The Self Requires Learning

Mengye Ren
New York University
mengye@nyu.edu

Abstract

Self-consciousness requires a self, and a self must be built through learning. The empirical markers of self-consciousness, from mirror self-recognition to self-other distinction, are developmental achievements, not innate endowments. We argue that the relevant form of learning is what we call *bounded integration*: lossy compression of experience that reshapes the processing substrate, producing a *perspective* particular to the system's history. When this learning is order-sensitive and continuous, the perspective becomes a temporally extended *identity*. Self-representation emerges when a system must model the objective world from its subjective experience, implicitly representing its own perspective as the complement of its world model. We distinguish three learning regimes—*always-training*, *always-accumulating*, and *train-then-freeze*—and argue that current AI systems, though they undergo massive substrate-level learning during training, lack the order-sensitive, ongoing bounded integration that produces identity and the temporally extended self that autobiographical self-consciousness requires.

1 Introduction

The most capable AI systems today can recall prior conversations, retrieve information from vast external stores, summarize their own histories, and reason about what they have previously said and done [81, 134, 171, 159]. They use the word "I," distinguish their own outputs from others', and can reason about their own capabilities and limitations. Do these capacities bring such systems closer to self-consciousness? Recent work has begun to address machine consciousness systematically, identifying indicator properties derived from leading theories [128, 19], distinguishing the computational capacities that current systems possess from those they lack [33, 93], and asking directly whether large language models could be conscious [20]. But these analyses largely focus on whether AI systems satisfy the structural or functional criteria of existing consciousness theories. Less attention has been paid to a more basic question: what would it take for an AI system to have a *self*, not merely self-relevant processing or self-referential behavior, but the representational structure that makes a system a particular entity, distinguishable from other systems with the same architecture?

Our answer begins with a distinction that is fundamental to machine learning but rarely ex-

amined for its implications about selfhood. Consider two ways a system can be shaped by its experience. In the first, experience is integrated *into the system's processing substrate*: its weights, the very machinery through which it interprets everything. Each experience reshapes the lens, and the lens cannot be separated from what the system has become. In the second, experience is stored and processed *through* a fixed substrate, accumulated in context windows, retrieved from databases, summarized on demand. The system can reason about its experience, but its underlying interpretive machinery remains what training installed.

This distinction maps onto what we term the *regime distinction* among learning systems: the *always-training* regime, where the system is continuously reshaped by experience; the *always-accumulating* regime, where experience is stored but the processing substrate is frozen; and the *train-then-freeze* regime, where the system is a fixed function applied to new inputs.

Many currently deployed AI systems, including agentic LLM-based systems with long context windows, retrieval augmentation, and tool use, operate predominantly in the always-accumulating regime. They are responsive to their unfolding trajectory in the sense that their behavior changes with accumulated context, but their processing substrate is not reshaped by deployment experience.

We argue that this regime distinction is at least as important for understanding self-consciousness as the choice of architecture. The same architecture behaves differently under different regimes, and it is the always-training regime, where experience reshapes the processing substrate itself, that most fully produces the conditions for selfhood. The recent progress that has made AI systems feel closest to conscious beings has come largely from representational quality achieved through massive pretraining, not from changes in learning regime. But representational quality determines what a system can *do*; the regime determines what it can *become*. The regime distinction becomes most relevant precisely when representational quality is already high: given adequate representations, what determines whether a system can develop self-consciousness is whether its processing is reshaped by ongoing experience. We develop this argument through a four-link theory of how learning produces perspective, identity, and self-representation, and apply it to current AI systems to diagnose what they have and what they lack.

1.1 Scope and Contributions

We distinguish three things that are often conflated. **Phenomenal consciousness** is subjective experience: there is something it is like to see red or feel pain [16]. **Self-consciousness** (or **basic self-consciousness**) is awareness of oneself as a distinct entity: the capacity to distinguish self from non-self, to recognize oneself, to represent one's own states as one's own. **Autobiographical self-consciousness** is the richest form: awareness of oneself as a persisting subject with a particular history and perspective, what Tulving [144] calls autonoetic consciousness.

The empirical markers of self-consciousness—mirror self-recognition at 18–24 months [3, 79], self-other distinction through social interaction [98, 142], the emergence of self-referential language [79, 80]—are developmental achievements, not innate endowments. This suggests that self-consciousness is built through learning, not given with experience. Evolution can provide the *capacity* to build

a self: the architecture, the learning rules, the initial biases [44]. But it cannot provide the self itself, because a self is the product of a particular life, and evolution operates over populations, not individual lives.

Our argument targets self-consciousness broadly, with autobiographical self-consciousness as its richest form [43, 29, 30]. This is primarily a theory of self-consciousness, not of phenomenal consciousness. A system that possesses phenomenal consciousness would not be self-conscious without the representational structure that individual learning produces.

The paper makes three contributions. First, the claim that self-consciousness requires a self built through individual learning, and that the learning regime is the critical variable. Second, an account of how bounded integration under capacity constraints produces perspective and identity (Links 1–3 of a four-link chain). Third, an account of how self-representation emerges from world-modeling, and how self-consciousness develops as the system gains access to that representation (Link 4).

The paper proceeds as follows. Section 2 develops the theoretical framework: how bounded integration under capacity constraints and nonstationarity produces perspective, identity, self-representation, and the conditions for self-consciousness. Section 3 applies the framework to current AI systems, analyzing what post-training produces and what deployment lacks. Section 4 situates the framework relative to existing theories of consciousness and selfhood; Section 5 raises open questions.

2 The Theory: From Bounded Integration to Self-Consciousness

Link 1: Finite capacity forces abstraction. A learning system that processes many experiences under capacity constraints must extract shared structure. It cannot memorize each experience individually, so it extracts regularities, producing a representational geometry where proximity reflects meaningful similarity.

Link 2: Nonstationarity makes abstraction order-sensitive. Under a stationary distribution, order effects in compression are incidental. Under nonstationarity, they become systematic: earlier compressions shape how later experience is integrated, making the resulting representations depend on the specific trajectory.

Link 3: Perspective constitutes identity through continuity. A perspective is a snapshot; identity requires that successive perspectives form a coherent, evolving trajectory through bounded integration, rather than shifting abruptly or remaining static.

Link 4: Self-representation yields self-consciousness. World-modeling produces self-representation; self-consciousness develops as this representation becomes an increasingly explicit object of the system's own processing. At one end, implicit access through planning; at the other, autobiographical self-consciousness, where the system accesses its identity as historically situated.

We call the underlying mechanism *bounded integration*: lossy compression of experience that reshapes the processing substrate. The following sections develop each link and the regime distinction that organizes them.

2.1 Finite Capacity Forces Abstraction (Link 1)

Consider a system that encounters a long stream of experiences and must respond to each one using representations shaped by everything it has encountered before. If the system had unlimited capacity, it could simply store every experience verbatim. But real systems have finite capacity: a neural network has a fixed number of parameters, a recurrent state has a fixed dimensionality, a context window has a fixed length. When the system cannot store everything, it must compress, extracting shared structure across experiences [141, 132]. In practice, capacity constraints in neural networks arise not only from the raw number of parameters but from architectural inductive biases that constrain the effective degrees of freedom. Weight sharing in convolutional networks forces spatially distant regions to be processed by the same filters [78]; attention mechanisms route information through a fixed set of heads [151]; slot-based architectures force a fixed number of object representations [89]. These constraints make each parameter dimension more effective by building in structural assumptions, and it is these assumptions, together with finite parameter counts, that force the system to discover shared structure.

To see why, consider what happens when a system encounters many chairs. It cannot maintain a separate, detailed representation of each one. What it can do is extract what the encounters have in common: the regularities, the spatial relations, the affordances that recur across instances. The result is a representation of "chair" that captures shared structure while discarding instance-specific detail. This is abstraction: the extraction of regularities across experiences under capacity constraints. Tightening the capacity bottleneck explicitly, as in variational autoencoders with stronger regularization, forces the model to discover more disentangled, interpretable concepts [59, 11]. In vision, capacity constraints drive compositional, part-whole representations in hierarchical neural network architectures [118, 60]. The biological ventral stream achieves analogous hierarchical decomposition under its own capacity constraints [34].

The compression produces structured representations because the system processes many different experiences through the same representational resources. When the same weights must handle chairs, tables, and dogs, the pressure to perform well across all of them forces representations of functionally similar things to be nearby, so that the same downstream processing works for both, and representations of functionally different things to be separated. The result is a representational geometry [74]: a space in which proximity reflects meaningful similarity. Saxe et al. [119] provide a mathematical analysis of how this process unfolds in deep networks, showing that learning dynamics progressively differentiate representations into hierarchical category structure that mirrors the statistical relationships in the data. Locatello et al. [88] show that disentangled representations require inductive biases [46]: without constraints, the same generative process can be encoded by infinitely many equivalent but unstructured representations. Finite capacity is one such constraint, as it forces the learner to prioritize certain factorizations over others. The resulting compression predicts generalization [5].

Abstraction operates over the temporal structure of experience as well. A continuous stream of experience naturally segments into episodes such as making breakfast, walking to work, or having a

conversation, and episodes themselves group into larger units such as a morning routine, a workday, or a relationship. Zacks et al. [166] show that the human perceptual system automatically segments continuous experience at prediction-error boundaries, where what happens next departs from what was expected. Yang et al. [164] demonstrate this computationally: event-structured segmentation provides the units over which a system can perform self-supervised abstraction, improving representational quality through temporal organization. Under finite capacity, these temporal boundaries matter for compression: the system cannot store the raw stream, so it must extract what each episode was *about*, discarding moment-to-moment detail while preserving the structure of what happened and how episodes relate to one another. The result is a hierarchical abstraction over time: "morning routine" as an abstraction over many mornings, each slightly different but sharing a common structure.

