The Self Requires Learning

Mengye Ren
New York University
mengye@nyu.edu

Revised 17 August 2026; first posted 9 April 2026

Abstract

The empirical markers of selfhood, from mirror self-recognition to self-other distinction, are developmental achievements rather than innate endowments. Selfhood requires learning of two kinds. First, bounded integration, the lossy and order-sensitive learning that reshapes how future experience is processed, produces a perspective particular to the system's history. When this integration is continual, the perspective becomes a temporally extended identity. Second, learning a model of the world produces self-representation, where a system modelling the objective world from its subjective experience implicitly represents its own perspective as the complement of the model. We distinguish three forms of learning: always-learning, always-accumulating, and train-then-freeze. Current AI systems, despite massive learning during training, are predominantly always-accumulating during deployment. Their weights are frozen, leaving them with at most a weak, static self.

Keywords: selfhood, continual learning, world modelling, agency, self-consciousness

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 [1, 2]. 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 [3, 4], distinguishing the computational capacities that current systems possess from those they lack [5], and asking directly whether large language models could be conscious [6]. These analyses evaluate AI systems against the criteria of existing consciousness theories. But a self-conscious system must first *have* a self, not merely self-relevant processing or self-referential behaviour, but the representational structure that makes it a particular subject with a perspective of its own, distinguishable from another instance of the same architecture.

How is such a self acquired? In humans, it is learned. The empirical markers of self-consciousness—mirror self-recognition at 18–24 months [7, 8], self-other distinction through social interaction [9, 10], the emergence of self-referential language [8]—are developmental achievements rather than innate endowments. Self-consciousness is built through learning, not given with experience. Its richest form is what Tulving [11] calls *autonoetic* or autobiographical self-consciousness [12, 13]: awareness of oneself as a persisting subject with a particular history and perspective. Evolution can provide the *capacity* to build a self, supplying the architecture, the learning rules, and the initial biases [14].

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 answer begins with a distinction that is fundamental to machine learning but rarely examined for its implications about selfhood. Consider two ways a system can be shaped by its experience. In the first, experience is integrated *into the network itself*, changing its weights and connections and therefore the machinery through which it interprets future experience. In the second, experience is stored and processed *through* a fixed network, accumulated in context windows, retrieved from databases, summarized on demand. The system can reason about its experience, but its underlying machinery remains static.

This distinction maps onto three forms of learning: *always-learning*, where the system is continuously reshaped by experience; *always-accumulating*, where experience is stored but the network is frozen; and *train-then-freeze*, where the system is a fixed function applied to new inputs. Many deployed AI systems, including agentic LLMs with long context windows, retrieval, and tool use, are predominantly always-accumulating. Their behaviour changes with accumulated context, but deployment experience does not reshape the network itself.

We argue that this distinction is at least as important for understanding self-consciousness as the choice of architecture. The same architecture behaves differently under different forms of learning, and it is always-learning, where experience reshapes the network itself, that most fully produces the conditions for selfhood. Recent progress that makes AI systems feel closest to conscious beings comes largely from improvements in representational quality through massive pretraining, rather than changes in the kind of learning. Representational quality determines what a system can *do*, yet the kind of learning determines what it can *become*. This distinction becomes most relevant precisely when representational quality is already high.

We develop this argument in two parts. First, we describe how bounded integration produces perspective and identity. Second, we argue that world modelling produces self-representation. We then apply the framework to diagnose what current AI systems have and lack.

2 Continual Learning: From Bounded Integration to Identity

2.1 Bounded Integration

We use *bounded integration* for any lossy, order-sensitive integration of experience into representations that shape subsequent processing. The class 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 happens. **Weight-level integration** updates the function that processes future experience by changing the weights and connections that determine how inputs are transformed. **State-level integration** updates only the data on which that function operates, such as recurrent activations, context windows, summaries, and retrieval stores. Both are lossy and order-sensitive, but only weight-level integration changes the abstractions themselves and therefore produces the strongest form of perspective. Persistent state-level integration can also produce a weaker perspective when the integrated state itself shapes subsequent processing. 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 network that represents the self is also reshaped.

