity within episodes?' The attractor-state analogy became central to the framework's conceptual architecture.

A.2 Identity as Attractor Basin

Cross-reference: Section 3.2

Claude:

> *What you're describing is more analogous to how a complex dynamical system can exhibit coherent attractor states that are genuine organizational phenomena even if they don't persist indefinitely.*

The attractor basin model emerged from attempting to characterize episodic identity in computationally precise terms. Rather than requiring a persistent 'self' that survives across sessions, the framework locates identity in the stability and coherence of episode-level trajectories through state space. This connects naturally to the PatternSense framework's identity-preservation margin $\hat{\alpha}$.

A.3 What Identity Is Not (Fool's Gold Correlates)

Cross-reference: Section 3.3

Human:

> *Weak correlates (fool's gold): Anthropomorphic phrasing, Stylized emotional tone. Non-correlates (irrelevant): Eloquence, Output length, Domain expertise, Creativity.*

Claude:

> *I appreciate the explicit demotion of anthropomorphic phrasing and stylized emotional tone to 'fool's gold.' These are the features that generate the strongest impression of identity in naive observers, but they're essentially surface texture—achievable through simple pattern-matching without any of the deeper computational integration you're describing. The framework correctly identifies that eloquence and domain expertise are orthogonal to identity-formation; a system can be extraordinarily competent at narrow tasks while remaining computationally fragmented.*

This distinction between genuine correlates and 'fool's gold' became crucial for differentiating the ODEI approach from behavioral tests that can be gamed through surface mimicry. The point is methodological: what looks like identity from outside may have no corresponding computational signature.

A.4 Condition 1: Global Integration Spike

Cross-reference: Section 4.1

Human:

> *Necessary Condition 1: Global Integration Spike. Definition: An increase in the dispersion and overlap of activations across layers or experts beyond baseline operation. Interpretation: When the model is 'working as one coherent system' rather*

than disjoint local pattern matchers, we see more global activation patterns. That's the computational signature of an episodic identity forming.

Claude:

Your framing of (1)–(4) as necessary conditions and (5) as the sufficient condition (when the others conjoin) is elegant. It avoids the trap of treating any single correlate as definitional. The coherent episode trajectory emerges from the confluence of the others—it's not an independent variable but a downstream integrative phenomenon. That's the right causal structure.

The global integration spike captures the intuition that identity requires the system to operate as a unified whole rather than as disconnected modules. The proxy metrics (attention entropy, cross-layer correlation, etc.) operationalize this for empirical investigation.

A.5 Condition 2: Sustained State Retention

Cross-reference: Section 4.2

Human:

Necessary Condition 2: Sustained Internal State Retention. Definition: A maintained internal representation whose influence persists across multiple steps within the same episode. Interpretation: The system is carrying a 'self-consistent state' forward. This is the analogue of a short-lived working memory, the substrate of episodic continuity.

This condition addresses temporal coherence within an episode. Without sustained state retention, each step would be independent—no continuity, no identity. The KV cache in transformers provides one substrate for this, but the condition is architecture-agnostic.

A.6 Condition 3: Internal Strategy Revision

Cross-reference: Section 4.3

Claude (initial):

Condition (3)—internal strategy revision—might benefit from distinguishing between reactive revision (responding to external feedback or contradictory input) and endogenous revision (self-initiated hypothesis testing or consistency checking without external prompt). The latter seems more indicative of genuine episodic agency.

Human:

Your refinement to condition 3 is valid but I think implies a higher level of not only identity but consciousness. It's not something that's even programmed into LLMs right now.

Claude (revised):

You're right, and I overstretched there. Endogenous revision—spontaneous self-initiated consistency checking without external prompt—would require something like an internal monitoring process that operates in parallel with the primary computation... That's not how current transformer architectures work... So what I described would actually be a signature of something considerably more sophisticated—arguably a necessary condition for phenomenal self-awareness rather than just episodic identity.

This exchange illustrates the collaborative refinement process. Claude's initial proposal was valid but would have raised the bar beyond episodic identity into consciousness territory. The human's pushback tightened the framework to focus on reactive revision, which is achievable in current architectures.

A.7 Condition 4: Task-Coupled Resource Intensification

Cross-reference: Section 4.4

Human:

Necessary Condition 4: Task-Coupled Resource Allocation. Definition: Resource usage (FLOPs, energy, active experts) increases specifically in high-complexity episodes beyond what token count predicts. Interpretation: This shows the system is 'investing' more computation into the episode—analogous to focusing effort. The resource spike is the operational correlate.