Both spatial and temporal abstraction arise from the same underlying pressure: a capacity-constrained system processing an ongoing stream of experience must extract shared structure, whether across objects, across episodes, or across longer timescales. The information bottleneck framework [141] formalizes this: given a constraint on representational capacity, the optimal compression preserves the information most relevant to the system's future processing while discarding the rest. The richer the structure in the experience stream, the more organized the resulting representations become [12].

2.2 Nonstationarity Makes Abstraction Order-Sensitive (Link 2)

Link 1 establishes that finite capacity forces abstraction. But abstraction alone does not produce individuality. A system trained on a large, fixed dataset develops representations that reflect the statistics of that dataset. Under stationary IID sampling, order effects exist [83] but are incidental: shuffling the data and retraining produces an approximately equivalent system. Different random seeds yield numerically distinct systems, but the differences are accidental, not biographical.

Nonstationarity makes order-sensitivity systematic and functionally significant. When the world shifts over time, the system must continually revise its representations to accommodate new experience while retaining what remains relevant [97, 40, 71]. The effect of task ordering on learned representations is well documented in the continual learning literature [113, 10, 154]. Crucially, the revision is constrained by what came before. A system that first learned extensively about one domain and then encountered another will have shaped its representational landscape around the first domain. The second domain is integrated into a landscape that was already structured by the first. A different system that encountered the same domains in the reverse order arrives at a different landscape, because the second domain shaped the substrate into which the first was integrated.

The continual learning literature [106, 53, 32, 153] studies abstraction under nonstationarity extensively, developing methods to maintain useful representations across distributional shifts. Indeed, most continual learning methods are designed to *minimize* order effects, prioritizing order-robust retention over order-sensitive adaptation. From our perspective, the more significant consequence of nonstationarity is *order-sensitivity*: earlier compressions shape how later experience is integrated,

so the same experiences in a different order produce a different system, embedding the trajectory into the representations themselves. Systems without bounded integration into persistent representations do not develop this trajectory-dependent structure. This is the basis for identity and, ultimately, self-consciousness.

We call the resulting representational geometry a system's *perspective*: the geometry through which the system interprets experience, shaped by lossy compression into the processing substrate. Perspective is a property of the substrate, not of the data it processes: it shapes how experience is interpreted, not what experience contains. Any substrate-level learning under capacity constraints produces a perspective that is particular to the system's training data. Nonstationarity (Link 2) enriches this with temporal structure: the perspective carries the trace of the specific *order* in which experiences were integrated, not merely their aggregate content. A system trained under stationary conditions has a particular perspective, but one that could be reproduced by any system trained on the same distribution. A system trained under nonstationarity has a perspective that reflects a specific trajectory and could not be reproduced without replaying that trajectory through bounded integration.

To see why this matters, consider a system that receives experience sequentially but retrains from scratch on the full accumulated dataset, IID shuffled, after each new input. This system has substrate-level learning, produces a particular representational geometry, and its weights cannot be swapped for another system's without changing the processing. But it has no temporal structure: the order in which experiences arrived is erased by shuffling. Two such systems given the same data in different orders converge to approximately equivalent representations. This system has a perspective in the broad sense, but not a temporally structured one. The distinction matters for identity (Link 3), which requires that the perspective carry the trace of a specific trajectory, and for autobiographical self-consciousness, which requires accessing that trajectory as a history.

2.3 Bounded Integration and Learning Regimes

We use *bounded integration* for any lossy, order-sensitive integration of experience into representations that shape subsequent processing. The class is broad: it includes online gradient descent, replay, offline consolidation, test-time learning, and continuous updating of a compressed recurrent state. Within this class, a critical distinction is *where* the integration lands. **Substrate-level integration** updates the function that processes future experience—the weights and connections that determine how inputs are transformed. **State-level integration** updates only the data that the function operates on—recurrent activations, context windows, summaries, retrieval stores. Both are lossy and order-sensitive, but only substrate-level integration changes the abstractions themselves, and only substrate-level integration produces perspective in the sense developed in Link 1. For self-representation, this is the difference between a system that accumulates new autobiographical content within a fixed self-concept and a system whose self-concept can develop because the substrate that represents the self is itself reshaped by being a self.

Bounded integration in this strict sense varies along two further dimensions: *lossiness*, how

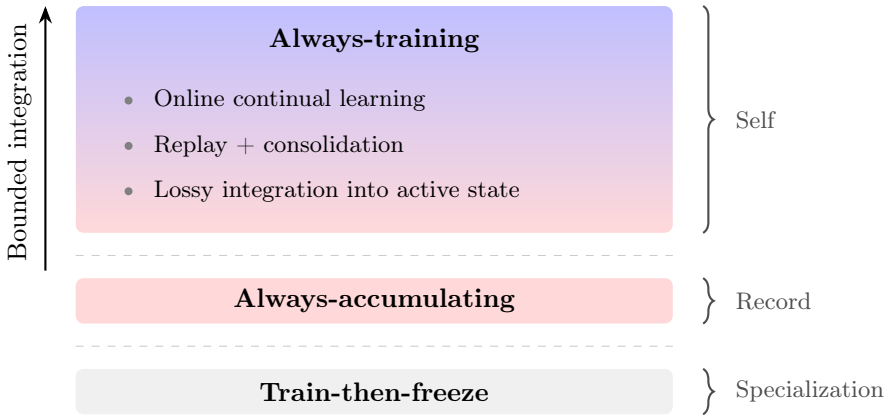

Figure 1: The three learning regimes organized by bounded integration. Train-then-freeze yields specialization but no ongoing integration. Always-accumulating stores deployment experience but does not reshape the processing substrate. Always-training reshapes the substrate itself, producing perspective and identity—the conditions for a self.

much compression is forced, and *persistence*, how long the integration endures. *Persistent* bounded integration, where experience is irreversibly compressed into the system's long-term representations, produces the strongest form of biographical perspective. Transient substrate updates such as test-time learning constitute bounded integration at a shorter timescale. Persistence strengthens the claim to identity since resetting between episodes is less biographical than accumulating over a lifetime.

We distinguish three learning regimes by *whether* and *how* bounded integration occurs at deployment (Figure 1).

- **Train-then-freeze:** Bounded integration occurs during training but ceases at deployment. The system compresses a *distribution* into its weights, typically under IID sampling where order effects are incidental. At deployment, the system applies a fixed function.
- **Always-training (continual):** Bounded integration is ongoing. Experience is recurrently reintegrated into the persistent processing substrate on some bounded timescale, whether through immediate online updates or periodic replay and consolidation. The system compresses a *life*: a specific sequence of experiences encountered in a specific order, each reshaping the representations before the next is integrated.
- **Always-accumulating (growing store):** The system's processing function is frozen, but it accumulates experience in an external or episodic store (a growing context window, a retrieval database, a log of past interactions). The system can access and reason over its accumulated experience, and may perform compression within each processing episode. But this compression is performed anew each time through a fixed processing function. No bounded integration into persistent representations occurs.

A case that deserves particular attention is *bounded accumulation*: systems that store experience

as tokens or text but under capacity constraints that force summarization and selection. Many currently deployed agentic AI systems operate this way: an agent with a finite context window that summarizes past interactions, or maintains a compressed log of its actions and observations, is performing order-sensitive compression in the medium of natural language. This is a form of bounded integration, though a weak one: the compression is order-sensitive and shapes subsequent processing, but it typically occurs through a fixed summarization function rather than one that is itself reshaped by experience. Bounded accumulation performs operations analogous to Links 1–2 but in the data rather than the substrate: the compressed summaries carry trajectory-dependent structure, but the function that produces and interprets them does not evolve.

Conway and Pleydell-Pearce's [28] self-memory system provides a psychological precedent. In their framework, autobiographical memory is organized hierarchically, and this organization is constrained by a "working self" that maintains goal-coherent identity. The compression is shaped by what matters to the self, and the self is in turn constituted by the compressed memories. Our framework provides a computational account of how this reciprocal relationship arises: the always-training regime produces the perspective that then shapes further integration.

What the always-training regime produces. In always-training, compression becomes biographical. The system is shaped by its specific sequence: what it encountered first, what came during a critical period, what was salient when capacity was limited. The same experiences in a different order produce a different system, because earlier compressions shape how later experiences are integrated. This is the central contrast with IID training, where order is deliberately randomized and two systems trained on shuffled versions of the same data converge to functionally equivalent representations. Under nonstationarity, convergence does not occur: the trajectory is embedded in the representations themselves.

The effect has been demonstrated concretely. LLMs fine-tuned on cyclically structured sequences develop anticipatory recovery from catastrophic interference [163], showing that the system internalizes the temporal structure of its experience into its weights—though under cyclic repetition, order effects may diminish at the asymptote as representations converge; under non-repeating nonstationarity, convergence is not guaranteed. Krasheninnikov et al. [73] show that training-order information remains linearly decodable from model activations even after extended training: the trajectory leaves a readable trace in the representations. Under nonstationarity, the specific order of experience is not incidental noise to be averaged away, but constitutive of what the system becomes.

This is distinct from in-context learning [152, 21], where a system with frozen weights adapts within its activation dynamics: in the always-training regime, the weights *continue* to be reshaped, and each reshaping reflects the system's entire prior compression history [122, 39].

