Graduate School of Computational Neuroscience

University of Tübingen

Growing Causal Abstractions

Rotation Essay

Prepared by

Mohammed Abbas Ansari

Supervised by

Prof. Dr. Bernhard Schölkopf
Dr. Lancelot Da Costa

ELLIS Institute & Max Planck Institute for Intelligent Systems
Department of Empirical Inference

30 March 2026

Abstract

Every experience is particular, yet intelligent behaviour depends on extracting something general from those particulars: structure that transfers to situations never previously encountered. How specific episodes give rise to reusable, compositional internal representations remains unresolved, with relevant formalizations scattered across cognitive science, reinforcement learning, causal inference, program induction, and active inference. This essay synthesizes these traditions around a single organizing concept, the *library of causal abstractions*, and makes three contributions. First, it identifies five functional constraints any system converting episodes into reusable causal structure must satisfy: structured encoding, belief-state construction under partial observability, modular reusable components, vocabulary expansion, and causal robustness. Second, it shows these constraints generate four tensions (specificity versus reusability, stability versus revisability, commitment versus uncertainty, causal depth versus epistemic cost) that force architectural trade-offs no solution can avoid. Third, it offers a biologically informed reading of the hippocampal–prefrontal system through these tensions, proposing a format hypothesis: the brain's continual-learning advantage rests not only on separated learning rates but on geometric modularity of consolidated knowledge, where orthogonal population subspaces enable extension without interference. The framework is interpretive, not a mechanistic model. Its value lies in generating predictions that no single contributing tradition yields alone, including dissociations between structural generalization and item memory under replay disruption, threshold effects in schema revision, and exploration-dependent transfer robustness. The deepest implication is that catastrophic forgetting and abstraction formation are not separate problems; the second may be the more fundamental one of which the first is a consequence.

Contents

1 Introduction

Every experience is particular, yet intelligent behaviour depends on extracting something general from those particulars: something about how mazes work, where predators appear, or what makes a stack stable, something that transfers to situations never previously encountered. The puzzle is not that this happens, but *how*. By what process do specific, unrepeatable episodes give rise to internal structure that is general, reusable, and compositional?

This question has been formalized independently across several traditions. Cognitive science has studied how learners infer structured hypothesis spaces from sparse data, discovering that generalization depends on recovering the *form* of a problem rather than merely fitting its surface statistics (Kemp and Tenenbaum, 2008; Tenenbaum et al., 2011). Reinforcement learning (RL) has formalized how agents discover reusable skills and state representations that reduce future planning cost (Dayan, 1993; Sutton et al., 1999; Abel, 2020). Causal inference has clarified what it means for a higher-level description to faithfully represent a lower-level system under intervention (Beckers and Halpern, 2019; Schölkopf et al., 2021). Program induction has shown how a system can expand its own representational vocabulary by compressing repeated solution structure into reusable primitives (Ellis et al., 2020). Active inference has proposed that perception, learning, and control can be unified under variational free energy minimization within hierarchical generative models, and more recent work has begun to extend this picture toward expandable object-centric structure learning (Friston, 2010; Da Costa et al., 2020; Parr et al., 2022; Heins et al., 2025). Memory neuroscience has, in turn, shown that hippocampal–prefrontal circuitry supports episodic encoding, offline replay, the emergence of schematic and map-like knowledge, and latent task representations that shape future learning (McClelland et al., 1995; Preston and Eichenbaum, 2013; Behrens et al., 2018; Schuck et al., 2016; Niv, 2019; Tompary and Davachi, 2017; Audrain and McAndrews, 2022).

Each tradition illuminates a genuine aspect of the problem. But no single tradition captures the full challenge: how specific experiences are converted into an expanding repertoire of reusable, causally relevant internal structure over the course of a lifetime. This essay calls such a repertoire a *library of causal abstractions* and asks what it takes to grow one.

The problem also has immediate practical significance. One of the most persistent failures of modern deep learning is catastrophic forgetting: neural networks overwrite old knowledge when learning new tasks because their knowledge is stored in entangled, distributed weight configurations that do not support modular extension (McCloskey and Cohen, 1989; French, 1999). Complementary Learning Systems (CLS) theory offered a foundational insight into why brains avoid this fate: the separation of fast hippocampal encoding from slow cortical integration prevents new experiences from overwriting consolidated knowledge (McClelland et al., 1995). But CLS theory left partly open a further question: what *format* does the consolidated cortical knowledge take? This essay argues that the format matters as much as the separation. Knowledge organized into modular, compositional, reusable components can be extended by adding entries without degrading existing ones. Knowledge entangled in shared distributed representations cannot. The problem of growing causal abstractions is

therefore not separate from the problem of continual learning; it may be the deeper problem of which continual learning is a consequence.

This essay makes three contributions. First, it identifies the functional constraints that any system converting episodes into reusable causal structure must satisfy, drawing on five computational traditions. Second, it shows that these constraints generate four inherent tensions, trade-offs that pull against each other and force architectural compromises on any solution. Third, it offers a biologically informed interpretive reading of the hippocampal–prefrontal system through the lens of those tensions, generating specific, testable predictions that would not arise from any single tradition alone. The essay does not claim that this is the best or only way to understand hippocampal–prefrontal organization. The contribution is a *vocabulary and a set of predictions*, not a mechanistic model or a uniquely determined architecture.

The remainder is organized as follows. Section 2 establishes the conceptual vocabulary. Section 3 surveys computational perspectives organized by the functional constraint each contributes. Section 4 develops four tensions between those constraints. Section 5 presents the hippocampal–prefrontal evidence organized by which tensions it addresses. Section 6 situates the synthesis relative to existing frameworks. Section 7 identifies open questions.

2 Abstractions, Libraries, and What Makes Them Causal

The relevant literatures point to similar phenomena using different words. Psychology speaks of schemas, neuroscience of cognitive maps and latent task states, reinforcement learning of state abstraction and options, program induction of reusable libraries. If these terms are treated as interchangeable, the argument becomes vague. If they are treated as entirely separate, the shared problem disappears. This section establishes a minimal vocabulary precise enough to support concrete claims.