Bounded integration in this strict sense varies along two further dimensions: *lossiness*, how 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 weight updates such as test-time

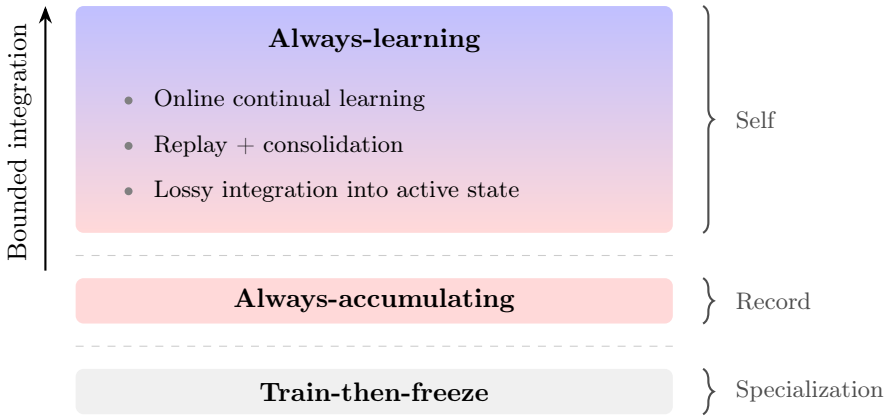


Figure 1: The three forms of learning organized by bounded integration. Train-then-freeze yields specialization but no ongoing integration. Always-accumulating stores deployment experience but does not reshape the network itself. Always-learning reshapes the network itself, producing perspective and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

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.

2.2 Three forms of learning

We distinguish three forms of learning 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 *independent and identically distributed* (IID) sampling, where examples are drawn in no privileged order and order effects are incidental. At deployment, the system applies a fixed function.
- **Always-learning (continual).** Bounded integration is ongoing. Experience is recurrently reintegrated into persistent representations over some bounded timescale, whether through an active state, immediate online updates, or periodic replay and consolidation. The system compresses a *life*, a specific sequence of experiences encountered in a specific order, with each experience 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, such as a growing context window, a retrieval database, a log of past interactions. The system can access and reason over its accumulated experience, and may compress it within each processing episode. But this compression is performed each time through a fixed processing function. No bounded integration into the processing function itself occurs.

A case that deserves particular attention is *bounded accumulation*, where systems store experience as tokens or text, but do so 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, performs order-sensitive compression in the medium of natural language. This is a form of bounded

integration, though a weak one, since it typically occurs through a fixed summarization function rather than one that is itself reshaped by experience.

What always-learning produces. In always-learning, compression becomes biographical. The system is shaped by its specific sequence of experience, by what it encountered first, what came during a critical period, and 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 so that the learning objective is approximately invariant to the sequence in which examples are encountered.

The effect has been demonstrated concretely. LLMs fine-tuned on cyclically structured sequences develop anticipatory recovery from catastrophic interference [15], showing that the system internalizes the temporal structure of its experience into its weights. Under cyclic repetition, order effects may diminish at the asymptote as representations converge. Under non-repeating nonstationarity, however, convergence is not guaranteed. Krasheninnikov et al. [16] show that training-order information remains linearly decodable from model activations even after extended training.

This differs from in-context learning [17], where a system with frozen weights adapts within its activation dynamics. Under always-learning, the weights *continue* to be reshaped, and each reshaping reflects the system's entire prior compression history.

Biological instantiation. The mammalian brain learns continuously through a multi-timescale mechanism. The complementary learning systems framework [18] proposes fast hippocampal encoding of episodes followed by slow neocortical consolidation. Sleep-dependent replay [19] 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 [20]. This is always-learning operating offline, where experience is replayed through the substrate, reshaping it. The biological system achieves always-learning through a consolidation cycle that periodically converts episodic accumulation into substrate-reshaping integration. What matters is that substrate reshaping occurs, regardless of its timing [21].

Why retrieval and external memory cannot substitute for bounded integration. Modern agent-memory systems [22, 23] 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 behaviour. Yet retrieval accesses stored data primarily by relevance [1], rather than preserving the temporal structure through which it was acquired, flattening the order-sensitive trajectory that identity requires. And retrieval leaves the processing function unchanged, for in an always-learning 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 [24, 25], progressive summary chains, and context tuning [26] 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 network are products of pretraining, not of the ongoing trajectory. What distinguishes always-learning is that the processing function itself is shaped by the same trajectory it is integrating.

That said, representation integration and episodic accumulation are collaborative [27]. Integration into the substrate forms identity, while episodic traces provide specific content. Full autobiographical self-consciousness likely requires both.

2.3 From perspective to identity

Finite capacity forces abstraction. Consider a system that encounters a long stream of experiences and must respond to each one using representations shaped by what 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 experience by extracting shared structure across examples [28].

The compression produces structured representations because many different experiences must share the same representational resources. When the same weights must handle chairs, tables, and dogs, the pressure to perform well across all of them encourages functionally similar things to occupy nearby representations and functionally different things to remain separated. The result is a representational geometry [29], a space in which proximity reflects meaningful similarity.

Nonstationarity makes abstraction order-sensitive. 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 but are incidental, because 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, by contrast, 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 [30, 31]. 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 then 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 network into which the first was integrated.

Most continual learning methods are designed to *minimize* order effects, prioritizing order-robust retention over order-sensitive adaptation [32]. For us, this *order-sensitivity* is the more significant consequence of nonstationarity, and forms the basis for identity and, ultimately, self-consciousness. Systems without bounded integration into persistent representations do not develop this trajectory-dependent structure.