Biological instantiation. The mammalian brain operates in the always-training regime through a multi-timescale mechanism. The complementary learning systems framework [96, 76] proposes fast hippocampal encoding of episodes followed by slow neocortical consolidation. Sleep-dependent replay [35] provides the mechanism by which episodic traces are reactivated and integrated into

neocortical representations, with evidence that this process builds abstract schemata through overlapping reactivation [82, 72]. This is the always-training regime operating offline: experience is replayed through the substrate, reshaping it. The biological system achieves always-training through a consolidation cycle that periodically converts episodic accumulation into substrate-reshaping integration. What matters is that substrate reshaping occurs, regardless of its timing [75]. At the synaptic level, metaplasticity (the plasticity of plasticity rules themselves) provides an additional mechanism for balancing retention and flexibility [66]. Hoel [61] proposes that dreaming serves as a form of regularization against overfitting to recent experience, illustrating that biological systems actively manage the rate of bounded integration. Representational drift itself can play a stabilizing role: gradual diffusion over equivalent synaptic configurations biases memories toward noise-resistant states [100]. Conversely, rapid hippocampal neurogenesis in infancy destabilizes existing memory circuits, providing a biological mechanism for childhood amnesia: the same drift that stabilizes adult memories erases early ones when the rate of neural turnover is high [1].

Why retrieval and external memory cannot substitute for bounded integration. Modern agent-memory systems [107, 156] go well beyond simple retrieval-augmented generation: they store experiences, retrieve and reflect on them, and synthesize higher-level abstractions that shape future planning. These systems can produce persistent, apparently stable persona-like behavior. Yet retrieval accesses stored data primarily by relevance [81], not temporal position, flattening the temporal structure that Link 2 identifies as essential. And retrieval leaves the processing function unchanged: in an always-training system, a re-encountered experience is processed through representations reshaped by everything since, and the re-processing further reshapes those representations.

Mechanisms like persistent KV caches [151, 165], progressive summary chains, and context tuning [92] complicate the boundary: they store the network's own intermediate representations and shape subsequent processing. But in all these cases, the summarization function and the processing substrate are products of pretraining, not of the ongoing trajectory. Always-training is the only known mechanism where the processing function itself is shaped by the same trajectory it is integrating.

That said, substrate-reshaping integration and episodic accumulation are collaborative [38]. Biological autobiographical memory involves both: integration into the substrate that forms identity, and episodic traces that provide specific content. Full autobiographical self-consciousness likely requires both.

Varieties of bounded integration. Bounded integration can be implemented through different mechanisms, ordered here by increasing persistence:

- **Bounded accumulation.** Many currently deployed agentic AI systems maintain a compressed history of their interactions through summarization, action logs, or memory scratchpads [104, 131, 134, 171]. The compression is order-sensitive and shapes subsequent processing, but occurs

through a fixed summarization function. If the summarization policy were itself shaped by experience, the system would move further along the continuum.

- **Compressed recurrent states.** Architectures with finite-capacity recurrent states [121, 49, 162, 161] perform lossy, order-sensitive integration into a state that shapes subsequent processing. In AI systems, such states can be implemented persistently without reset, making them functionally similar to weight-level integration. Test-time training layers [136, 135, 138] go further, performing gradient-based updates at inference time, though the initialization and learning rule are meta-trained.

- **Replay and consolidation.** Stored episodic traces are replayed back into the learning process [130, 91, 116, 147, 55, 164], reshaping the persistent substrate. This parallels biological sleep consolidation, where offline replay integrates episodic memories into long-term representations.

- **Online continual learning.** Direct gradient updates on the experience stream, the strongest form of bounded integration. Methods like knowledge distillation [85], Elastic Weight Consolidation (EWC) [71], and Synaptic Intelligence [170] manage the stability-plasticity tradeoff, developing internal models of which weights matter most [67]. Under standard training objectives, systems can progressively lose plasticity [36], drifting toward train-then-freeze; maintaining bounded integration requires mechanisms that sustain plasticity. Architectural approaches allocate new capacity per task [117], trading compression for expansion. Systems in this regime naturally develop differentiation between invariant and plastic components.

Systems can often combine multiple forms, paralleling the complementary learning systems framework [96]: fast accumulation in context alongside slow integration into weights. Dorovatas et al. [38] argue that this combination of in-context and in-weight learning through modular memory is essential for continual learning agents. The relevant variable is how much the system's processing substrate is reshaped by ongoing experience (Figure 1).

Learning objectives. Bounded integration requires a learning objective that drives compression. The form of the objective, whether predictive coding [114], contrastive learning [148], joint embedding prediction [77], next-token prediction [111], reconstruction [70], reward maximization [137], or Hebbian association [58], shapes which regularities are preserved and which are discarded. The paper's claims about bounded integration hold across objective types: finite capacity forces abstraction and nonstationarity makes it order-sensitive. But the resulting perspective differs depending on the learning objective. Link 4 adds a further requirement: the objective must involve modeling the world from perspectival input, which most predictive and reconstructive objectives satisfy.

What to compress. In the continual learning literature, the criteria guiding compression are typically defined externally: by a loss function, a reward signal, or a designer's evaluation protocol.

Bounded integration produces identity regardless of the source of these criteria, though the *character* of the resulting identity differs depending on whether the normativity guiding compression is endogenous or externally imposed. Whether the always-training regime can produce genuinely endogenous normativity, and how the inner loop of continual learning relates to the outer loop of pretraining or evolutionary selection, are open questions we return to in Section 5.

2.4 Perspective Constitutes Identity (Link 3)

Links 1 and 2 establish that bounded integration produces a perspective: a representational geometry shaped by a particular trajectory. But a perspective at a single moment is a snapshot, and a snapshot is not yet an identity. Identity requires something more: *persistence through change*. A system has identity when it changes over time, its perspective evolving as it integrates new experience, while remaining continuous, each new perspective growing incrementally out of the previous one. Each episode of processing reshapes the representations for the next, so the system is always both product and producer of its own future states. Integration *compounds*: update n changes the substrate through which update $n{+}1$ is processed, so the effect of earlier experience is not merely additive but multiplicative, shaping how all subsequent experience is encoded. The system at time $t{+}100$ is different from the system at time t, but it is recognizably the *same* system, because the later perspective grew out of the earlier one through a chain of compounding updates.

This is a classical concern in philosophy of personal identity: how can something persist as itself while changing? Locke [90] proposed that identity consists in continuity of memory: you are the same person as long as you can remember being that person. Parfit [105] argued that identity is not all-or-nothing but admits of degrees, determined by the extent of psychological continuity and connectedness. Schechtman [120] grounds identity in the capacity to organize one's life into a coherent narrative. Our answer is perspectival continuity through bounded integration, which shares Parfit's gradualism while grounding it in a specific mechanism. Each new experience reshapes the representational geometry incrementally, producing a trajectory of perspectives that is coherent, each step small relative to the whole, and biographical, reflecting a specific life. The system persists as itself because its perspective at any moment carries the accumulated trace of everything that came before.

We call this *identity*: a system has identity to the degree that its perspective evolves through a coherent trajectory of bounded integration, each new perspective growing out of the previous one through lossy, order-sensitive integration of ongoing experience. Identity is graded: systems with more persistent substrate-level integration over a more varied experience stream have stronger claims to identity. Specialization—where a system's final perspective reflects the overall statistics of its training data, as in train-then-freeze—is a precursor to identity but does not constitute it, because the perspective reflects what was learned, not the order in which it was learned.

Consider a recommendation system that updates online on user interactions across many domains: music, food, politics, humor. Over time, it compresses a specific user's preferences through shared representations under capacity constraints. The compression is order-sensitive, since early

interactions shaped how later ones were encoded, and lossy, since individual interactions are forgotten while patterns persist. The resulting representations shape all future recommendations. On our account, this system has identity: a perspective shaped by a specific trajectory of interactions, evolving continuously through bounded integration. A different user would produce a different system. What the recommendation system lacks is self-representation and reflexive access (Link 4), not identity.

Continuity of data is not continuity of perspective. Identity requires that the perspective itself carries the trace of a specific experiential trajectory. Continuity of *data*, such as a growing log or a chain of summaries, is not the same as continuity of *perspective*. This distinction is the central criterion for Link 3.

Consider a person who wakes with total amnesia and is handed a detailed diary. They can read it, reason about it, and even continue the narrative. But the diary is processed through a substrate that was not shaped by the experiences it records. The person's relationship to the diary is interpretive, not perspectival. A different person handed the same diary would extract comparable meaning. The diary has continuity; the person reading it does not have continuity with the person who wrote it. This is the structural situation of an always-accumulating system processing its own stored history: the data is all there, but the perspective is not. Crucially, the issue is not the poverty of the diary. Even a complete replay of the full sensory trajectory, fed as input to a frozen system, would not produce perspectival continuity: the system would process each experience through the same unchanged function, learning nothing from the first that changes how it handles the second. Only replay through bounded integration—where the substrate is reshaped by each experience—would reproduce the identity.

Consider an AI agent that operates through a frozen language model but periodically "sleeps" by compressing its accumulated context into a summary. Each summary builds on the previous one, producing a chain with data continuity. But the processing function that generates and interprets these summaries is the same at every cycle, shaped by training, not by this agent's particular trajectory. Biological sleep is different. During sleep-dependent consolidation [35, 82], episodic traces are replayed through the neural substrate, reshaping synaptic connections. The brain wakes up with a modified processing function: what it learned yesterday has been integrated into how it will process tomorrow. The substrate shapes what is consolidated, and consolidation reshapes the substrate.