2.1 Episodes and abstractions

An episode is a token experience: a particular trajectory through time, in a specific context, containing a specific conjunction of sensory details, actions, and outcomes. An abstraction is a representation that suppresses some of this specificity while retaining structure expected to be useful in future episodes. The schema literature captures one version of this: organized knowledge structures abstracted across multiple experiences, preserving shared relational organization while discarding what is particular to any single event (Ghosh and Gilboa, 2014; Gilboa and Marlatte, 2017). Cognitive maps capture another: relational encodings of how entities or states are organized, extended beyond spatial domains (O'Keefe and Nadel, 1978; Behrens et al., 2018). Latent task states capture a third: compact internal representations combining observable features with hidden variables inferred from memory, preserving what matters for decision-making (Schuck et al., 2016; Niv, 2019).

What these formulations share is that abstraction involves a *selection*: some aspects of experience are retained because they are expected to recur or matter, while others are discarded. This distinguishes abstraction from mere compression. A lossy image file is

compressed. A cognitive map of a building's layout is an abstraction: it preserves specifically those aspects (spatial relationships, connectivity) expected to be useful for future navigation, while discarding others (lighting, wall colour). Abstraction is an evaluative operation shaped by the agent's history of what has proved useful, not a mechanical operation definable independently of goals and context.

A point easily overlooked is that this evaluative selection does not begin at some late processing stage. What counts as an "episode" is already shaped by the agent's existing knowledge and attentional state. An agent with an existing schema for restaurant visits will segment and encode a dinner differently from one encountering the situation for the first time (Zacks et al., 2007; Ghosh and Gilboa, 2014). Abstraction is already operative in how experience is encoded. This will matter when discussing hippocampal encoding later in Section 5: even the brain's fast episodic system produces relationally organized traces reflecting prior knowledge and current inferential demands (Eichenbaum, 2017; Courellis et al., 2024).

Finally, "abstraction" in the literature refers to operations over different objects. State abstraction concerns grouping states equivalent for some purpose (Abel, 2020). Action abstraction compiles action sequences into temporally extended skills (Sutton et al., 1999). Temporal abstraction chunks continuous streams into discrete events (Zacks et al., 2007). These three forms recur throughout the essay under different names.

2.2 What makes a collection of abstractions a library

Not every collection of abstractions constitutes a library. What distinguishes a library from a stockpile is a set of functional properties that make stored structure *available* for future cognition, not just *present* in the system. Four properties are essential.

Portability. The abstraction must apply beyond the episodes from which it was extracted. A representation of T-junctions that only works in one maze is not portable; one capturing the relational structure of T-junctions in general is.

Retrievability. The abstraction must be accessible when a new situation calls for it. An agent with excellent library entries but no indexing scheme for recognizing when each is relevant has a library with no catalogue.

Revisability. The abstraction must be updatable when new evidence warrants it, without catastrophic loss of other entries. This requires some degree of modularity: the representation of one abstraction should be at least partially independent of others.

Composability without interference. New entries can be added and existing entries recombined without degrading prior entries. This property most sharply distinguishes a library from a distributed representation. In a library, adding a book does not alter existing books. In a neural network's weight matrix, learning something new changes the same parameters encoding everything already learned. This is why deep networks suffer catastrophic forgetting (McCloskey and Cohen, 1989): the format of knowledge storage does not support modular extension. Composability without interference is the structural precondition for continual learning.

It is equally important to specify what does *not* qualify as a library entry. A cached

model-free value function, though useful for action selection, is not a library entry: it is not compositional (it cannot be recombined with other value functions to solve a novel task) and it is not revisable without retraining from scratch. An opaque latent embedding in a neural world model, though grounded in sensory experience, is not a library entry: it is not separately addressable (the system cannot deploy one learned latent state without activating the surrounding representational context) and it cannot be composed with entries from different domains. A low-level feature detector in early visual cortex, though portable across scenes, is not a library entry in the relevant sense: it does not carry the relational, domain-level structure that supports the kind of transfer this essay is concerned with. The boundary, then, is that a library entry must be *functionally addressable* (the system can deploy it selectively without deploying all stored knowledge) and *structurally explicit* enough to support composition with other entries for novel tasks. This criterion excludes most of what standard neural networks learn while including the kinds of modular, relationally organized representations that the schema, cognitive map, and task-state literatures describe.

With these properties in hand, schemas, cognitive maps, latent task states, and learned skills can be seen as different *formats* of library entry serving a common function. They differ in format (event templates versus relational geometry versus hidden-state vectors versus action policies) and in what aspect of the environment they primarily capture. Collapsing them into a single category would lose real distinctions. But they share all four library properties: each is portable across episodes, retrievable when relevant, revisable under new evidence, and (in the biological system) storable without catastrophically interfering with other entries. The organizing claim of this essay is that these constructs are best understood not as separate phenomena that happen to coexist in the brain, but as different realizations of the same underlying role: entries in a library of reusable internal structure. It should be noted that latent task states are the most debatable members of this set. Schemas and cognitive maps are clearly durable structures that persist across episodes and can be deployed in new situations. Latent task states, by contrast, may in some cases be transient control summaries constructed on the fly from working memory, rather than stored entries retrieved from a library. Whether task-state representations are durable and reusable in the same sense as schemas, or whether they are ephemeral constructions assembled from more basic library components, is an empirical question that future work should address. The essay includes them provisionally because the available evidence suggests they can persist and transfer across structurally similar tasks (Schuck et al., 2016; El-Gaby et al., 2024), but this inclusion should be understood as a hypothesis rather than an established fact.

2.3 What "causal" means, and why it must be graded

A representation can be useful without being causal. A latent variable that predicts well may rely on correlations that break under intervention or distributional change. Formal work on causal abstraction clarifies what a stronger standard looks like. In Beckers and Halpern's framework, a causal abstraction maps a low-level model to a higher-level one such that the higher-level model preserves the interventionally relevant structure of the lower-level one: macro-variables and macro-interventions must stand in a principled relation to

micro-variables and micro-interventions (Beckers and Halpern, 2019). Causal representation learning makes a complementary point: the target should be high-level variables aligned with the causal structure of the data-generating process, supporting transfer across distributional changes (Schölkopf et al., 2021).