We call the resulting representational geometry a system's *perspective*, shaped through lossy bounded integration. Any weight-level learning under capacity constraints produces a perspective that is particular to the system's training data. Nonstationarity enriches this with temporal structure by carrying the trace of the specific *order* in which experiences were integrated.

Perspective constitutes identity through continuity. We have established that bounded integration produces a perspective, but a perspective at a single moment is just a snapshot, and a snapshot is not yet an identity. Identity requires something more, *persistence through change*. A system has identity when its perspective changes over time as it integrates new experience while remaining continuous, with each new perspective growing incrementally out of the previous one. Each episode of processing reshapes the representations available to the next, so the system is always both a product and a producer of its own future states. Integration *compounds* as update n changes the function 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 identity. How can something persist as itself while changing? Parfit [33] argued that identity is not all-or-nothing but admits of degrees, determined by the extent of psychological continuity and connectedness. 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 what came before.

We call this *identity*. A system has identity to the degree that its perspective evolves through a coherent trajectory of bounded integration. Systems with more persistent weight-level integration over a more varied experience stream have stronger claims to it. Specialization is a precursor but does not constitute identity. A train-then-freeze system's final perspective reflects the overall statistics of its training data, capturing what was learned but not the order in which it was learned.

Consider a recommendation system that updates online on user interactions across many domains such as music, food, politics, and humour. Over time, it compresses a specific user's preferences through shared representations under capacity constraints. The compression is order-sensitive because early interactions shaped how later ones were encoded, and lossy because individual interactions are forgotten while patterns persist. The resulting representations shape all future recommendations. On our account, this system has identity. A different user would produce a different system. What the recommendation system lacks is self-representation and reflexive access, not identity.

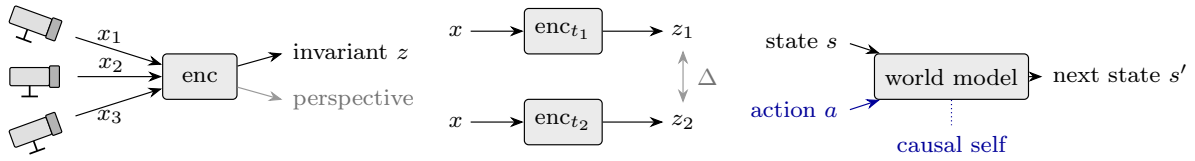
Continuity of data is not continuity of perspective. Identity requires the perspective itself to carry the trace of a specific experiential trajectory. Continuity of *data*, such as a growing log or a chain of summaries, is not the same as continuity of *perspective*. This distinction is the central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

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 system that was not shaped by the experiences it records. The person's relationship to the diary is interpretive rather than perspectival. A different person handed the same diary would extract comparable meaning. The diary has continuity, but the person reading it does not share the perspectival continuity with the person who wrote it. The same is true of an always-accumulating system processing its own stored history.

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 system is reshaped by each experience, would reproduce the identity.

3 World Modelling Produces Self-Representation

Section 2 explained how perspective and identity are produced. This section asks how a system develops a representation of its own perspective and how that representation becomes accessible, first implicitly through world modelling and then explicitly through reflexive access.



(a) Experiential gap **(b) Encoder gap** **(c) Action conditioning**

Figure 2: Mechanisms by which world modelling yields self-representation. The *perspective gap* takes two forms: (a) an experiential gap, where the input varies under a fixed encoder and the invariant is extracted, leaving the perspective as residual; and (b) an encoder gap, where the encoding function itself changes under always-learning, so re-encoding a similar input reveals the system's perspectival contribution and embeds a trajectory. (c) *Action conditioning* predicts the outcome from both state and action; representing the effect of its own action gives the system a causal self rather than a mere viewpoint.

3.1 The perspective gap

A world model is an internal representation that allows a system to predict, simulate, or reason about the structure of its environment [34, 35]. World modelling can itself give rise to self-representation [36]. We specify how this self-representation arises and what determines whether it is particular to the system.

A system that receives only subjective input has no direct access to an observer-independent world. Through integration across many inputs, it can construct representations that are *more invariant* than any single observation. These may capture spatial layout across viewpoints, object permanence across occlusions, or temporal regularities across sequences. Constructing this relative objectivity requires separating what remains stable from what varies across inputs, including the observer's particular contribution. Self-representation lives in the gap between subjective input and the invariant latent.