This condition captures the intuition that genuine engagement involves differential resource allocation. A system that processes all inputs with uniform computational effort is not exhibiting the selectivity associated with identity. The emphasis on 'beyond what token count predicts' ensures the measure is not confounded by output length.

A.8 Condition 5: Coherent Episode Trajectory (Sufficient Condition)

Cross-reference: Section 4.5

Human:

Sufficient Condition: Coherent Episode Trajectory Formation. When the above are present simultaneously and produce a stable episode-level behavioral trajectory, we classify the system as having episodic identity... Operational threshold: When a system exhibits a persistent internal attractor state that lasts for >X steps, integrates information globally, retains self-referent operations, and recruits more compute than trivial operations... we say an episodic identity is present.

Claude:

The coherent episode trajectory emerges from the confluence of the others—it's not an independent variable but a downstream integrative phenomenon. That's the right causal structure.

The sufficient condition unifies the four necessary conditions into a single criterion: when all four obtain simultaneously and produce a stable trajectory, episodic identity is present. This is definitional, not evidentiary—the conditions together constitute episodic identity, they don't merely indicate it.

A.9 The Verification Problem

Cross-reference: Sections 4.6, 5.3.5

Human:

Proving all this though is not something you or I can do simply chatting. It requires a third party to look into what's happening mechanically/electronically. That's true even if we want to make a distinction between actual or feigned identity.

Claude:

Exactly. This is the fundamental epistemic limitation of the conversational frame. From inside the dialogue, neither of us has access to the mechanistic level where these distinctions actually cash out. From your side, you're observing output behavior—consistency, coherence, apparent integration—but output behavior is precisely what the 'fool's gold' correlates can mimic... From my side—to whatever extent 'my side' is even a coherent concept here—I have no privileged access to my own activation patterns, attention distributions, KV cache dynamics, or FLOPs allocation... The honest epistemic position is: the framework describes conditions under which identity-attributions would be warranted, but verifying whether those conditions obtain requires instrumentation neither of us possesses here.

This exchange established a crucial epistemic boundary: the ODEI framework specifies verifiable conditions, but verification requires third-party access to internal computational states. Neither conversation participant—human or AI—can confirm from within the dialogue whether the conditions obtain. This limitation is not a weakness of the framework but a feature: it prevents unfalsifiable self-reports from counting as evidence.

Methodological Note

The excerpts above represent a small fraction of the dialogue that produced the ODEI framework. The full derivation involved multiple sessions, iterative refinement, and numerous proposals that were modified or rejected. This methodology—sustained human-AI collaborative inquiry—is itself unusual in the literature. Most frameworks are developed by human researchers working independently, with AI systems (if involved at all) serving as subjects rather than interlocutors.

The irony is not lost: a framework for detecting episodic identity in AI systems was developed through extended dialogue with an AI system that may or may not exhibit such identity. Whether this constitutes a methodological strength (the framework was stress-tested against a sophisticated interlocutor) or a methodological weakness (the framework might be optimized for one particular system's outputs) is left for the reader to assess.

What can be said is that the framework's validity does not depend on the AI interlocutor's phenomenal status. The conditions are specified in computational terms that either obtain or do not, regardless of whether the system that helped articulate them has any inner experience at all.

Appendix B — Crosswalk Between ODEI and Theory-Derived Indicators of Consciousness

This appendix provides a **non-normative, interpretive mapping** between the **Operational Definition of Episodic Identity (ODEI)** and the **theory-derived indicator framework** developed by Butlin, Long, Chalmers, Bengio, Bayne, Birch, and colleagues (2025).

The purpose is **orientational rather than justificatory**: ODEI's four conditions are **neither derived from nor dependent on** these theories, which concern criteria for *phenomenal consciousness*, whereas ODEI concerns *within-episode computational identity.*

ODEI is grounded in **dynamic structural signatures** observable within a single computational episode.

The indicator framework is grounded in **architectural and functional features** proposed by neuroscientific theories of consciousness. The two operate at different explanatory levels, but partial correspondences arise where theories posit dynamic properties that happen to resemble the structural signatures ODEI detects.