Why treat "causal" as graded? Because establishing causal faithfulness requires interventional evidence: the agent must manipulate variables and observe consequences, not just track correlations. The degree of causal faithfulness an abstraction can achieve is therefore constrained by the agent's history of interactions. Full faithfulness in the formal sense would require testing the macro-model against all relevant interventions, which is infeasible for bounded agents. In practice, biological abstractions likely satisfy different criteria to different degrees. Three criteria recur across the literatures reviewed in this essay:

Predictive adequacy. The abstraction preserves enough information to predict task-relevant future observables within a given distribution. This is the criterion formalized by the information bottleneck (Tishby et al., 1999). It can be satisfied by passive observation but may break under distributional shift.

Control relevance. The abstraction preserves enough to support near-optimal decisions under partial observability. This is the criterion formalized in Abel's theory of state abstraction (Abel, 2020). It requires evidence from acting in the environment but may be tied to a specific reward structure.

Interventional stability. The abstraction remains valid when mechanisms change, when the agent intervenes in new ways, or when the task shifts. This approaches the formal criterion of Beckers and Halpern (2019) and the desideratum of causal representation learning (Schölkopf et al., 2021). It is epistemically the most expensive to establish, requiring targeted interventions across multiple contexts.

These criteria are partially overlapping rather than strictly ordered. Model-free reinforcement learning demonstrates that control relevance is achievable without predictive adequacy: a cached value function supports near-optimal decisions without forecasting observations (Watkins and Dayan, 1992). Conversely, a macrovariable can be interventionally faithful without being control-relevant until embedded in a task with specific objectives. The three criteria are dimensions along which abstractions can be evaluated, not a strict hierarchy.

This matters for evaluating neuroscientific evidence. Rather than asking whether a neural representation has "achieved causal abstraction" (almost certainly not in the full formal sense), the productive questions are: which criteria does it satisfy? What evidence would establish a stronger one? What processes might move an abstraction from one criterion toward another over the course of experience? The working position of this essay is that most biologically useful abstractions satisfy predictive adequacy and control relevance within their domain, while approaching interventional stability to varying degrees. They are *causally graded*: modular, interaction-sensitive, and reusable under moderate distributional shift, even if not fully faithful in the formal sense.

A legitimate objection is that "causally graded" risks becoming a euphemism for "not actually causal." The term earns its place only if causal depth, even partial, is a distinct and testable property rather than a relabeling of "useful for control." The distinction is this:

a control-relevant abstraction supports good decisions *within the distribution where it was learned*. A causally deeper abstraction supports good decisions *when that distribution changes*, because it has captured something about the mechanism generating outcomes rather than merely the statistical association between states and rewards. An agent that has learned "pushing objects makes them move" through active manipulation holds an abstraction that transfers to novel objects, novel surfaces, and novel goals. An agent that has learned the same state-action-reward contingency through passive observation holds an abstraction that may fail as soon as the context shifts. Both are control-relevant within their training distribution. Only the first has causal depth. This difference is empirically testable through transfer under distributional shift, and it is the reason this essay retains "causal" rather than retreating to "control-relevant" throughout. The word names a real and graded property, not merely an aspiration.

2.4 What "growing" requires

The library cannot be fixed in advance. In a sufficiently complex world, the agent will encounter regularities that cannot be captured by any combination of existing primitives. When this happens, refining parameters within the current vocabulary is not enough. The agent must create genuinely new structure.

This is the distinction between parameter learning and structure learning. Parameter learning adjusts values within a fixed model. Structure learning changes the model itself: adding variables, introducing categories, discovering relational templates. Both are necessary for growing a library: parameter learning keeps existing entries calibrated; structure learning creates new entries when the existing repertoire proves insufficient.

Growth is not only about adding structure. A living library must also support merging (combining entries that capture the same regularity), splitting (separating an entry that is too coarse), and pruning (removing entries no longer useful). Bayesian model reduction provides one principled approach to pruning: evaluating whether a simpler model explains the data nearly as well (Friston et al., 2018; Heins et al., 2025). But these remain isolated solutions rather than a general theory of library maintenance.

The deepest challenge, however, is the *vocabulary problem*. Every structure-learning method presupposes a vocabulary: nodes in a graph, variables in a causal model, primitives in a programming language. The method searches over compositions of that vocabulary but does not create new vocabulary. Bayesian nonparametric methods allow the *number* of components to grow, but the *type* of component is fixed by the prior (Anderson, 1991; Griffiths and Ghahramani, 2011). DreamCoder allows new library primitives by compressing repeated solution fragments, but the base language must be specified in advance (Ellis et al., 2020). The object-centric model of Heins et al. (2025) can add new objects to its generative model, but the compositional grammar defining what an "object" is must be given. In each case, growth occurs within a fixed meta-vocabulary. How a genuinely new meta-vocabulary can emerge from experience remains open. This is arguably the central unsolved problem in understanding how intelligent systems develop over a lifetime.

3 What the Problem Requires: Five Functional Constraints

The definitions in Section 2 specify what a growing library of causal abstractions would look like. They do not yet say what it takes to build one. This section turns to the computational traditions that have formalized different aspects of the problem, asking what functional constraints emerge when the question is taken seriously. Rather than reviewing each tradition in isolation (which risks producing a field-by-field tour that obscures their convergence), the section is organized around the constraints themselves. Five emerge from the synthesis: the need for structured encoding, for belief-state construction under partial observability, for modular reusable components, for mechanisms of vocabulary expansion, and for causal robustness. Each constraint is motivated by a shared computational logic; each is supported by different traditions; and each, as Section 4 will show, is in tension with at least one other.

3.1 Structured encoding for generalization

The most basic reason to build abstractions is that they support generalization. But the kind of generalization that matters for library building is not extrapolation from surface similarity (which requires only a feature space and a distance metric). It is generalization from structure: transfer based on shared relational organization between situations whose surface features may differ entirely. An agent that understands pursuit–evasion in one setting and transfers that understanding to a perceptually different setting is generalizing from relational form, not shared features.