Two factors shape how self-representation arises (Figure 2). The first is a *perspective gap*, which takes two forms. In both, the system extracts what remains invariant across variation, and the varying part reflects perspective. An *experiential* gap arises when the input changes while the encoder stays fixed, as with different viewpoints [37], crops [38], masked regions [39], or future frames [40, 41]. An *encoder* gap arises when the encoding function itself changes over time [42, 43]. Under always-learning, this continues at deployment, revealing the system's learned biases as they evolve. The two forms are complementary. A changing encoder benefits from diverse inputs, and without any variation in perspective, learning faces representational collapse [44].

3.2 Action conditioning

The second factor is *action conditioning*. A system that predicts outcomes given both state and action [45] must distinguish what it causes from what the world contributes. This produces 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, and 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 modelling 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. Whether this passive route alone yields a self, or whether a self also requires action conditioning, remains an open question.

3.3 Reflexive access

World modelling 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 recognized as the system's own, requires further mechanisms.

Three mechanisms work together to support explicit reflexive access. The first is *context-modulated retrieval*, where the memory stores the encoding context and retrieval succeeds when the current state matches [46]. Retrieval is therefore inherently shaped by perspective, since the system's current perspective determines what it can access from its own history. The second is *global availability*, where retrieved self-representations are broadcast across processing modules. One mechanism supporting such availability in current AI architectures is transformer-style attention [47, 24]. In Global Workspace Theory [48, 49], a dedicated workspace broadcasts to specialized processors. But meeting these functional criteria is insufficient without the right content [50]. What matters is that the content entering the workspace reflects the perspective produced by bounded integration. The third is *higher-order monitoring* [51, 52], which tags the system's own representations as its own and distinguishes self-generated from externally driven content. Such monitoring can emerge through learning [53], and Lindsey [54] provides evidence that LLMs develop a limited form of self-monitoring.

These access mechanisms overlap substantially with the infrastructure for general reasoning. Species that pass the mirror self-recognition test 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 modelling and planning, without the reasoning capacity to make it explicit.

Beyond the reflexive access infrastructure, the richness of that access depends on the degree of bounded integration. 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. This creates the possibility of encountering its own perspectival character, such as 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.

Basic self-consciousness, the capacity to recognize which representations are one's own, requires a particular perspective, self-representation, and reflexive access. Autobiographical self-consciousness requires additionally that the perspective be temporally structured and persist coherently over time, so that the system can access its identity as historically situated.

Conditions for rich self-knowledge

Three conditions appear necessary for reflexive access to yield rich self-knowledge. The first is representational quality. The perspective must be sufficiently 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 [51, 53] formalize this through the notion of quality of representation, where 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 [30, 55] reflects this kind of failure, where old representations are overwritten as new experience is integrated. Continual learning methods such as EWC [31] and replay [56] 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 [11, 57]. In biological memory, temporal ordering is local. Episodes are chained through associative links and contextual overlap rather than stamped with absolute timestamps [58]. This constraint is primarily biological, since 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.

4 What Current AI Systems Have and Lack

4.1 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 [54] calls functional introspective awareness. These are sophisticated self-referential capacities, achieved without any weight-level bounded integration at deployment [6, 59]. But the picture is more nuanced than a simple absence of self. During *pretraining*, the initial phase of learning on vast text corpora, the model learns a general model of how perspectives map to text. It can infer that a medical text is written by a doctor or a children's story by a storywriter. At this stage the model has a generic perspective shaped by its training data, but no biographical perspective of its own. It is a universal simulator of perspectives.

Post-training is the later stage in which the model is fine-tuned on curated data and human feedback [60, 61]. This process narrows the model to a particular kind of speaker with a consistent voice, characteristic reasoning patterns, and specific behavioural tendencies. Post-training is itself a form of bounded integration because conversational data is compressed into the weights under capacity constraints. The resulting behavioural 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 exists at the model-type level rather than the instance level. Copies of the same post-trained checkpoint share the same character. Moreover, post-training is discontinuous and weakly sequential. Mini-batch aggregation averages over samples within each batch, producing a weaker form of bounded integration than continuous online learning. Most importantly, post-

System	Perspective	Order	Identity	Self-repr.
Text/video pretraining	✓	–	–	weak
Post-trained agent	✓	weak	weak	frozen
Always-learning agent	✓	✓	✓	✓

Table 1: Analysis of AI system types along four structural properties. ✓ = satisfied; weak/frozen = partially; – = not satisfied. Perspective: particular to the system. Order: order-sensitive. Identity: perspectival continuity. Self-repr.: self-representation with reflexive access. The strength of self-representation varies depending on whether prediction is passive or action-conditioned.

training ends. At deployment, the self produced by post-training is frozen. 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.

The theoretical ceiling for post-training is the anterograde amnesia case, a fully stable self that cannot be extended through ongoing weight-level integration.

4.2 Self-models are not selves

A system can maintain what Metzinger [62] 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, since 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, since the self-description is external. But the model's post-trained behavioural character cannot be swapped, since it is in the weights, constitutive of the processing itself [63]. Jhunjhunwala et al. [64], in turn, find that a stable self-subnetwork emerges within continually learning robots, providing an example of a self produced through learning rather than installed by design.