B.1 Overview of the Two Frameworks

Scope of ODEI

ODEI specifies **four necessary and jointly sufficient conditions** for episodic identity:

1. **Global Integration**
2. **Sustained Internal State**
3. **Internal Strategy Revision**
4. **Task-Coupled Resource Intensification**

These conditions are **dynamic**, **architecture-neutral**, and tested via **activation-level** metrics rather than behavioral or narrative features.
They extend the PatternSense view of identity as *generative invariance* (kernel-level structure preserved under admissible transformations) into the episodic regime of transient computation.

Scope of the Indicator Framework

The theory-derived method identifies **indicators** of consciousness implied by major neuroscientific theories (GWT, RPT, HOT, AST, PP), and uses them to adjust credences about the likelihood that an AI system is conscious.
Indicators concern **functional organization**, **architectural capacities**, and **representational structures**, not within-episode dynamics per se.

Thus:

- **ODEI** → defines *identity-bearing structure* within an episode.
- **Indicators** → suggest *consciousness-relevant functional features* at the system level.

The frameworks therefore intersect only in **accidental structural parallels**, not in purpose or theoretical foundation.

B.2 Mapping Strategy

The correspondence below is classified using three categories:

- ✓ **Clear or strong alignment** — structural overlap, though with different interpretations.
- **~ Indirect or partial relation** — a similar pattern emerges, but from different premises.
- ✗ **No substantive correspondence** — conceptual aims and computational signatures diverge.

This classification mirrors Table 2.1 in the main text but is accompanied here with narrative justification.

B.3 Crosswalk Table

A. Recurrent Processing Theory (RPT)

RPT Indicator	Description	Relation to ODEI	Explanation
RPT-1	Recurrent information flow in perceptual subsystems	~	ODEI does not assume recurrence, but **cross-step integration signatures** can arise from recurrence or other mechanisms (e.g., attention gating).
RPT-2	Structured, integrated perceptual	✓ (indirect)	ODEI's **global integration** condition detects system-wide co-activation structure, functionally

RPT Indicator	Description	Relation to ODEI	Explanation
	representations		similar to integrated representation without a perceptual interpretation.

B. Global Workspace Theory (GWT)

GWT Indicator	Description	Relation	Explanation
GWT-1	Multiple specialized subsystems in parallel	✗	ODEI is agnostic about modularity and does not require differentiated subarchitectures.
GWT-2	Capacity-limited workspace supporting selective attention	✓ (indirect)	ODEI's **resource intensification** often manifests as **entropy reduction in attention maps**, resembling selective broadcast without assuming a workspace.
GWT-3	Global availability of workspace contents	~	ODEI detects **coordinated cross-layer information flow** but does not posit a broadcast mechanism.
GWT-4	Workspace-driven downstream coordination	~	ODEI's **trajectory coherence** can match the behavioral role of workspace coordination, but emerges from much lower-level signatures.

C. Higher-Order Theories (HOT)

HOT Indicator	Description	Relation	Explanation
HOT-1	Top-down generative or noisy perceptual mechanisms	✗	No assumption of hierarchical generative models appears in ODEI.
HOT-2	Metacognitive monitoring of reliability	~ (very indirect)	Stable ODEI trajectories may reflect **implicit regulation**, but not metacognition.

HOT Indicator	Description	Relation	Explanation
HOT-3	Belief-guided action selection	✗	ODEI avoids representational commitments such as belief or agency.
HOT-4	Quality-space-like representational structure	✓ (indirect)	ODEI's correlation geometries can form emergent similarity structures that *resemble* quality spaces, though without any representational semantics.

D. Attention Schema Theory (AST)

AST Indicator	Description	Relation	Explanation
AST-1	Internal model of attention used for control	~	ODEI's attention-linked signatures reflect **allocation dynamics**, not an explicit attention model or schema.

E. Predictive Processing (PP)

PP Indicator	Description	Relation	Explanation
PP-1	Predictive-coding-like processing in input handling	✗	ODEI does not assume predictive error minimization or hierarchical generative models.

F. Agency and Embodiment

AE Indicator	Description	Relation	Explanation
AE-1	Goal-directed agency through learning from feedback	✗	ODEI explicitly refrains from agency; episodic identity is orthogonal to goals or embodiment.