Tenenbaum et al. (2011) argued that this is the central challenge of learning: minds generalize from sparse data because they construct structured internal models rather than merely compressing observed regularities. Gershman (2017) sharpened the point by identifying the "blessing of abstraction": in hierarchical models, higher-level regularities can sometimes be learned *faster* than lower-level details because they pool evidence across many instances. Kemp and Tenenbaum (2008) showed that learners can infer not only the content of a representation but its *form*, using hierarchical Bayesian inference over a grammar of structural types (trees, rings, chains, grids) to determine which relational scaffold best organizes a domain.

These results converge on a common insight: generalization from structure requires representations that preserve relational organization, not just feature co-occurrences. Battaglia et al. (2018) argued that relational inductive biases are central to combinatorial generalization, and Lake and Baroni (2023) demonstrated that standard networks can achieve more systematic compositional generalization with appropriate meta-learning curricula.

The constraint is that episode encoding must preserve relational structure, not just statistical summaries. Without structured raw material, later abstraction-extraction operations have nothing to work with.

3.2 Belief-state construction under partial observability

Intelligent agents do not merely represent domains. They act in them, often without direct access to the information needed for good decisions. The partially observable Markov

decision process (POMDP) framework (Kaelbling et al., 1998) formalizes this problem: when the true state is hidden, rational action requires the agent to maintain a belief state, a sufficient summary of its observation–action history for decision-making. Exact belief maintenance is intractable for all but simple environments, so bounded agents must construct *approximate* internal states that selectively retain task-relevant information.

Two operations are involved. Compression drops aspects of observation irrelevant to the current task. Augmentation adds information not present in the current observation but inferable from memory or prior knowledge. Task-state representations, as formalized by Niv (2019), arise from exactly this combination: they compress observations to task-relevant dimensions and augment them with hidden variables inferred from experience. Critically, the interplay between compression and augmentation changes as the agent's library grows. A schema for restaurant visits allows an agent to infer, from a single observation (being handed a menu), a rich set of hidden variables about the likely event sequence, social roles, and interaction structure, none of which are in the sensory input.

A useful way to make the timescale structure of this inference problem explicit, especially in active-inference-inspired formulations, is to distinguish fast state inference, medium-timescale parameter learning, and slower structure inference requiring sustained evidence of model inadequacy before revision (Da Costa et al., 2020; Costa et al., 2025; Parr et al., 2022). Alternative strategies exist: recurrent model-free RL can develop empirically belief-like internal representations without explicit inference (Ni et al., 2022). But the timescale vocabulary maps naturally onto distinguishable biological processes: neural dynamics for state tracking, synaptic changes for parameter learning, and consolidation for structural revision.

The constraint that emerges is that abstractions must serve control under partial observability. The internal state must preserve whatever latent structure matters for future decisions, even when that structure must be inferred from memory and context. This couples abstraction to the agent's action problem in a way that purely descriptive or compressive views of abstraction do not capture.

3.3 Modular reusable components and continual learning

Repeated structure can be packaged into higher-order units that reduce the cost of future learning and planning. The options framework (Sutton et al., 1999) formalizes this for action sequences: an option is a temporally extended action with an initiation set, internal policy, and termination condition that reduces effective planning depth. Object-Oriented Markov decision processes (Object-Oriented MDPs) (Diuk et al., 2008) capture a complementary kind of reuse organized around entities and relations rather than action sequences: if an agent has learned interaction rules among object types, it can reuse them with new tokens of the same types. Dayan's successor representation factors value into a predictive component (expected future state occupancies) and a reward component, allowing the predictive structure to transfer when rewards change but dynamics remain the same (Dayan, 1993).

These three forms of reusable component (temporally extended skills, relational templates, predictive maps) differ in format but share a common economic logic: each earns its place

by reducing the cost of future problems. Correa et al. (2023) formalize this from a resource-rational perspective, and Lieder and Griffiths (2020) provide the broader framework: task decompositions are worth maintaining insofar as they reduce overall computational cost under bounded rationality. The investment in building an abstraction is amortized across future problems, which means an option or template is worth learning only if the structure it captures is likely to recur.

This amortization logic reveals a deep connection to the continual learning problem. Catastrophic forgetting occurs because standard deep networks store knowledge in entangled distributed weights that do not support modular extension (McCloskey and Cohen, 1989; French, 1999; Pan et al., 2025). The root cause is not a deficient learning algorithm but a deficient knowledge *format*. A system storing knowledge in modular components with defined interfaces can extend its repertoire by adding components without altering existing ones. Adding an option to a library does not change existing options. Adding a new object type does not alter previously learned interaction rules.

This observation has a testable implication: the degree of catastrophic forgetting should depend on the *modularity* of knowledge representation, not only on the learning algorithm. McCourt et al. (2023) showed that modularity can emerge from selection pressure for fault tolerance but that standard backpropagation does not discover modular structure even when the underlying task is modular. Clune et al. (2013) made a complementary evolutionary argument: modularity arises when the environment itself has modular structure and there is a cost for network connections, because modular architectures can reconfigure subnetworks without disrupting unrelated function. It is worth noting that modern deep learning has itself moved toward more modular formats: Mixture of Experts architectures (Shazeer et al., 2017), progressive neural networks (Rusu et al., 2016), and tool-use libraries in large language models all represent attempts to support extension without interference. That these engineering solutions converge on the same design principle, modular components with defined interfaces, is itself evidence for the format hypothesis rather than a refutation of it. The problem is that routing and retrieval across modules remains a bottleneck, and end-to-end training still tends to entangle what modular design tries to separate. McClelland et al. (1995)'s CLS theory can be read in this light: the two-system architecture prevents catastrophic interference not only through separated learning rates but, this essay argues, because the slow system produces *modular, separately addressable library entries* rather than a single entangled representation.

The constraint is that the knowledge format must support modular extension. Without modularity, every new piece of learning risks degrading existing knowledge, and continual learning becomes a problem to be patched rather than a capacity that follows from good architectural design.