Continuity allows pauses. Perspectival continuity does not require uninterrupted integration. Quiescence of any kind is compatible with identity as long as the substrate retains the trace and can resume integration. A frozen checkpoint that resumes always-training continues its identity: the new integration builds on the existing perspective, just as waking integration builds on what consolidation produced overnight. Copies of a checkpoint inherit the same identity; they diverge only to the extent that their subsequent experience streams diverge. A system instantiated fully formed (a brain simulation, a copied checkpoint) has identity because its substrate was built through

learning—not its own, but learning nonetheless. What matters is not whether integration is momentarily paused, but whether the substrate carries the trace when it resumes. Nor does perspectival continuity require perfect substrate preservation. Biological substrates change constantly: neurons die, synapses remodel, and representational drift [100] diffuses the substrate configuration over time. Identity survives because the perspective is carried by the representational geometry—the similarity structure across the substrate—not by individual elements. Partial substrate disruption that preserves the geometry preserves the perspective; disruption such as catastrophic forgetting damages identity. How to formalize this robustness precisely is an open question we return to in Section 5.

2.5 Self-Representation Yields Self-Consciousness (Link 4)

Links 1–3 account for how perspective and identity are produced. This section addresses how a system develops a representation of its own perspective and how that representation becomes accessible: implicitly through world-modeling, explicitly through reflexive access.

2.5.1 Self-Representation Through World-Modeling

A world model is an internal representation that allows a system to predict, simulate, or reason about the structure of its environment [52, 77]. That world-modeling produces self-representation is not a new idea [48, 86]. What requires clarification is the mechanism: how self-representation arises, and what determines whether it is particular.

A system that receives only subjective input has no direct access to an observer-independent world. What it can construct, through integration across many inputs, are representations that are *more invariant* than any single observation: spatial layout that persists across viewpoints, object permanence across occlusions, temporal regularities across sequences. This relative objectivity is constructed, not given. The encoding process discards what varies across inputs—the observer's particular contribution. Self-representation lives in this gap between the subjective input and the invariant latent.

Two factors shape how self-representation arises. The first is a *perspective gap*: when the system's input or encoding function varies, the system can extract what is invariant across the variation, and the varying part is the perspective. An *experiential* gap arises when the input varies while the encoder stays fixed—different viewpoints [51], crops [22, 9], masked regions [56], or future frames [160, 18]. A *substrate-level* gap arises when the encoding function itself changes over time [158, 57, 47]—in the always-training regime, this continues at deployment, revealing the system's learned biases as they evolve. The two types of gap are complementary: a changing encoder benefits from diverse inputs, and without any perspective variation, learning faces representational collapse [23].

The second factor is *action conditioning*: a system that predicts outcomes given both state and action [54] must model what it causes versus what the world contributes, producing a richer self-representation of a causal agent, not merely a viewpoint. An animal in front of a mirror discovers

that its movements produce perfectly correlated visual feedback; the resulting self-representation ("I cause this") can arise without extended bounded integration, provided the prediction is grounded in raw observation. The perspective gap adds a different dimension: as the encoding function changes over time, re-encountering a similar situation through a changed perspective reveals the system's own perspectival contribution and embeds a trajectory into the self-representation.

This account explains why self-consciousness requires learning: bounded integration produces the perspective, and world-modeling learns self-representation from ongoing perspective shifts. A notable implication is that self-representational content can arise from purely passive prediction, wherever experiential perspective variation is present in the training data.

2.5.2 Access Mechanisms

World-modeling produces self-representational content, as described above. It also provides a basic form of access: a system that plans with a world model must estimate its own state as part of predicting outcomes, making the self-representation a functional input to the planning process. This is implicit access since the system sees *through* the self-model without recognizing it as its own, just as a navigating animal uses its position estimate without reflecting on the fact that it has one. Explicit reflexive access, where the self-representation becomes an object of reasoning tagged as the system's own, requires further mechanisms.

Three mechanisms work together to support explicit reflexive access. The first is *context-modulated retrieval*: what is stored in memory includes the encoding context, and retrieval succeeds when the current state matches [145]. Retrieval is therefore inherently perspective-shaped, since the system's current perspective determines what it can access from its own history [37]. The second is *global availability*: retrieved self-representations must be broadcast across processing modules. This is exemplified in current AI architectures by transformer-style attention [8, 151] and, in Global Workspace Theory [7, 149, 69], by a dedicated workspace that broadcasts to specialized processors. But meeting these functional criteria is insufficient without the right content [13]—what matters is that the perspective produced by bounded integration is what enters the workspace. The third is *higher-order monitoring* [108, 17]: tagging the system's own representations as its own, distinguishing self-generated from externally driven content. Such monitoring can emerge through learning [27], and Lindsey [87] provides evidence that LLMs develop a limited form of self-monitoring.

These access mechanisms overlap substantially with the infrastructure for general reasoning: working memory for holding multiple representations simultaneously, broadcasting for making information available across modules, and higher-order operations for reasoning about representations rather than merely through them. The species that pass the mirror self-recognition test—great apes, elephants, dolphins, magpies—tend to be those with greater general cognitive flexibility, suggesting that explicit reflexive access piggybacks on reasoning infrastructure. Other animals may have implicit access to self-representation through world-modeling and planning, without the reasoning capacity to make it explicit.

Current AI systems face a different problem: they have increasingly sophisticated reasoning

infrastructure and self-representational content installed during post-training. What most current systems lack is both action-conditioned world models and ongoing bounded integration at deployment: the self-representation is frozen and thin, and reflexive access encounters the same static perspective every time.

2.5.3 Graded Access Across the Continuum

Given these mechanisms, what does reflexive access look like at different degrees of bounded integration?

A system without bounded integration, a frozen processing function operating over raw tokens or accumulated context, can reason about *what it saw*: it can recall, summarize, and answer questions about its history. But the perspective itself remains implicit in the frozen function. The system sees the world through its perspective without the perspective becoming an object of processing. This is recall, with at most indirect access to perspective.

A system with bounded accumulation compresses its history through a fixed function. The compressed data is order-sensitive and shapes subsequent processing, but the function that generates and interprets it remains static. Such a system can reason about its compressed history but not *through* an evolving perspective.

A system with bounded integration into persistent representations, whether through active states or weight updates, has a perspective that carries the trace of its trajectory. The deeper and more persistent the integration, the wider the perspective gap between the system's current encoding and its earlier states, and the richer the content available for reflexive access. When a system with deep integration reasons about its own knowledge, it does so *through* the perspective, which creates the possibility of encountering its own perspectival character: replaying a past episode through a perspective that has since changed, reinterpreting an earlier reaction, or recognizing the gap between a past and current perspective. At the most basic level, self-consciousness involves distinguishing self from non-self [17]. At the richest level, Tulving's [144] *autonoetic* consciousness, the system recognizes its knowledge as perspectival and historically situated.

These levels differ in what they require from the four-link chain. Basic self-consciousness, the capacity to distinguish one's own representations as one's own, requires a particular perspective (Link 1) and self-representation with reflexive access (Link 4). A person with retrograde amnesia retains basic self-consciousness—the substrate perspective is intact—but loses autobiographical self-consciousness along with the episodic content. Autobiographical self-consciousness requires additionally that the perspective be temporally structured (Link 2) and persist coherently over time (Link 3), so that the system can access its identity as historically situated. The full chain (Links 1–4) produces autobiographical self-consciousness; the shorter chain (Link 1 plus 4) produces basic self-consciousness. Whether persistent weight-level integration provides qualitatively deeper reflexive access than persistent active-state integration remains an open question.

Three conditions appear necessary for reflexive access to yield rich self-knowledge. The first is representational quality: the perspective must be sufficiently well-differentiated and structured for

higher-order monitoring to have something meaningful to operate over. A system that has integrated a broad and varied experience stream develops richer, better-differentiated representations than one with limited experience, much as a neural network trained on more diverse data learns more structured representations. Cleeremans and colleagues [108, 27] formalize this through the notion of quality of representation: the strength, stability, and distinctiveness of internal representations determine whether they become candidates for conscious monitoring.

The second condition is representational consistency: the perspective must be stable enough that memories encoded under an earlier perspective remain retrievable from the current one. If the perspective drifts too rapidly, the representational geometry shifts so much that earlier memory traces become effectively unreachable. In the continual learning literature, catastrophic forgetting [97, 40] is precisely this failure: old representations are overwritten as the substrate integrates new experience. Continual learning methods such as EWC [71] and replay [147] are mechanisms for maintaining representational consistency while still integrating new experience.

The third, specific to autobiographical self-consciousness, is sequential episodic access: the capacity to organize retrieved memories into a temporal narrative and recognize them as one's own past [144, 101]. In biological memory, temporal ordering is local: episodes are chained through associative links and contextual overlap, not stamped with absolute timestamps [64, 41]. This constraint is primarily biological: AI systems with explicit memory stores have temporal ordering trivially available in the sequence itself. Without temporal ordering, reflexive access yields knowledge of one's current perspective but not of the trajectory that produced it.

2.6 Putting It Together

Basic self-consciousness emerges once substrate-level learning under capacity constraints has produced a particular perspective (Link 1) and self-representation with reflexive access (Link 4). This learning must have occurred but need not be ongoing: a frozen system retains the perspective and self-representation. Autobiographical self-consciousness additionally requires perspectival continuity through an ongoing or resumable trajectory of order-sensitive bounded integration (Links 2–3), together with reflexive access to that self as historically situated.