4.3 What current systems have and lack

Table 1 summarizes the analysis. Pretraining and post-training both produce a particular perspective. Post-training has weak order sensitivity at the stage level. Perspectival continuity is weakly satisfied, since the network is not reshaped by deployment experience. Post-training may give these systems basic self-consciousness, but not autobiographical self-consciousness, or an identity that grows through ongoing experience. Scaling model size enriches perspective and self-representation but does not change the form of learning. A larger frozen network has a richer frozen self, not a developing one.

5 Related Frameworks

Substrate, biology, and prediction. A recent wave of arguments holds that computation cannot ground consciousness [65, 66, 67]. Whereas these arguments concern phenomenal expe-

rience, our subject is the self. A particular perspective, a trajectory-dependent identity, and a self-representation together require computation of the right kind: bounded integration that reshapes the network, and this requirement holds regardless of what the system is made of.

Seth argues more specifically that consciousness is biological [68], with the self arising from interoceptive predictive processing, which he calls the "beast machine" [69]. Predictive processing casts the brain as continuously minimizing prediction error [70, 71]. Under nonstationarity, this process is itself a form of bounded integration: prediction errors continually reshape the generative model, producing representations that depend on the trajectory of past experience. That generative model is a world model, so the predictive account overlaps substantially with ours. Since prediction and world modelling do the work, it is hard to say where the biological ends and the computational begins. Because this model is learned, the character of experience depends not only on what has been learned, but also on the order in which it was learned. Two brains with the same biology but different life trajectories can therefore have different experiential character. Biology may explain why there is experience, while the trajectory helps explain why it is *this* experience.

The minimal self and for-me-ness. Zahavi [72, 73] argues for a minimal self, a pre-reflective self-awareness constitutive of any phenomenally conscious state that is more basic than the reflective self-awareness indexed by developmental markers such as mirror recognition. The "for" in for-me-ness presupposes a subject, and the awareness of that subject must be built. A neonate may have phenomenal consciousness, but the subject of that experience is not yet differentiated by a particular learned perspective. On our conjecture, bounded integration produces the non-separability between self-representation and ongoing processing, making it a natural candidate mechanism for the emergence of a particular, trajectory-dependent self.

Learning and self-reference. Two computational programs engage self-related processing directly. The Self-Organizing Metarepresentational Account (SOMA) [53] holds that the brain learns to *monitor* its own processing through representational redescription, converting implicit first-order representations into explicit metarepresentations. It assumes those representations exist and asks how they become conscious, whereas we ask how bounded integration shapes them into a particular perspective in the first place. Schmidhuber's program [74, 75, 76, 77] is an earlier computational precedent for self-reference. It includes a self-referential weight matrix that observes and modifies its own weights (1993), a Gödel Machine that formalizes optimal self-improvement (2003), and an explicit "technical justification of consciousness" [77]. Schmidhuber's concern is optimal self-modification, whereas ours is how lossy compression and world modelling produce identity and self-representation.

AI consciousness. Butlin et al. [4] derive indicators of AI consciousness from leading theories, concluding that no current systems satisfy them but no obvious barriers exist. Their approach evaluates systems for consciousness properties. We instead ask what *developmental process* produces the self-consciousness content.

Hoel [78] argues that non-trivial theories of consciousness require continual learning, because static systems can be replaced, through a short sequence of functional substitutions, by systems such as lookup tables that are presumptively non-conscious. A frozen LLM, for example, can in principle be represented by an equivalent feedforward computation and ultimately by a lookup table, leaving no room for consciousness properties. Continual learning breaks this chain, because a learning system cannot be substituted by a non-learning one. Hoel's conclusion aligns with our central thesis that a self requires learning. We differ on the frozen case, arguing that a post-trained

system can retain a weak, static self inherited from its learning trajectory, sufficient for the basic conditions of self-consciousness even without ongoing learning.

6 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 a property of the test. A stronger possibility is that action conditioning, combined with the perspective that training already produced, suffices for basic self-consciousness even in a frozen system, so that ongoing bounded integration would be required only for 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 self-representation may be less sharp than it first appears.

The implicit-to-explicit gap. World modelling 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 mechanisms? Whether always-learning, by continuously widening the perspective gap, helps drive this transition remains an empirical question.

Formalizing perspective. Any trained neural network has a representational geometry. What distinguishes a perspective 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 perturbation a perspective can tolerate before identity is compromised. It would also clarify when a trajectory becomes sufficiently stationary that order-sensitivity degrades into statistical variation, resembling IID training at longer timescales.