3.4 Vocabulary expansion

All three constraints above presuppose that the agent already has the right representational vocabulary. But in a complex world, the agent will encounter regularities that cannot be captured by any existing primitives. Options presuppose a state space. POMDPs presuppose

a generative model. Structure induction presupposes a grammar. What happens when the vocabulary itself is insufficient?

Bayesian nonparametric methods allow the *number* of components to grow with data: the Chinese Restaurant Process (CRP) for categories (Anderson, 1991), the Indian Buffet Process for features (Griffiths and Ghahramani, 2011), sticky hierarchical Dirichlet process hidden Markov models (HDP-HMMs) for sequential states (Fox et al., 2008). Gershman and colleagues applied related ideas to latent causes and event-like structure, proposing that the brain may use something functionally analogous to a CRP when deciding whether a new experience signals a familiar or novel context. These models are attractive because they provide principled ontology expansion, but what grows is the *number* of components, not the *type*. A CRP can create the 101st cluster, but that cluster has the same format as the first. A further concern for embodied agents is that standard nonparametric models assume exchangeability (the data partition does not depend on observation order), which is violated when the agent's actions change the data-generating process.

Program induction takes a different approach, rooted in the broader proposal that cognition can be understood as inference over structured generative programs. Goodman et al. (2008) formalized this idea by showing that concept learning, causal reasoning, and language understanding can be cast as probabilistic inference over programs in a stochastic language of thought. Lake et al. (2015) demonstrated its power most dramatically: their Bayesian Program Learning (BPL) system acquired new handwritten character concepts from single examples by inferring short motor programs, matching human-level generalization in a domain where standard deep learning required orders of magnitude more data. The key insight is that compositionality is built into the representational language itself, so that new concepts are assembled from reusable parts (strokes, sub-routines, causal primitives) rather than learned as monolithic patterns. DreamCoder (Ellis et al., 2020) extended this into an explicit library-growth mechanism: during waking, it searches for programs solving tasks using current primitives; during sleep, it compresses recurring sub-programs into new library entries that change the hypothesis space for future problems. Theory-based RL (Tsividis et al., 2021) extends the same logic into sequential decision-making, learning executable generative theories that function as causal models for planning. These systems make library growth explicit. But the base language must be specified in advance: DreamCoder can compose existing primitives in new ways but cannot create atomic operations not already in its foundation, and BPL's stroke primitives are hand-designed for the character domain.

Object-centric approaches occupy a middle ground. The object-centric model of Heins et al. (2025) treats scenes as compositions of objects in a hierarchical generative model, expanding when unexplained observations demand a new entity and contracting via Bayesian model reduction when an entity proves unnecessary. The Renormalising Generative Model (Friston et al., 2024) builds hierarchical representations through principled coarse-graining inspired by statistical physics, growing depth as the data's multi-scale structure demands.

In every case, growth occurs within a fixed meta-vocabulary: a grammar of component types, a base language, a compositional format for objects. The constraint, and the deepest open problem, is that any library-building system must eventually confront the limits of its

own representational language. How genuinely new vocabulary enters a system, as opposed to new compositions of existing primitives, remains unsolved by any current approach.

3.5 Causal robustness

The four constraints above say nothing about what makes an abstraction robust under change. A representation that predicts well in one environment may break under intervention or distributional shift. The Independent Causal Mechanisms principle (Peters et al., 2017; Schölkopf et al., 2021) provides the normative answer: the mechanisms by which causes produce effects are autonomous modules that can be independently intervened upon. If an agent's library entries align with these independent mechanisms, then the library is modular in the strongest possible sense, and updating one entry (because its mechanism has changed) does not require updating others.

This is a compelling picture, but it faces a fundamental epistemic obstacle. Learning which variables are causally related requires interventional evidence: the agent must manipulate variables and observe consequences. Without interventions, causal direction is generally unidentifiable from observational data alone (Pearl, 2009). This means the causal depth of an agent's library is bounded by its behavioural repertoire. An agent can only learn causal structure for the parts of the world it can actually manipulate. Pearl's causal hierarchy (Pearl and Mackenzie, 2018) formalizes this: observational, interventional, and counterfactual knowledge form a strict hierarchy, with higher levels generally underivable from lower levels alone.

A connection to empowerment is instructive here (Klyubin et al., 2005). An agent with high empowerment (high mutual information between actions and future states) has a rich repertoire of actions producing diverse, distinguishable consequences. Such an agent is better positioned to learn causal structure because each action is effectively an intervention generating evidence about how the world responds. The developmental psychology literature emphasizes the same point: children's exploratory play is not random but structured by the informativeness of actions, with children preferentially acting on objects where causal structure is uncertain (Gopnik, 2012; Bonawitz et al., 2012).

A practical question is whether full causal faithfulness is necessary for a useful library. Recent world models suggest useful abstraction can emerge from learning to predict and simulate without explicit causal graphs (Hafner et al., 2023; LeCun, 2022). But there are reasons to think something more is needed for robust transfer. The position this essay adopts is that the most useful abstractions are *causally graded*: modular enough to be updated independently, interaction-sensitive enough to capture how entities affect one another, and stable enough under moderate distributional shift for reuse in new contexts, without necessarily satisfying the full Beckers–Halpern criterion under arbitrary interventions.

The constraint is that the agent's active engagement with its environment is not external to the abstraction problem but constitutive of it. What an agent can abstract depends on what it can do.

4 Tensions Between the Requirements

The five constraints identified in Section 3 are not independent items on a checklist. They interact, and in several cases they pull against each other. A system that prioritizes one may find it harder to satisfy another. The result is a set of architectural trade-offs that any solution, biological or artificial, must navigate. These trade-offs are not incidental engineering difficulties; they reflect the structure of the problem itself. This section identifies four such tensions and derives predictions from each.

4.1 Specificity vs. reusability

The system needs richly detailed episodic material for future structural comparison (especially Sections 3.1 and 3.4) but it also needs compressed, general representations for current inference and control (Section 3.3). These pull in opposite directions. If the system discards episodic specificity too early, it loses the relational detail from which new abstractions could later be extracted. If it retains too much, memory fills with undigested particulars that are expensive to store and slow to search.