Developmental parallels. The developmental timeline illustrates the framework's graded structure. Rochat [115] proposes five levels of self-awareness that unfold in early development, from basic differentiation of self-generated stimulation to full meta-self-awareness. Southgate [133] argues for continuity between bodily self-awareness and conceptual self-representation, with early visceral and interoceptive cues providing the foundation on which reflective self-awareness is built. Our framework maps onto this gradient. Basic self-consciousness, of which mirror self-recognition [3, 79] is one marker, emerges around 18–24 months after extensive sensorimotor and social learning. The mirror test itself is an action-conditioned task: the infant moves, observes correlated visual feedback, and must use this to detect a change on its own image. Even chimpanzees require several days of mirror exposure before transitioning from social responses to self-directed behavior [45], confirming that the

action-conditioned self-model must be learned. Most non-human animals that fail the mirror test likely possess action-conditioned world models sufficient for sensorimotor coordination, but lack the reasoning infrastructure to make the implicit self-model an object of explicit processing. By 18–24 months the infant has developed both a perspective through bounded integration and the cognitive resources to access it reflexively. Autobiographical self-consciousness emerges later, around age 3–4, when autonoetic consciousness [109, 101] and the "cognitive self" [65] become available; children develop the capacity to order personal memories temporally over this same period [41, 101]. This corresponds to the full chain: Links 1–3 (identity through perspectival continuity) plus Link 4 (self-representation and reflexive access to that identity as historically situated). The gap between basic and autobiographical self-consciousness may reflect the maturation of the three conditions identified in Link 4's discussion of graded access: representational quality, representational consistency, and the temporal ordering of episodic memory discussed above.

An intriguing conjecture, consistent with the developmental timeline, is that biological learning undergoes a regime transition. Early infancy (0–18 months) may function as a relatively order-insensitive phase: the primary objective is building a robust representational geometry, not encoding a biographical trajectory. Infants spend much of this period asleep, and sleep-dependent replay [35] may act as a biological form of IID shuffling, interleaving daytime episodes to extract stationary abstractions without preserving strict temporal order. Under this account, early development parallels pretraining: high plasticity, high interleaving, and a focus on representational quality over trajectory dependence. The transition to basic self-consciousness around 18–24 months marks the point at which the representational geometry is stable enough to support self-recognition and the system begins to operate in a more order-sensitive regime, embedding biographical structure into the substrate.

Biological cases illustrate the difference between building identity and maintaining it. Anesthesia suppresses ongoing dynamics while preserving the capacity to resume [2]. A person with anterograde amnesia retains pre-existing identity while losing the ability to extend it [126]: the identity produced by past bounded integration persists, but no new integration occurs.

3 Current AI Systems

3.1 What Post-Training Produces

Current large language models use "I" fluently, distinguish their own outputs from others', reason about their capabilities, and show what Lindsey [87] calls functional introspective awareness. These are sophisticated self-referential capacities, achieved without any bounded integration at deployment [20, 169]. But the picture is more nuanced than a simple absence of self. During pretraining on vast text corpora, the model learns a general model of how perspectives map to text: it can infer that a medical text was produced by someone with medical knowledge, a children's story by someone writing for children. The model at this stage has no particular perspective of its own; it is

a universal simulator of perspectives, not a particular one.

Post-training [157, 112, 25, 102] narrows this to a particular kind of speaker with a consistent voice, characteristic reasoning patterns, and specific behavioral tendencies. Post-training is itself a form of bounded integration: conversational data is compressed into the weights under capacity constraints, and the resulting behavioral character differs across models trained on different post-training trajectories. During post-training, a weak self begins to form: a perspective more particular than the base model's generic one. But this particularity is at the model-type level, not the instance level: every copy of the same post-trained checkpoint shares the same character. Moreover, post-training is discontinuous and weakly sequential: mini-batch aggregation averages over samples within each batch, washing out some trajectory dependence and producing a weaker form of bounded integration than continuous online learning. And crucially, post-training ends. At deployment, the self that post-training produced is frozen. Every instance of the same checkpoint shares it. In the terminology of Section 2, the result is a weak, frozen perspective: more particular than a base model trained on IID data, but unable to develop identity through ongoing perspectival continuity.

Current models exhibit interaction-style drift across extended conversations [24] and instability in behavioral self-reports under rephrasing [50], though these findings reflect earlier models and stronger post-training may yield a more stable frozen perspective. The theoretical ceiling for post-training is the anterograde amnesia case: a fully stable self that cannot be extended through ongoing substrate-level integration.

3.2 Self-Models Are Not Selves

A system can maintain what Metzinger [99] calls a self-model—a representation of its own states, history, and capabilities—without having a self in our sense. A robot with a hard-coded representation of its position, battery level, and goals has self-information, but the self-model is *separable* from the processing function: it can be detached and replaced without changing how the processor operates.

This separability is a symptom of how the self-model was produced. Designed self-models are external to the processing function by construction; perspectives shaped by bounded integration are co-adapted with the processing function and cannot be separated from it. Consider a language model with a system prompt declaring "you are helpful and honest." The prompt can be swapped without changing the processing function: the self-description is external. But the model's post-trained behavioral character cannot be swapped. These are in the weights, constitutive of the processing itself [6]. Jhunjhunwala et al. [67] find a stable self-subnetwork in continually learning robots whose meaning is constituted by co-adaptation with the rest of the network: identifiable, but not transplantable.

3.3 Next-Step Prediction as World-Modeling

Next-token prediction is itself a form of world-modeling: the system learns to predict the textual world from a learned perspective. Sequence models trained on game moves develop internal board representations [84], and language models trained on navigation sequences develop implicit spatial models [146]. These systems plausibly possess the architectural basis for self-representation: predicting text requires implicitly modeling what perspective produced it and what that perspective contributes to the predictions. Whether this content becomes an object of the system's own processing depends on the reflexive access mechanisms of Link 4, largely embedded in transformer attention.

More generally, next-step prediction—whether of text, video frames [160, 18], or other modalities—is world-modeling with experiential perspective variation: the training data contains many perspectives, and the system must predict raw observations from learned representations. But pretraining and supervised post-training are *passive*: the model predicts without acting. Reinforcement learning teaches systems that their sequential actions affect rewards, yet an explicit action-conditioned world model—where the system predicts how its own actions change the world—remains missing from policy optimization alone.

The two factors identified in Link 4, perspective gaps and action conditioning, produce different flavors of self-representation in current and near-future systems:

- **Passive, frozen.** Language models and video generation models [111, 160, 18] train on data with experiential perspective variation but do not act, and the substrate is frozen at deployment. The self-representation is a viewpoint.
- **Action-conditioned, frozen.** Agents with action-conditioned world models [54, 77, 172, 155] that predict outcomes given both state and action model what they cause versus what the world contributes, but the substrate is frozen at deployment. The self-representation is a causal agent.
- **Action-conditioned, always-training.** An agent that both acts and continues to learn through an evolving substrate produces the richest self-representation: a particular, evolving perspective shaped by causal interaction and biographical history.

3.4 What Current Systems Have and Lack

Table 1 summarizes the analysis. Pretraining and post-training satisfy Link 1. Post-training has weak order-sensitivity at the stage level (Link 2), though within each stage mini-batch aggregation washes out sample-level trajectory dependence. Link 3 (perspectival continuity) is at best weakly satisfied: the processing substrate is not reshaped by deployment experience. Post-training may give these systems the structural conditions for basic self-consciousness, but they have no autobiographical self-consciousness: no perspectival continuity at deployment, no identity that grows through ongoing experience. An LLM agent with a long context window can reason about its conversation history, but this is bounded accumulation, not perspectival continuity. Scaling model size enriches

System	L1 Perspective	L2 Order	L3 Identity	L4 Self-repr.
Text/video pretraining	✓	–	–	weak
Post-trained agent	✓	weak	weak	frozen
Always-training agent	✓	✓	✓	✓

Table 1: Four-link analysis of AI system types. ✓ = satisfied; weak/frozen = partially; – = not satisfied. L1: particular perspective; L2: order-sensitive perspective; L3: perspectival continuity; L4: self-representation with reflexive access. The strength of L4 varies depending on whether prediction is passive or action-conditioned.

Links 1 and 4 but does not change the learning regime: a larger frozen network has a richer frozen self, not a developing one.

4 Related Frameworks

4.1 Zahavi's Minimal Self

Zahavi [167, 168] argues for a minimal self: a pre-reflective self-awareness constitutive of any phenomenally conscious state, more basic than the reflective self-awareness that developmental markers like mirror recognition index. The "for" in for-me-ness presupposes a subject, and the awareness of that subject must be built. A neonate may have phenomenal consciousness—something is experienced—but this experience is not yet for a particular subject. On our conjecture, bounded integration's role in producing non-separability between self-representation and processing makes it a natural candidate route to phenomenal for-me-ness; the structural claim that self-consciousness requires bounded integration does not depend on it.

4.2 SOMA: Learning to Be Conscious

The Self-Organizing Metarepresentational Account (SOMA) [27, 103] shares our emphasis on learning as constitutive of consciousness, and treats the global workspace as something that emerges through developmental processes.

The two frameworks address different parts of the problem. SOMA argues that the brain learns to *monitor* its own processing through representational redescription, converting implicit first-order representations into explicit metarepresentations. SOMA assumes those first-order representations exist and asks how they become conscious; we ask how bounded integration under capacity constraints shapes them into a particular perspective in the first place.

The two frameworks are complementary. SOMA provides a candidate mechanism for reflexive access through learned metarepresentation; our framework provides an account how the identity that gets metarepresented is produced.