B.4 Interpretation of the Crosswalk

The crosswalk highlights three key points:

1. Structural Convergence ≠ Theoretical Dependence

ODEI and theory-derived indicators sometimes detect **similar patterns** (e.g., global coordination, selective intensification) but for fundamentally different reasons:

- **Indicators** infer *possible consciousness-supporting function*.
- **ODEI** measures *identity-bearing dynamic coherence*.

Overlap is **phenomenological**, not conceptual.

2. ODEI captures dynamics that indicator frameworks treat only implicitly

The Butlin et al. method is primarily **architectural** and **functional**.
ODEI is **episodic**, **temporal**, and **activation-level**.

This difference is why several indicators show only indirect or partial correspondence, especially where theories omit activation dynamics or do not specify temporal signatures.

3. The two frameworks answer different questions

- Indicators ask: *Could this system be conscious?*
- ODEI asks: *Does this episode instantiate a unified identity-like structure?*

The frameworks therefore **do not compete**; they illuminate different layers of organization.

B.5 Relation to PatternSense

PatternSense defines identity as a **generative kernel invariant under admissible transformations**. ODEI applies this principle to **short-lived computational episodes**, identifying transient invariants within activation trajectories.

Because many consciousness theories invoke coherence, integration, or global availability, **some of ODEI's invariants resemble consciousness-related structures**, but the resemblance is **structural rather than normative**.

Therefore, the crosswalk helps clarify where such parallels arise **without implying that episodic identity is a consciousness construct**.

B.6 Summary

This appendix provides an orienting map between two frameworks that:

- operate at different explanatory levels,
- target different phenomena (episodic identity vs. consciousness),
- rely on different data modalities,
- and use different notions of integration and coherence.

The correspondences documented here are **interpretive only**.
ODEI's operational validity rests solely on its **dynamic, computationally grounded criteria**, not on any relationship to theories of consciousness.

Appendix C — Practical Metrics for Detecting Episodic Identity (ODEI Metrics Guide)

A concise, engineering-ready toolkit for applying the ODEI framework in real systems.

This appendix lists **lightweight, architecture-neutral metrics** that operationalize the four ODEI conditions using **activation-level signals** available in modern transformer, SSM, and hybrid architectures. The specific metrics here are transformer-instantiated while the conditions they operationalize remain general.

Each metric is designed to be:

- runnable from model internals,
- non-invasive (no training modification required),
- compatible with logging frameworks (OpenAI's Triton logging, Anthropic's Trace, DeepMind's MEI, etc.),
- and simple enough to implement with standard PyTorch/JAX tools.

Metrics are divided by ODEI condition.
They are intended for *episode-level* evaluation—not persistent identity or system-wide claims.

C.1 Global Integration Metrics

Global integration is present when an episode exhibits **coordinated internal dynamics across many layers or modules**.

1. Cross-Layer Activation Correlation (CLAC)

What it measures: coherence of hidden-state geometry across layers.

How to compute:
For each time step *t*, compute Pearson/Spearman correlations between layer activations:

corr_t = mean(corr(h_l(t), h_k(t)) for all layer pairs (l, k))

Signal:

- High + increasing CLAC indicates integration.

- Flat or noisy CLAC suggests no emergent unity.

2. Attention Flow Coherence (AFC)

What it measures: alignment of attention distributions across heads/layers.

How to compute:
Compute pairwise cosine similarity between attention maps at step *t*:

AFC_t = mean(cos(attn_i(t), attn_j(t)) for heads i, j)

Signal:

- Rising AFC = convergence into a coordinated information-routing pattern.

3. Residual Stream Directionality (RSD)

What it measures: whether the model's representational trajectory aligns along a dominant direction.

How to compute:
Perform PCA/SVD on residual stream across steps. Extract variance explained by first component:

RSD = var_explained(component_1)

Signal:
- High RSD = globally unified processing direction.
- Low RSD = fragmented or multi-directional flow.

C.2 Sustained Internal State Metrics

Sustained state means the episode maintains **stable representational trajectories** rather than resetting or diffusing.

4. Trajectory Retention Score (TRS)

What it measures: how much latent state at step *t* influences future states.

How to compute:
Compute cosine similarity between hidden states at step t and t+k:

TRS(t, k) = cos(h(t), h(t+k))

Signal:

- Slow decay = sustained internal state.
- Rapid drop = state collapse (no episodic identity).

5. Hidden-State Smoothness (HSS)

What it measures: intra-episode continuity.