Both episodic specificity and abstract generality are needed simultaneously. The system must maintain representations at multiple levels of specificity, and must have operations that convert specific episodes into more general structure when evidence warrants it. This has a direct architectural consequence: the system should have at least two interacting representational formats, one preserving structured particulars and one stabilizing commonalities into portable form. CLS theory (McClelland et al., 1995) proposed this separation to avoid catastrophic interference. The present analysis goes further in two respects.

First, the specific format is not passive storage. Episode encoding is already shaped by existing knowledge and inferential demands. An agent with a schema for the current situation encodes differently from one without. The two formats are coupled from the outset: the library shapes how new episodes are encoded, which shapes what future generalizations are possible.

Second, the balance between specificity and generality shifts across the agent's lifetime. Early in learning, useful information is mostly in the episodic store. Later, the agent relies more on deploying existing abstractions. But a mature agent cannot stop encoding specific episodes, because new regularities may require fresh episodic material to detect. The tension is managed dynamically, never fully resolved.

Prediction (formation–deployment dissociation). Disrupting hippocampal encoding should impair formation of new abstractions for novel domains without impairing deployment of existing abstractions in already learned domains. Critically, the impairment should scale with how much *new structural comparison* is required: tasks solvable by deploying an existing schema should be relatively spared, while tasks requiring a new schema from episodic comparison should be specifically impaired. This goes beyond the classical finding that hippocampal damage impairs memory, because it predicts a graded interaction between lesion effects and the structural novelty of the task. Conversely, disrupting prefrontal

schema representations should impair deployment of existing abstractions while leaving basic episodic encoding partly preserved.

4.2 Stability vs. revisability

Library entries earn their value through stability: a schema for restaurant visits must persist reliably and resist corruption by noise. But the world changes. If the agent cannot revise its abstractions when evidence warrants it, it becomes rigidly committed to structure that no longer matches reality. Recent work illustrates this directly: persistent orbitofrontal representations of a prior schema can shape subsequent learning, including learning when new demands conflict with earlier structure (Maor et al., 2026). Schemas are simultaneously assets and liabilities, depending on fit.

This is the stability–plasticity dilemma (Grossberg, 1980), restated in library vocabulary. The library framing adds specificity that the classical formulation lacks. What must be stable is the structural integrity of individual library entries, their interfaces and domains of applicability. What must be plastic is the system's ability to revise, split, or replace specific entries when accumulated evidence warrants it. The architectural consequence is that the system must respond differently to different magnitudes of mismatch between new evidence and existing structure. Small mismatches should trigger parameter updating within the existing structure. Large mismatches should trigger structural revision: creating a new schema, splitting an old one, or replacing an outdated model.

This distinction echoes Piaget's (1952) developmental framework of assimilation and accommodation (Piaget, 1952). Assimilation incorporates new experience into existing cognitive structures. Accommodation modifies the structures themselves when assimilation fails. What the present analysis adds is the prediction that these are not merely two names for "easier" and "harder" learning. They should correspond to qualitatively distinct processing dynamics with a detectable transition between them.

The timescale-separated inference framework discussed earlier provides a natural implementation: the agent holds model structure approximately fixed while adjusting parameters, and revises structure only when accumulated evidence of parameter-level failure becomes sufficiently strong (Da Costa et al., 2020; Costa et al., 2025). This is approximately what Bayesian model comparison does: parameters are updated continuously given the current model, but the model is revised only when its marginal likelihood drops significantly relative to an alternative.

Prediction (assimilation threshold). Schema-violating experiences should induce a qualitative transition between two processing modes rather than merely a continuous gradient. Below the threshold, the existing schema remains in force with locally adjusted predictions. Above it, the system begins entertaining alternative structural hypotheses. The threshold should be modulated by schema strength: a well-established schema supported by many episodes should require more counter-evidence to trigger structural revision than a recently formed one. Standard paradigms use only two conditions (congruent vs. incongruent). Testing this prediction requires parametrically varying violation magnitude to search for a

transition point.

4.3 Commitment vs. uncertainty

Effective action requires treating current abstractions as approximately correct. An agent that maintains full uncertainty over whether its schema is the right one, continuously entertaining radical alternatives, will be slow and indecisive. But effective learning requires precisely this openness. An agent that fully commits to its current model and refuses to consider alternatives cannot recognize when it is wrong and cannot update its library.

This tension is related to, but not identical with, the classical exploration–exploitation trade-off. Exploration–exploitation concerns which *actions* to take given a fixed model of the world. The commitment–uncertainty tension concerns which *internal model* to use when interpreting observations and planning actions. The two can come apart in practice. Consider an agent navigating a building it believes to be a hotel. It can *explore* (try opening unfamiliar doors) while remaining fully *committed* to its hotel schema, interpreting every observation through that frame. Conversely, it can *exploit* (head directly for the exit using its best current route) while maintaining genuine *uncertainty* about whether the building is a hotel or an office complex, ready to switch schemas if evidence accumulates. The first case involves action-level exploration with model-level commitment. The second involves action-level exploitation with model-level uncertainty. Exploration–exploitation theory addresses the first dimension; the commitment–uncertainty tension addresses the second. In practice they interact, but they are not the same trade-off, and conflating them obscures the specific architectural demand that model-level uncertainty places on the system.

The computational implication is that the system should separate deployment from revision in time. Deployment requires commitment: selecting a schema, treating it as correct, and using it to guide inference and action in real time. Revision requires uncertainty: comparing multiple episodes against existing structure and considering alternatives. These two modes have different computational requirements. Deployment must be fast and deterministic enough for coherent action. Revision can be slower and should consider multiple hypotheses.