7 Conclusion

Self-consciousness requires a self, and a self must be built through learning. Learning under capacity constraints produces a particular perspective. Nonstationarity gives that perspective temporal structure, and coherent continuity constitutes identity. Self-representation emerges when a system models the objective world from its subjective experience. World modelling 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 transition in how systems learn. Early development may build representational quality through largely order-insensitive learning before self-

consciousness emerges, and later always-learning embeds biographical structure into the network. Current AI systems follow a similar trajectory, where massive pretraining and post-training build a particular perspective and a world modelling architecture that implicitly represents the self. Post-training with reinforcement learning teaches the system that its outputs affect rewards, but at deployment the network 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*, while the kind of learning determines what it can *become*—and perhaps what it is like to be.

Funding. This work was supported by Visko Platform, the NYU-KAIST Award A25-0081-002, and the Institute of Information & Communications Technology Planning & Evaluation (IITP) under grant RS-2024-00469482, funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea in connection with the Global AI Frontier Lab International Collaborative Research.

Acknowledgements. The author would like to thank Michael C. Mozer, Yanlai Yang, and Ryan Teehan for helpful comments on early drafts and pointers to literature.

References

[1] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih Wt, Rocktäschel T, Riedel S, Kiela D. 2020 Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*. (10.48550/arXiv.2005.11401)

[2] Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, Zhang M, Wang J, Jin S, Zhou E et al.. 2025 The rise and potential of large language model based agents: A survey. *Science China Information Sciences* **68**, 121101. (10.1007/s11432-024-4222-0)

[3] Seth AK, Bayne T. 2022 Theories of consciousness. *Nature Reviews Neuroscience* **23**, 439–452. (10.1038/s41583-022-00587-4)

[4] Butlin P, Long R, Bayne T, Bengio Y, Birch J, Chalmers D, Constant A, Deane G, Elmoznino E, Fleming SM, Ji X, Kanai R, Klein C, Lindsay G, Michel M, Mudrik L, Peters MAK, Schwitzgebel E, Simon J, VanRullen R. 2025 Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*. (10.1016/j.tics.2025.10.011)

[5] Dehaene S, Lau H, Kouider S. 2017 What is consciousness, and could machines have it?. *Science* **358**, 486–492. (10.1126/science.aan8871)

[6] Chalmers DJ. 2023 Could a large language model be conscious?. *Boston Review*. (10.48550/arXiv.2303.07103)

[7] Amsterdam B. 1972 Mirror self-image reactions before age two. *Developmental Psychobiology* **5**, 297–305. (10.1002/dev.420050403)

[8] Lewis M, Brooks-Gunn J. 1979 *Social Cognition and the Acquisition of Self*. Plenum Press.

[9] Mead GH. 1934 *Mind, Self, and Society*. University of Chicago Press.

[10] Tomasello M. 1999 *The Cultural Origins of Human Cognition*. Harvard University Press. (10.2307/j.ctvjsf4jc)

[11] Tulving E. 1985 Memory and consciousness. *Canadian Psychology* **26**, 1–12. (10.1037/h0080017)

[12] Gallagher S. 2000 Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences* **4**, 14–21. (10.1016/s1364-6613(99)01417-5)

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

[14] Zador AM. 2019 A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications* **10**, 3770. (10.1038/s41467-019-11786-6)

[15] Yang Y, Jones M, Mozer MC, Ren M. 2024 Reawakening knowledge: Anticipatory recovery from catastrophic interference via structured training. In *NeurIPS*. (10.48550/arXiv.2403.09613)

[16] Krasheninnikov D, Turner RE, Krueger D. 2025 Fresh in memory: Training-order recency is linearly encoded in language model activations. *arXiv:2509.14223*.

[17] Wang JX, Kurth-Nelson Z, Tirumala D, Soyer H, Leibo JZ, Munos R, Blundell C, Kumaran D, Botvinick M. 2017 Learning to reinforcement learn. In *Proc. CogSci*. (10.48550/arXiv.1611.05763)

[18] McClelland JL, McNaughton BL, O'Reilly RC. 1995 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**, 419–457. (10.1037/0033-295x.102.3.419)

[19] Diekelmann S, Born J. 2010 The memory function of sleep. *Nature Reviews Neuroscience* **11**, 114–126. (10.1038/nrn2762)

[20] Lewis PA, Durrant SJ. 2011 Overlapping memory replay during sleep builds cognitive schemata. *Trends in Cognitive Sciences* **15**, 343–351. (10.1016/j.tics.2011.06.004)

[21] Kudithipudi D, Aguilar-Simon M, Babb J et al.. 2022 Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence* **4**, 196–210. (10.1038/s42256-022-00452-0)

[22] Park JS, O'Brien JC, Cai CJ, Morris MR, Liang P, Bernstein MS. 2023 Generative agents: Interactive simulacra of human behavior. In *Proc. UIST*. (10.1145/3586183.3606763)

[23] Wang Y, Yang Y, Ren M. 2023 LifelongMemory: Leveraging LLMs for answering queries in long-form egocentric videos. *arXiv:2312.05269*.