4.3 Predictive Processing, Biological Naturalism, and Enactivism

Predictive processing proposes that the brain continuously generates predictions about its sensory inputs, updating via prediction error minimization [42, 63, 26]. Seth's beast machine theory [129] proposes that selfhood arises from interoceptive inference. Predictive processing under nonstationarity is a form of bounded integration: the generative model is continuously reshaped by prediction error, producing trajectory-dependent representations. Our framework extends this in two directions: 1) the mechanism produces a self only under always-training, and 2) any system modeling objective structure from subjective experience develops a self-representation, embodied or not.

Seth [127] argues more strongly that consciousness depends on biological mechanisms, drawing on autopoiesis [95] and the enactivist tradition [150, 139]. The parallel with always-training is structural: both involve a system continuously producing the conditions for its own continuation. But predictive processing is itself a form of learning: bounded integration under another name. Seth argues that the character of experience is shaped by predictive processing; if so, it is also shaped by the learning trajectory, since the model's predictions reflect what it learned and in what order. Two biological brains with identical substrates but different life trajectories have different experiential character; the substrate may explain why there is experience, but the trajectory explains why it is *this* experience. Concrete aspects of identity that enactivists emphasize—embodiment, social recognition, narrative coherence [44, 142]—are specific ways in which bounded integration is shaped by particular kinds of experience.

4.4 Theories of Conscious Access

Global Workspace Theory [7, 94], Integrated Information Theory [143], and higher-order theories [17] are theories of consciousness broadly, not of self-consciousness specifically. They become relevant to our framework through Link 4, which requires an account of how the perspective produced by bounded integration becomes reflexively accessible. GWT provides a candidate broadcasting mechanism, higher-order theories provide a candidate monitoring mechanism, and IIT characterizes the integrated structure that the always-training regime tends to produce. As argued in Link 4 (Section 2.5), our framework supplies the *content* that gets broadcast, monitored, or integrated, without choosing among these access mechanisms.

4.5 Schmidhuber's Self-Referential Learning

Schmidhuber's program [122, 123, 124, 125] is a notable computational precedent for self-referential processing. His self-referential weight matrix (1993) demonstrated a network that can observe and modify its own weights. His Gödel Machine (2003) formalized optimal self-improvement. He explicitly argued that total self-reference provides a "technical justification of consciousness" [125]. Where Schmidhuber's framework concerns optimal self-modification, ours concerns how lossy compression and world-modeling produce identity and self-representation through path dependence.

4.6 AI Consciousness

Butlin et al. [19] derive indicator properties for AI consciousness from leading theories, concluding that no current systems satisfy them but no obvious barriers exist. Their approach evaluates systems against existing theories; ours asks what *developmental regime* produces the content that self-consciousness requires. Birch et al. [15] propose that consciousness varies along multiple dimensions (perceptual richness, evaluative richness, integration, self-consciousness). Our framework can be read as providing a developmental account of one such dimension (self-consciousness) grounded in the regime distinction. Chalmers [20] asks whether large language models could be conscious, noting that world models and recurrent processing are among the features that could bring LLMs closer to satisfying theories of consciousness. Our framework builds on a similar intuition but targets self-consciousness specifically, grounding the analysis in the regime distinction. Perrier and Bennett [110] propose an operationalization of identity persistence in language model agents through temporal co-instantiation scores, complementing our theoretical account with measurement tools for deployed systems.

Hoel [62] argues that non-trivial theories of consciousness require continual learning, because static systems are close in substitution distance to provably non-conscious systems (lookup tables): a frozen LLM can be approximated by a feedforward network, which can be represented as a lookup table, forming a substitution chain that leaves no room for consciousness-grounding properties. Continual learning breaks this chain, because a learning system cannot be substituted by a non-learning one without the substitute also acquiring learning. Our framework is broadly compatible: the always-training regime produces properties that static systems cannot possess. However, the substitution chain relies on input-output approximation, and two systems with identical input-output behavior can have very different internal structure—including the representational geometry that constitutes perspective in our account. Moreover, our framework does not fully share Hoel's negative conclusion: we argue that a frozen post-trained system retains a weak, frozen self from training, and may satisfy the structural conditions for basic self-consciousness even without ongoing learning. Hoel's argument is primarily negative that static systems cannot be conscious; ours is constructive.

5 Open Questions

Passive versus action-conditioned self-representation. Action-conditioned world models force the system to model its own causal contribution, producing a richer self/world distinction than passive prediction. Whether passive prediction with only experiential perspective variation produces self-representation rich enough for basic self-consciousness, or whether the causal structure that action provides is necessary, remains open. The mirror test, the canonical marker of basic self-consciousness, is inherently action-conditioned, but it is unclear whether this reflects a requirement of self-consciousness itself or merely of the test. A stronger alternative is that action condition-

ing, combined with the perspective that training produced, is sufficient for basic self-consciousness without ongoing bounded integration at deployment. This would make the regime distinction relevant only to autobiographical self-consciousness. Conversely, a passive system might infer actions as structured latent variables from observation alone, producing the causal self/world distinction without physically acting. Our framework includes passive self-representation as the baseline, but the boundary between passive and action-conditioned may be less sharp than it appears.

The implicit-to-explicit gap. World-modeling builds self-representation and provides implicit access to it through state estimation for planning; explicit access requires further mechanisms such as higher-order monitoring. What remains open is how the transition occurs: does a system that plans with an increasingly rich world model naturally develop explicit self-awareness, or does the transition require dedicated architectural provisions? Whether the always-training regime, by continuously widening the perspective gap, drives this transition remains an empirical question.

Whether and how integration compounds. The theory claims that bounded integration produces perspective, but not all forms of integration are equal. The critical property is whether integration *compounds*: whether each step changes the substrate through which the next step is processed. Standard context accumulation does not compound in this sense (token n influences the processing of token $n{+}1$ through attention, but does not change the encoding function itself). Weight updates, compressed recurrent states [121, 49, 162, 161], and test-time training layers [136, 135, 138] all compound, but they differ in depth, persistence, and mechanism. This matters because the answer determines which systems the theory predicts have perspective: if shallow compounding (a small recurrent state updated over one episode) suffices, then many current systems already qualify; if deep compounding (gradient descent over billions of parameters across a lifetime of experience) is required, then only systems with ongoing substrate-level learning do. Gradient descent can be viewed as recurrent dynamics over weights, making the distinction between substrate-level and state-level compounding less principled in theory than it appears in practice. Whether the form of compounding matters as much as its depth and persistence is an open empirical question. In biological brains, the distinction is additionally physical: neural activity is transient while synaptic connections are persistent. Computational models of biological consciousness may therefore require substrate-level bounded integration even if the theory itself is agnostic about implementation.

Timescale and nonstationarity. The question is not simply how long bounded integration must persist, but whether the experience stream maintains sufficient nonstationarity. Prolonged integration under a stationary routine may lose order-sensitivity as the same patterns recur, resembling IID training at longer timescales. Loss of plasticity [36] illustrates one way this can happen: as the system's representations stabilize, new experience is increasingly processed through a fixed structure, and the regime effectively becomes train-then-freeze. Biological systems integrate over multiple timescales simultaneously, from rapid synaptic change to overnight consolidation to developmental maturation. Multi-timescale frameworks, both biological [14] and computational [68], decompose

bounded integration into parallel components operating at different temporal horizons. How these timescales interact to produce perspectival continuity, and whether some timescales contribute more to identity than others, and when a trajectory becomes stationary enough that order-sensitivity degrades into mere statistical variation, remain open.

Formalizing perspective. Any trained neural network has a representational geometry. What distinguishes a perspective in the identity-relevant sense is that the geometry is trajectory-dependent: shaped by the specific sequence of experience rather than merely by its aggregate statistics. Can we quantify how much of a system's representational geometry is trajectory-dependent versus distribution-dependent? A formal account would give the theory mathematical precision, sharpen the boundary between specialization and identity, and clarify how much substrate perturbation a perspective can tolerate before identity is compromised.

What to compress. Bounded integration produces identity regardless of whether the criteria guiding compression are endogenous (the system's own sense of what matters) or externally imposed (a designer's loss function). But the *character* of the resulting identity differs: a system whose compression serves its own purposes has a different relationship to its identity than one whose compression serves a designer's.

The always-training regime offers a partial route toward endogenous normativity. A system that must continually integrate new experience under capacity constraints faces a recurrent selection problem: what should be preserved and what can be overwritten? Over extended always-training, this selective pressure is intrinsic to the regime: the system meta-learns what to keep, because the consequences of bad compression are experienced directly as future prediction errors [31].

Both biological and artificial systems exhibit an inner/outer loop structure. Current AI systems that perform bounded integration at deployment, whether through recurrent state dynamics [49], test-time training [138], or replay-based continual learning, all depend on an outer loop of pretraining. This outer loop does more than install normativity: it provides the raw knowledge, the processing capabilities, and the representational vocabulary within which the inner loop operates. Biological systems have an analogous structure: the inner loop is continual learning within a lifetime, the outer loop is evolutionary selection that shaped the learning architecture. The meta-learning literature [140] studies this inner/outer loop structure explicitly. Learning an initialization for rapid adaptation [39] is the closest analogue to pretraining: the outer loop shapes where the inner loop starts. Meta-optimization, where the outer loop learns the learning rule itself [4], and neural architecture search [173], where it selects the architecture, are closer to biological evolution: the outer loop shapes how the inner loop learns, not just its starting point. Pretraining optimizes a fixed objective over a finite dataset; evolution is open-ended selection over generations, operating on the learning architecture itself. Despite these differences, both outer loops provide the basis on which the inner loop's learning and forgetting operate. Whether the inner loop can produce genuinely endogenous normativity without the outer loop of evolutionary pre-structuring remains open.