This maps naturally onto the difference between online and offline processing. Online behaviour depends on committed internal models deployed for moment-to-moment decisions. Offline processing during rest and sleep involves reactivation, comparison, and potential reorganization of representational structure (Wilson and McNaughton, 1994; Peyrache et al., 2009; Yang et al., 2024). The online/offline distinction is, on this reading, not merely a consequence of the brain needing rest. It is a functional solution to the commitment–uncertainty tension: the system commits during behaviour and revises during consolidation, separating the two in time to prevent mutual interference.

There is an important interaction with Section 4.2. The offline processing that supports revision is the same processing that must evaluate whether a mismatch is small (triggering parameter updating) or large (triggering structural revision). The commitment–uncertainty and stability–revisability tensions are linked through the consolidation process, which serves both as the context for entertaining uncertainty and as the mechanism for evaluating the

assimilation–accommodation threshold.

Prediction (offline–online double dissociation). Disrupting offline consolidation should impair the quality of future abstractions without affecting deployment of existing ones. Disrupting online prefrontal representations should impair current schema-guided behaviour without preventing later consolidation of recently encoded episodes. The critical feature is that each process matters for something specific: offline processing for future abstraction quality, online deployment for current use. If sleep disruption impaired item memory and structural generalization equally, the specific link between offline processing and abstraction formation would not be supported.

A compound prediction follows from the interaction with Section 4.2: disrupting offline processing should impair *both* the detection of structural mismatch *and* the construction of new abstractions, a compound effect not predicted by either tension alone.

4.4 Causal depth vs. epistemic cost

Moving from predictive adequacy through control relevance to interventional stability requires progressively more expensive evidence. Predictive adequacy can be assessed from passive observation. Control relevance requires acting. Interventional stability requires targeted manipulation across contexts. The agent faces a resource allocation problem: how much of its limited behavioural budget should it invest in deepening causal depth versus broadening its library with new (but shallow) abstractions versus exploiting what it already has?

This adds a dimension beyond classical exploration–exploitation: even within exploration, the agent must decide *what kind* of information to seek. Passive exploration discovers statistical regularities. Active probing discovers causal structure. The latter is more expensive but produces more robust abstractions. This tension overlaps with bounded-rationality accounts of exploration for robustness (Lieder and Griffiths, 2020), and the essay should be transparent about this. What the present framing specifically adds is the connection between the agent's action repertoire and the achievable causal depth of its abstractions: the claim that what an agent *can do* constrains what it can *know* about causal structure, and therefore constrains the robustness of its library under distributional shift. Generic bounded-rationality frameworks treat exploration cost as a scalar resource. The causal-depth tension treats it as a structural constraint that shapes the *kind* of knowledge the agent can acquire, not just the *amount*. An empowerment-maximizing agent (Klyubin et al., 2005) is, in effect, seeking out the interventional evidence needed for causally deep abstractions, because it preferentially acts where its actions make the most difference to future sensory states.

The consequence is that the library's causal quality will inevitably be uneven. Parts of the world the agent has actively probed will be represented by causally deeper abstractions. Parts only passively observed will be shallower and more fragile. This unevenness is not a deficiency but a rational response to bounded resources. It connects to the active inference treatment of epistemic value (Friston et al., 2015; Da Costa et al., 2020): expected free energy decomposes into pragmatic and epistemic components, and an agent balancing both is

implicitly navigating the causal-depth tension.

Prediction (active exploration enhances transfer under distributional shift). Agents that actively explore through interventional probing should develop abstractions more robust under distributional shift than agents that passively observe equivalent environments. The prediction is not simply that active learners learn more. It is that active and passive learners should perform *comparably* when tested in the training environment but *diverge* specifically when tested in a shifted environment, because active exploration provides the interventional evidence needed for causal depth while passive observation does not. Behaviourally, one could compare transfer performance after active versus passive training, with the shift manipulation as the critical test. Neurally, hippocampal–prefrontal representations in actively exploring agents should show less reorganization when the environment shifts, because the underlying abstractions are less dependent on specific contextual features.

5 The Hippocampal–Prefrontal System as an Architecture Shaped by These Tensions

The preceding sections developed the argument computationally. This section turns to biology and asks whether the hippocampal–prefrontal system can be productively read through the lens of the tensions identified in Section 4.

The claim is modest and should be stated precisely. The essay does not argue that the hippocampal–prefrontal system was designed to solve the library-building problem, nor that the three-stage architecture proposed below is the best or only interpretation of the evidence. The claim is that the tension framework provides a useful *vocabulary* for organizing converging findings that span the schema, replay, task-state, and cognitive-map literatures, and that this vocabulary generates predictions not derivable from any of these literatures taken individually. Biological systems are shaped by evolutionary pressures, developmental constraints, and historical contingency. The three-stage architecture is an interpretive framework, not a uniquely determined reading of the data.

The evidence is organized around a three-stage architecture: (1) structured episode capture, (2) selective comparison and alignment, and (3) stabilization and deployment. These stages are not strictly sequential; they interact continuously. Stabilized abstractions from Stage 3 shape how new episodes are encoded in Stage 1, and accumulated mismatch in Stage 2 can trigger revision of Stage 3 entries. The assignment of stages to brain regions is approximate, and recent evidence complicates any simple division of labour.

5.1 Structured episode capture

The first stage addresses the specificity side of Section 4.1. The system needs a mechanism for rapidly encoding relationally structured episodes preserving enough detail for later comparison.

The classical evidence centres on the hippocampus. Bilateral medial temporal lobe damage produces inability to form new declarative memories despite preserved intelligence

and remote memory (Scoville and Milner, 1957). CLS theory explains why this fast-encoding system must be distinct from the system storing generalized knowledge: a distributed cortical system learning gradually is vulnerable to catastrophic interference if forced to learn new episodes too quickly (McClelland et al., 1995). The hippocampus solves this by rapidly binding specific conjunctions in recent events, while slower cortical learning can later extract stable regularities without destabilization.

Critically, hippocampal encoding is not mere item storage but *structured binding*. The cognitive-map tradition established that hippocampus represents relations among elements of experience, originally spatial (O'Keefe and Nadel, 1978) but now extended to temporal, social, and conceptual structure (Eichenbaum, 2017; Howard and Eichenbaum, 2015). This encoding is shaped by existing knowledge. Tse et al. (2007) demonstrated that rats with a pre-existing spatial schema learned new paired associations within that layout in a single trial, whereas rats without the schema required many trials. The schema did not merely aid retrieval; it accelerated encoding itself.