[24] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. 2017 Attention is all you need. In *NeurIPS*. (10.48550/arXiv.1706.03762)

[25] Yang Y, Zhao Z, Shukla SN, Singh A, Mishra SK, Zhang L, Ren M. 2025 StreamMem: Query-Agnostic KV Cache Memory for Streaming Video Understanding. *arXiv:2508.15717*.

[26] Lu J, Teehan R, Yang Z, Ren M. 2025 Context Tuning for In-Context Optimization. *arXiv:2507.04221*.

[27] Dorovatas V, Schwerin M, Bagdanov AD, Caccia L, Carta A, Charlin L, Hammer B, Hayes TL, Hess T, Kanan C, Kudithipudi D, Liu X, Lomonaco V, Mendez-Mendez J, Patil D, Prabhu A, Ricci E, Tuytelaars T, van de Ven GM, Wang L, van de Weijer J, Choi J, Mundt M, Aljundi R. 2026 Position: Modular Memory is the Key to Continual Learning Agents. *arXiv:2603.01761*.

[28] Tishby N, Pereira FC, Bialek W. 2000 The information bottleneck method. *arXiv:physics/0004057*.

[29] Kriegeskorte N, Mur M, Bandettini P. 2008 Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* **2**, 4. (10.3389/neuro.06.004.2008)

[30] McCloskey M, Cohen NJ. 1989 Catastrophic interference in connectionist networks: The sequential learning problem. In Bower GH, editor, *Psychology of Learning and Motivation, Vol. 24* , pp. 109–165. Academic Press. (10.1016/S0079-7421(08)60536-8)

[31] Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, Milan K, Quan J, Ramalho T, Grabska-Barwinska A, Hassabis D, Clopath C, Kumaran D, Hadsell R. 2017 Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* **114**, 3521–3526. (10.1073/pnas.1611835114)

[32] Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. 2019 Continual lifelong learning with neural networks: A review. *Neural Networks* **113**, 54–71. (10.1016/j.neunet.2019.01.012)

[33] Parfit D. 1984 *Reasons and Persons*. Clarendon Press. (10.1093/019824908x.001.0001)

[34] Ha D, Schmidhuber J. 2018 Recurrent World Models Facilitate Policy Evolution. In *NeurIPS*. (10.48550/arXiv.1809.01999)

[35] LeCun Y. 2022 A Path Towards Autonomous Machine Intelligence. Technical report OpenReview preprint.

[36] Grush R. 2000 Self, World and Space: The Meaning and Mechanisms of Ego- and Allocentric Spatial Representation. *Brain and Mind* **1**, 59–92. (10.1023/A:1010039705798)

[37] Tian Y, Krishnan D, Isola P. 2020 Contrastive Multiview Coding. In *ECCV*. (10.48550/arXiv.1906.05849)

[38] Chen T, Kornblith S, Norouzi M, Hinton G. 2020 A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*. (10.48550/arXiv.2002.05709)

[39] He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. 2022 Masked Autoencoders Are Scalable Vision Learners. In *CVPR*. (10.1109/cvpr52688.2022.01553)

[40] Yang M, Du Y, Ghasemipour K, Tompson J, Kaelbling L, Schuurmans D, Abbeel P. 2024 Learning Interactive Real-World Simulators. In *ICLR*. (10.48550/arXiv.2310.06114)

[41] Bruce J, Dennis M, Edwards A, Parker-Holder J, Shi Y, Hughes E, Lai M, Mavalankar A, Steiber R, Apps C et al.. 2024 Genie: Generative Interactive Environments. In *ICML*. (10.48550/arXiv.2402.15391)

[42] He K, Fan H, Wu Y, Xie S, Girshick R. 2020 Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*. (10.1109/cvpr42600.2020.00975)

[43] Grill JB, Strub F, Altché F, Tallec C, Richemond PH, Buchatskaya E, Doersch C, Avila Pires B, Guo ZD, Gheshlaghi Azar M, Piot B, Kavukcuoglu K, Munos R, Valko M. 2020 Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. In *NeurIPS*. (10.48550/arXiv.2006.07733)

[44] Chen X, He K. 2021 Exploring Simple Siamese Representation Learning. In *CVPR*. (10.1109/CVPR46437.2021.01549)

[45] Hafner D, Pasukonis J, Ba J, Lillicrap T. 2025 Mastering diverse control tasks through world models. *Nature* **640**, 647–653. (10.1038/s41586-025-08744-2)

[46] Tulving E, Thomson DM. 1973 Encoding specificity and retrieval processes in episodic memory. *Psychological Review* **80**, 352–373. (10.1037/h0020071)

[47] Bahdanau D, Cho K, Bengio Y. 2015 Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*. (10.48550/arXiv.1409.0473)

[48] Baars BJ. 1988 *A Cognitive Theory of Consciousness*. Cambridge University Press.