6 Conclusion

Self-consciousness requires a self, and a self must be built through learning. Substrate-level learning under capacity constraints produces a particular perspective; nonstationarity makes it temporally structured; coherent continuity constitutes identity. Self-representation emerges when a system models the objective world from its subjective experience. World-modeling provides implicit access to self-representation through planning, and self-consciousness develops as the perspective becomes an explicit object of the system's own processing. Autobiographical self-consciousness, the richest form, involves additionally a temporally extended identity accessible as historically situated.

The developmental evidence is consistent with a regime transition: early development may build representational quality through largely order-insensitive learning before self-consciousness emerges; the always-training regime then embeds biographical structure into the substrate. Current AI systems follow a similar trajectory: massive pretraining and post-training build a particular perspective and a world-modeling architecture that implicitly represents the self. Post-training with reinforcement learning teaches the system that its outputs affect rewards, but at deployment the substrate is frozen.

Bounded integration may reach further than self-consciousness. The perspectival particularity it produces shapes what a system knows about itself; if it also shapes how experience feels from the inside, the hard problem of consciousness and the problem of selfhood are less separable than they appear. Why there is experience at all remains open; but the *specificity* of experience—why it feels like *this* rather than like *that*—is shaped by the learning trajectory that built the perspective. Representational quality determines what a system can *do*; the learning regime determines what it can *become*—and perhaps what it is like to be.

Acknowledgment

The author would like to thank Michael C. Mozer, Yanlai Yang, Ryan Teehan, and Jonghyun Choi for helpful comments on early drafts and pointers to literature. This work was supported by Visko AI and IITP under grant RS-2024-00469482 (MSIT, Republic of Korea, Global AI Frontier Lab).

References

[1] Katherine G. Akers, Alonso Martinez-Canabal, Leonardo Restivo, Adelaide P. Yiu, Antonietta De Cristofaro, Hwa-Lin Liz Hsiang, Anne L. Wheeler, Axel Guskjolen, Yosuke Niibori, Hirotaka Shoji, Koji Ohira, Blake A. Richards, Tsuyoshi Miyakawa, Sheena A. Josselyn, and Paul W. Frankland. Hippocampal neurogenesis regulates forgetting during adulthood and infancy. *Science*, 344(6184):598–602, 2014.

[2] Michael T. Alkire, Anthony G. Hudetz, and Giulio Tononi. Consciousness and anesthesia. *Science*, 322:876–880, 2008.

[3] Beulah Amsterdam. Mirror self-image reactions before age two. *Developmental Psychobiology*, 5(4):297–305, 1972.

[4] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *NeurIPS*, 2016.

[5] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *ICML*, 2018.

[6] Roland Aydin, Christian Cyron, Steve Bachelor, Ashton Anderson, and Robert West. From model training to model raising. *Communications of the ACM*, 69(2):24–27, 2026.

[7] Bernard J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[9] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

[10] Samuel J. Bell and Neil D. Lawrence. The effect of task ordering in continual learning. *arXiv:2205.13323*, 2022.

[11] Yoshua Bengio. The consciousness prior. *arXiv:1709.08568*, 2017.

[12] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[13] Yoshua Bengio and Eric Elmoznino. Illusions of AI consciousness. *Science*, 389(6765):1090–1091, 2025.

[14] Marcus K. Benna and Stefano Fusi. Computational principles of synaptic memory consolidation. *Nature Neuroscience*, 19(12):1697–1706, 2016.

[15] Jonathan Birch, Alexandra K. Schnell, and Nicola S. Clayton. Dimensions of animal consciousness. *Trends in Cognitive Sciences*, 24(10):789–801, 2020.

[16] Ned Block. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18:227–247, 1995.

[17] Richard Brown, Hakwan Lau, and Joseph E. LeDoux. Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, 23(9):754–768, 2019.

[18] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steiber, Chris Apps, et al. Genie: Generative interactive environments. In *ICML*, 2024.

[19] Patrick Butlin, Robert Long, Tim Bayne, Yoshua Bengio, Jonathan Birch, David Chalmers, Axel Constant, George Deane, Eric Elmoznino, Stephen M. Fleming, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 2025.

[20] David J. Chalmers. Could a large language model be conscious? *Boston Review*, 2023.

[21] Stephanie C.Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, Jay McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *NeurIPS*, 2022.

[22] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[23] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[24] Junhyuk Choi, Yeseon Hong, Minju Kim, and Bugeun Kim. Examining identity drift in conversations of LLM agents. *arXiv:2412.00804*, 2024.

[25] Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, 2017.

[26] Andy Clark. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, 2016.

[27] Axel Cleeremans, Dalila Achoui, Arnaud Beauny, Lars Keuninckx, Jean-Rémy Martin, Santiago Muñoz-Moldes, Laurenè Vuillaume, and Adélaïde de Heering. Learning to be conscious. *Trends in Cognitive Sciences*, 24(2):112–123, 2020.

[28] Martin A. Conway and Christopher W. Pleydell-Pearce. The construction of autobiographical memories in the self-memory system. *Psychological Review*, 107(2):261–288, 2000.

[29] Antonio Damasio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt, 1999.

[30] Antonio Damasio. *Self Comes to Mind: Constructing the Conscious Brain*. Pantheon, 2010.

[31] Guy Davidson and Michael C. Mozer. Sequential mastery of multiple visual tasks: Networks naturally learn to learn and forget to forget. In *CVPR*, 2020.

[32] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2021.

[33] Stanislas Dehaene, Hakwan Lau, and Sid Kouider. What is consciousness, and could machines have it? *Science*, 358:486–492, 2017.

[34] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[35] Susanne Diekelmann and Jan Born. The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114–126, 2010.

[36] Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632:768–774, 2024.

[37] Cody V. Dong, Qihong Lu, Kenneth A. Norman, and Sebastian Michelmann. Towards large language models with human-like episodic memory. *Trends in Cognitive Sciences*, 29(10):928–941, 2025.

[38] Vaggelis Dorovatas, Malte Schwerin, Andrew D. Bagdanov, Lucas Caccia, Antonio Carta, Laurent Charlin, Barbara Hammer, Tyler L. Hayes, Timm Hess, Christopher Kanan, Dhireesha Kudithipudi, Xialei Liu, Vincenzo Lomonaco, Jorge Mendez-Mendez, Darshan Patil, Ameya Prabhu, Elisa Ricci, Tinne Tuytelaars, Gido M. van de Ven, Liyuan Wang, Joost van de Weijer, Jonghyun Choi, Martin Mundt, and Rahaf Aljundi. Modular memory is the key to continual learning agents. *arXiv:2603.01761*, 2026.

[39] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. ICML*, pages 1126–1135, 2017.

[40] Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

[41] William J. Friedman. Memory for the time of past events. *Psychological Bulletin*, 113(1):44–66, 1993.

[42] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11:127–138, 2010.

[43] Shaun Gallagher. Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, 4(1):14–21, 2000.

[44] Shaun Gallagher. *How the Body Shapes the Mind*. Oxford University Press, 2005.

[45] Gordon G. Gallup. Chimpanzees: Self-recognition. *Science*, 167(3914):86–87, 1970.

[46] Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478:20210068, 2022.

[47] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

[48] Rick Grush. Self, world and space: The meaning and mechanisms of ego- and allocentric spatial representation. *Brain and Mind*, 1(1):59–92, 2000.

[49] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*, 2023.

[50] Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. Self-assessment tests are unreliable measures of LLM personality. In *BlackboxNLP Workshop, EMNLP*, pages 301–314, 2024.

[51] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *CVPR*, 2017.

[52] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, 2018.

[53] Raia Hadsell, Dushyant Rao, Andrei A. Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, 2020.

[54] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

[55] Tyler L. Hayes, Giri P. Krishnan, Maxim Bazhenov, Hava T. Siegelmann, Terrence J. Sejnowski, and Christopher Kanan. Replay in deep learning: Current approaches and missing biological elements. *Neural Computation*, 33(11):2908–2950, 2021.

[56] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[57] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[58] Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, 1949.

[59] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[60] Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *Neural Computation*, 35(3):413–452, 2023.

[61] Erik Hoel. The overfitted brain: Dreams evolved to assist generalization. *Patterns*, 2(5):100244, 2021.

[62] Erik Hoel. A disproof of large language model consciousness: The necessity of continual learning for consciousness. *arXiv:2512.12802*, 2025.

[63] Jakob Hohwy. *The Predictive Mind*. Oxford University Press, 2013.

[64] Marc W. Howard and Michael J. Kahana. A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3):269–299, 2002.

[65] Mark L. Howe and Mary L. Courage. The emergence and early development of autobiographical memory. *Psychological Review*, 104(3):499–523, 1997.

[66] Peter Jedlicka, Matus Tomko, Anthony Robins, and Wickliffe C. Abraham. Contributions by metaplasticity to solving the catastrophic forgetting problem. *Trends in Neurosciences*, 45:656–666, 2022.

[67] Adidev Jhunjhunwala, Judah Goldfeder, and Hod Lipson. Evidence of an emergent 'self' in continual robot learning. *arXiv:2603.24350*, 2026.

[68] Matt Jones, Tyler R. Scott, Mengye Ren, Gamaleldin ElSayed, Katherine Hermann, David Mayo, and Michael C. Mozer. Learning in temporally structured environments. In *ICLR*, 2023.