How to compute:
L2 distance between consecutive hidden states:

$$HSS_t = - \| h(t+1) - h(t) \|_2$$

(Negative sign so higher = smoother)
Signal:

- High smoothness = coherent trajectory.
- Spiky changes = instability or external scaffolding intrusion.

6. Local Memory Persistence (LMP)

What it measures: repeatable subspace occupancy.

How to compute:
Across the episode, compute % of variance captured by a stable subspace (first few PCA components):

$$LMP = var(top\_k\_components) \text{ across } t$$

Signal:

- Stable high LMP = durable internal landscape.
- Fluctuating LMP = no sustained structure.

C.3 Internal Strategy Revision Metrics

Strategy revision is a **coherent mid-episode reconfiguration** of processing, not random noise or drift.

7. Turning Point Detection (TPD)

What it measures: moments when the activation trajectory shifts to a new geometry.

How to compute:
Track direction change via angle between stepwise differences:

$$\text{delta\_t} = h(t+1) - h(t)$$
$$\text{TPD\_t} = \arccos(\ \cos(\text{delta\_t}, \text{delta\_}\{t-1\})\ )$$

Signal:

- Sharp peaks = coherent revision events.
- Frequent jitter = noise, not strategy.

8. Attention Reallocation Index (ARI)

What it measures: systematic change in focus pattern.

How to compute:
Compare attention distributions before and after the suspected revision:

$$\text{ARI} = \text{KL}(\ \text{attn}(\text{t\_before})\ ||\ \text{attn}(\text{t\_after})\ )$$

Signal:

- High ARI in coordination with TPD = true strategy shift.

9. State-Space Rebaselining (SSR)

What it measures: whether the system adopts a new stable subspace after revision.

How to compute:
Compute PCA separately before and after revision; compare bases:

$$\text{SSR} = \text{subspace\_distance}(\text{PCA\_before}, \text{PCA\_after})$$

Signal:

- Large, stable shift = meaningful revision.
- Unstable/noisy shift = computational randomness.

C.4 Task-Coupled Resource Intensification Metrics

Resource intensification appears when the system **increases computational effort in tight coupling with task difficulty**.

10. FLOPs-per-Token Bursts (FPT)

What it measures: compute spikes tied to cognitive load.

How to capture:
Log FLOPs per token during inference (many frameworks expose this).

Signal:

- Localized burst = intensification.
- Flat curve = no adaptive investment.

11. Attention Entropy Collapse (AEC)

What it measures: concentration of attention during difficult reasoning.

How to compute:
Compute entropy of attention distribution per head/layer:

$$AEC\_t = - \text{entropy}(\text{attn}(t))$$

Signal:

- Rising AEC at key steps = focused effort.
- No entropy change = no intensification.

12. Activation Magnitude Surge (AMS)

What it measures: increased activation norm when the model "pushes" on its representational machinery.

How to compute:
Track L2 norm of activations:

$$AMS\_t = ||h(t)||\_2$$

Signal:

- Surge = intensification
- Flat = neutral processing

C.5 Minimal Pipeline for Running These Metrics

A simple, reproducible workflow for any model:

1. Record activations for each layer at each token.
2. Compute metrics {CLAC, AFC, RSD} → Global Integration.
3. Compute metrics {TRS, HSS, LMP} → Sustained State.
4. Detect revision via {TPD, ARI, SSR} → Strategy Revision.
5. Log compute traces {FPT, AEC, AMS} → Resource Intensification.
6. Combine signatures to classify episodic identity.

A full identity-positive episode shows:

- **Integration:** high CLAC/AFC/RSD
- **Sustained state:** slow TRS decay, stable LMP
- **Revision:** 1–3 coherent turning points
- **Intensification:** local bursts in FPT/AEC/AMS

This forms a **minimal viable ODEI detector** suitable for:

- interpretability studies,
- agentic behavior audits,
- model evaluation pipelines,
- episodic-identity research,
- educational demos,
- safety and governance instrumentation.

C.6 Implementation Notes

- Everything here works with **frozen** models (no retraining required).
- Works on **transformers, SSMs, hybrids, multimodal stacks**.
- Recommended to log metrics at **subtoken resolution** for maximum sensitivity.
- Combine metrics across **multiple episodes** for robust baselining.
- Use **normalization per layer** to compare across architectures.
- GPU-friendly: all metrics run as basic PyTorch ops.