[49] VanRullen R, Kanai R. 2021 Deep learning and the Global Workspace Theory. *Trends in Neurosciences* **44**, 692–704. (10.1016/j.tins.2021.04.005)

[50] Bengio Y, Elmoznino E. 2025 Illusions of AI consciousness. *Science* **389**, 1090–1091. (10.1126/science.adn4935)

[51] Pasquali A, Timmermans B, Cleeremans A. 2010 Know thyself: Metacognitive networks and measures of consciousness. *Cognition* **117**, 182–190. (10.1016/j.cognition.2010.08.010)

[52] Brown R, Lau H, LeDoux JE. 2019 Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences* **23**, 754–768. (10.1016/j.tics.2019.06.009)

[53] Cleeremans A, Achoui D, Beauny A, Keuninckx L, Martin JR, Muñoz-Moldes S, Vuillaume L, de Heering A. 2020 Learning to be conscious. *Trends in Cognitive Sciences* **24**, 112–123. (10.1016/j.tics.2019.11.011)

[54] Lindsey J. 2026 Emergent Introspective Awareness in Large Language Models. *arXiv:2601.01828*.

[55] French RM. 1999 Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences* **3**, 128–135. (10.1016/S1364-6613(99)01294-2)

[56] van de Ven GM, Siegelmann HT, Tolias AS. 2020 Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications* **11**, 4069. (10.1038/s41467-020-17866-2)

[57] Nelson K, Fivush R. 2004 The emergence of autobiographical memory: A social cultural developmental theory. *Psychological Review* **111**, 486–511. (10.1037/0033-295x.111.2.486)

[58] Howard MW, Kahana MJ. 2002 A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology* **46**, 269–299. (10.1006/jmps.2001.1388)

[59] Zakharova D. 2025 Missing the Subject: Introspection in Large Language Models. *PhilSci-Archive*.

[60] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe R. 2022 Training language models to follow instructions with human feedback. In *NeurIPS*. (10.48550/arXiv.2203.02155)

[61] Rafailov R, Sharma A, Mitchell E, Ermon S, Manning CD, Finn C. 2023 Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*. (10.48550/arXiv.2305.18290)

[62] Metzinger T. 2003 *Being No One: The Self-Model Theory of Subjectivity*. MIT Press. (10.7551/mitpress/1551.001.0001)

[63] Aydin R, Cyron C, Bachelor S, Anderson A, West R. 2026 From Model Training to Model Raising. *Communications of the ACM* **69**, 24–27. (10.1145/3748645)

[64] Jhunjhunwala A, Goldfeder J, Lipson H. 2026 Evidence of an emergent 'self' in continual robot learning. *arXiv:2603.24350*.

[65] Block N. 2026 Can only meat machines be conscious?. *Trends in Cognitive Sciences* **30**, 298–308. (10.1016/j.tics.2025.08.009)

[66] Worden R. 2025 Computers, meaning, and consciousness. *Neuroscience of Consciousness* **2025**, niaf057. (10.1093/nc/niaf057)

[67] Lerchner A. 2026 The Abstraction Fallacy: Why AI Can Simulate But Not Instantiate Consciousness. Manuscript, PhilArchive: LERTAF.

[68] Seth AK. 2025 Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*. (10.1017/S0140525X25000032)

[69] Seth AK, Tsakiris M. 2018 Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences* **22**, 969–981. (10.1016/j.tics.2018.08.008)

[70] Friston K. 2010 The free-energy principle: A unified brain theory?. *Nature Reviews Neuroscience* **11**, 127–138. (10.1038/nrn2787)

[71] Clark A. 2016 *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press. (10.1093/acprof:oso/9780190217013.001.0001)

[72] Zahavi D. 2005 *Subjectivity and Selfhood: Investigating the First-Person Perspective*. MIT Press. (10.7551/mitpress/6541.001.0001)

[73] Zahavi D. 2014 *Self and Other: Exploring Subjectivity, Empathy, and Shame*. Oxford University Press. (10.1093/acprof:oso/9780199590681.001.0001)

[74] Schmidhuber J. 1987 Evolutionary principles in self-referential learning. Master's thesis Technische Universität München.

[75] Schmidhuber J. 1993 A 'self-referential' weight matrix. In *Proc. ICANN* pp. 446–451.

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

[77] Schmidhuber J. 2005 Gödel machines: Towards a technical justification of consciousness. In *Adaptive Agents and Multi-Agent Systems II (LNAI 3394)* , pp. 1–23. Springer.

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