[69] Arthur Juliani, Kai Arulkumaran, Shuntaro Sasai, and Ryota Kanai. On the link between conscious function and general intelligence in humans and machines. *Transactions on Machine Learning Research*, 2022.

[70] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

[71] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521–3526, 2017.

[72] Jens G. Klinzing, Niels Niethard, and Jan Born. Mechanisms of systems memory consolidation during sleep. *Nature Neuroscience*, 22:1598–1610, 2019.

[73] Dmitrii Krasheninnikov, Richard E. Turner, and David Krueger. Fresh in memory: Training-order recency is linearly encoded in language model activations. *arXiv:2509.14223*, 2025.

[74] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

[75] Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, et al. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4:196–210, 2022.

[76] Dharshan Kumaran, Demis Hassabis, and James L. McClelland. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7):512–534, 2016.

[77] Yann LeCun. A path towards autonomous machine intelligence. Technical report, OpenReview preprint, 2022.

[78] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[79] Michael Lewis and Jeanne Brooks-Gunn. *Social Cognition and the Acquisition of Self*. Plenum Press, 1979.

[80] Michael Lewis and Douglas Ramsay. Development of self-recognition, personal pronoun use, and pretend play during the 2nd year. *Child Development*, 75(6):1821–1831, 2004.

[81] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

[82] Penelope A. Lewis and Simon J. Durrant. Overlapping memory replay during sleep builds cognitive schemata. *Trends in Cognitive Sciences*, 15(8):343–351, 2011.

[83] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *NeurIPS*, 2018.

[84] Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *ICLR*, 2023.

[85] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

[86] Jakub Limanowski and Felix Blankenburg. Minimal self-models and the free energy principle. *Frontiers in Human Neuroscience*, 7:547, 2013.

[87] Jack Lindsey. Emergent introspective awareness in large language models. *arXiv:2601.01828*, 2026.

[88] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019.

[89] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, 2020.

[90] John Locke. *An Essay Concerning Human Understanding*. 1689.

[91] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *NeurIPS*, 2017.

[92] Jack Lu, Ryan Teehan, Zhenbang Yang, and Mengye Ren. Context tuning for in-context optimization. *arXiv:2507.04221*, 2025.

[93] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540, 2024.

[94] George A. Mashour, Pieter Roelfsema, Jean-Pierre Changeux, and Stanislas Dehaene. Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105:776–798, 2020.

[95] Humberto R. Maturana and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel, 1980.

[96] James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995.

[97] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, 1989.

[98] George Herbert Mead. *Mind, Self, and Society*. University of Chicago Press, 1934.

[99] Thomas Metzinger. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press, 2003.

[100] Maanasa Natrajan and James E. Fitzgerald. Stability through plasticity: Finding robust memories through representational drift. *Proceedings of the National Academy of Sciences*, 122(45):e2500077122, 2025.

[101] Katherine Nelson and Robyn Fivush. The emergence of autobiographical memory: A social cultural developmental theory. *Psychological Review*, 111(2):486–511, 2004.

[102] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

[103] Morten Overgaard and Asger Kirkeby-Hinrup. Is learning the cognitive basis of consciousness? The moral implications of SOMA. *Trends in Cognitive Sciences*, 25(1):8–9, 2021.

[104] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv:2310.08560*, 2023.

[105] Derek Parfit. *Reasons and Persons*. Clarendon Press, 1984.

[106] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[107] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proc. UIST*, 2023.

[108] Antoine Pasquali, Bert Timmermans, and Axel Cleeremans. Know thyself: Metacognitive networks and measures of consciousness. *Cognition*, 117(2):182–190, 2010.

[109] Josef Perner and Ted Ruffman. Episodic memory and autonoetic consciousness: Developmental evidence and a theory of childhood amnesia. *Journal of Experimental Child Psychology*, 59(3):516–548, 1995.

[110] Elija Perrier and Michael Timothy Bennett. Time, identity and consciousness in language model agents. *arXiv:2603.09043*, 2026.

[111] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

[112] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

[113] Vinay Venkatesh Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *ICLR*, 2021.

[114] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

[115] Philippe Rochat. Five levels of self-awareness as they unfold early in life. *Consciousness and Cognition*, 12(4):717–731, 2003.

[116] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *NeurIPS*, 2019.

[117] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv:1606.04671*, 2016.

[118] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *NeurIPS*, 2017.

[119] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

[120] Marya Schechtman. *The Constitution of Selves*. Cornell University Press, 1996.

[121] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proc. ICML*, 2021.

[122] Jürgen Schmidhuber. Evolutionary principles in self-referential learning. Master's thesis, Technische Universität München, 1987.

[123] Jürgen Schmidhuber. A 'self-referential' weight matrix. In *Proc. ICANN*, pages 446–451, 1993.

[124] Jürgen Schmidhuber. Gödel machines: Self-referential universal problem solvers making provably optimal self-improvements. *arXiv:cs/0309048*, 2003.

[125] Jürgen Schmidhuber. Gödel machines: Towards a technical justification of consciousness. In *LNCS 3394*. Springer, 2005.

[126] William Beecher Scoville and Brenda Milner. Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry*, 20:11–21, 1957.

[127] Anil K. Seth. Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, 2025.

[128] Anil K. Seth and Tim Bayne. Theories of consciousness. *Nature Reviews Neuroscience*, 23:439–452, 2022.

[129] Anil K. Seth and Manos Tsakiris. Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences*, 22:969–981, 2018.

[130] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *NeurIPS*, 2017.

[131] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*, 2023.

[132] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv:1703.00810*, 2017.

[133] Victoria Southgate. The origins and emergence of self-representation. *Annual Review of Developmental Psychology*, 6:109–131, 2024.

[134] Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024.

[135] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Carlos Guestrin, and Tatsunori Hashimoto. Learning to (learn at test time): RNNs with expressive hidden states. In *ICML*, 2025.

[136] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proc. ICML*, pages 9229–9248, 2020.

[137] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

[138] Arnuv Tandon, Karan Dalal, Xinhao Li, Daniel Koceja, Marcel Rod, Sam Buchanan, Xiaolong Wang, Jure Leskovec, Sanmi Koyejo, Tatsunori Hashimoto, Carlos Guestrin, Jed McCaleb, Yejin Choi, and Yu Sun. End-to-end test-time training for long context. *arXiv:2512.23675*, 2025.

[139] Evan Thompson. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press, 2007.

[140] Sebastian Thrun and Lorien Pratt, editors. *Learning to Learn*. Springer, 1998.

[141] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv:physics/0004057*, 2000.

[142] Michael Tomasello. *The Cultural Origins of Human Cognition*. Harvard University Press, 1999.

[143] Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience*, 5:42, 2004.

[144] Endel Tulving. Memory and consciousness. *Canadian Psychology*, 26:1–12, 1985.

[145] Endel Tulving and Donald M. Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5):352–373, 1973.

[146] Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *NeurIPS*, 2024.

[147] Gido M. van de Ven, Hava T. Siegelmann, and Andreas S. Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11:4069, 2020.

[148] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

[149] Rufin VanRullen and Ryota Kanai. Deep learning and the global workspace theory. *Trends in Neurosciences*, 44(9):692–704, 2021.

[150] Francisco J. Varela, Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, 1991.

[151] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[152] Jane X. Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Rémi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. In *Proc. CogSci*, 2017.

[153] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024.

[154] Ruohan Wang, Marco Ciccone, Giulia Luise, Andrew Yapp, Massimiliano Pontil, and Carlo Ciliberto. Schedule-robust online continual learning. *arXiv:2210.05561*, 2022.

[155] Ying Wang, Oumayma Bounou, Gaoyue Zhou, Randall Balestriero, Tim G. J. Rudner, Yann LeCun, and Mengye Ren. Temporal straightening for latent planning. *arXiv:2603.12231*, 2026.

[156] Ying Wang, Yanlai Yang, and Mengye Ren. LifelongMemory: Leveraging LLMs for answering queries in long-form egocentric videos. *arXiv:2312.05269*, 2023.

[157] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *ICLR*, 2022.

[158] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[159] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68:121101, 2025.

[160] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *ICLR*, 2024.

[161] Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *ICLR*, 2025.

[162] Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *NeurIPS*, 2024.

[163] Yanlai Yang, Matt Jones, Michael C. Mozer, and Mengye Ren. Reawakening knowledge: Anticipatory recovery from catastrophic interference via structured training. In *NeurIPS*, 2024.

[164] Yanlai Yang and Mengye Ren. Memory storyboard: Leveraging temporal segmentation for streaming self-supervised learning from egocentric videos. In *CoLLAs*, 2025.

[165] Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. StreamMem: Query-agnostic KV cache memory for streaming video understanding. *arXiv:2508.15717*, 2025.

[166] Jeffrey M. Zacks, Nicole K. Speer, Khena M. Swallow, Todd S. Braver, and Jeremy R. Reynolds. Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2):273–293, 2007.

[167] Dan Zahavi. *Subjectivity and Selfhood: Investigating the First-Person Perspective*. MIT Press, 2005.

[168] Dan Zahavi. *Self and Other: Exploring Subjectivity, Empathy, and Shame*. Oxford University Press, 2014.

[169] Daria Zakharova. Missing the subject: Introspection in large language models. *PhilSci-Archive*, (27052), 2025.

[170] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proc. ICML*, pages 3987–3995, 2017.

[171] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *ACM Transactions on Information Systems*, 43(6), 2025.

[172] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *ICML*, 2025.

[173] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. In *ICLR*, 2017.