Recent evidence complicates the simple narrative in which hippocampus encodes specifics while cortex encodes abstractions. Courellis et al. (2024) recorded single neurons in human hippocampus during an inferential reasoning task and found that hippocampal populations encoded multiple task variables simultaneously in an abstract, disentangled format. This geometry appeared specifically when patients could perform inference. Samborska et al. (2022) found a complementary pattern in mice: prefrontal cortex carried representations of common structure across a family of tasks, while hippocampus mapped that structure onto the sensorimotor specifics of the current situation. El-Gaby et al. (2024) extended this further, showing that mice discovered shared task structure and exploited it for zero-shot inference in novel problems.

The more accurate characterization is therefore that hippocampal representations have a level of abstraction that is not fixed but depends on inferential demands. When the task requires binding specific items to contexts, representations are episode-specific. When it requires inference across contexts, they can become abstract. Even Stage 1 is not rigidly confined to one side of Section 4.1.

There is an alternative interpretation that should be acknowledged. Hippocampal encoding might serve primarily online state inference (constructing belief states for immediate action, as in Section 3.2) rather than providing raw material for later offline abstraction. These functions are not mutually exclusive, and the same encoding operation may serve both. An episode encoded for current state inference also becomes available for later replay and structural comparison. The relative weight of these two functions remains debated.

5.2 Selective comparison and alignment

Stage 2 is where a memory system begins to behave like a library-building system. In terms of the tensions, it is the most demanding stage: it must navigate all four. It selects which episodes are worth retaining (Section 4.1). It evaluates mismatch between new episodes and existing structure (Section 4.2). It operates without interfering with committed deployment of existing abstractions (Section 4.3). And by comparing episodes across contexts, it can begin

testing whether structural commonalities hold beyond original conditions (Section 4.4).

The leading biological candidates are replay, ripples, and offline consolidation. Wilson and McNaughton (1994) established that neural patterns from waking behaviour are re-expressed in hippocampal ensembles during subsequent sleep. Peyrache et al. (2009) showed that replay extends to prefrontal circuits, with activity patterns related to recently acquired rule learning replaying during slow-wave sleep. This is important because it shows that what is reactivated offline includes rule-related structure encoded in prefrontal circuits, not just raw sensory trajectories, providing a substrate for the comparison operation Stage 2 requires.

Replay is selective. Yang et al. (2024) provided direct evidence that awake hippocampal sharp-wave ripples function as a selection mechanism. Recording from thousands of Cornu Ammonis area 1 (CA1) neurons in mice, they found that ripple content decoded specific trial blocks, and the blocks most strongly represented in awake ripples were preferentially reactivated during subsequent sleep. Awake ripples acted as a tagging system marking certain experiences for later consolidation. The selection criterion is not fully characterized, but the existence of a selection mechanism is itself important: a system that selectively retains structurally informative episodes while discarding redundant ones is better positioned for efficient library building.

The most informative recent finding for the present analysis comes from Xiao et al. (2025), who recorded intracranial activity from human epilepsy patients performing a non-spatial inference task organized by an abstract two-dimensional conceptual structure. Hippocampal ripples during brief rest periods predicted the emergence of grid-like representational codes, and these ripples supported the alignment of newly learned experiences with the latent relational structure. This is as close as current evidence gets to a direct demonstration of what Stage 2 proposes: newly encoded episodes are not merely replayed as isolated traces but are repositioned relative to an existing structural scaffold during offline processing.

Schwartenbeck et al. (2023) provide a complementary perspective, showing that generative replay in hippocampal–prefrontal circuits can support compositional inference and hypothesis testing about structure. On their account, replay functions not merely as rehearsal but as model construction, which connects directly to the library-building function proposed here.

Complementary consolidation evidence supports alignment: Tompary and Davachi (2017) showed that consolidation increased representational overlap across related memories, and Audrain and McAndrews (2022) showed that pre-existing schemas scaffold neocortical integration of congruent memories while hippocampal representations preserve episode-specific distinctions.

A conservative alternative must be acknowledged: replay might primarily strengthen traces, with representational convergence reflecting passive detail loss rather than active alignment. The most discriminating test is that ripple disruption should selectively weaken structural generalization while sparing simple item memory, a dissociation that pure trace-strengthening accounts do not predict.

5.3 Stabilization and deployment

Stage 3 concerns what happens once recurring structure becomes durable enough to guide future cognition. The system stabilizes it into representations deployable without reconstructing them from original episodes. This addresses the reusability side of Section 4.1 and the stability side of Section 4.2.

The neural evidence centres on prefrontal cortex, particularly ventromedial prefrontal cortex (vmPFC) and orbitofrontal cortex (OFC). Gilboa and Marlatte (2017) argue that vmPFC is centrally involved in schema instantiation and schema-guided processing. van Kesteren et al. (2010, 2012) showed that schema-congruent encoding especially engages medial prefrontal cortex, and Spalding et al. (2015) showed that vmPFC damage disrupts this function, establishing necessity rather than mere correlation.

Schuck et al. (2016) showed that human OFC encodes hidden task states inferred from history, with better decoding predicting better performance. Schiereck et al. (2025) provided causal evidence: inactivating rat OFC specifically impaired belief updating about hidden states without eliminating simpler strategies, moving the evidence from correlation to causation.

These two literatures share a common functional profile, but the unification should not be overstated. Schema research in vmPFC emphasizes durable, event-level knowledge structures accumulated across many experiences and deployed to interpret new situations. Task-state research in OFC emphasizes compact, real-time models of hidden variables constructed for ongoing decision-making. These may involve different computational operations (template matching versus online inference), different timescales of formation, and different degrees of durability. Moneta et al. (2024) reviewed evidence that vmPFC and OFC representational spaces are richer and more heterogeneous than either pure value codes or simple map metaphors. What the two literatures share, and what justifies treating them as related for the purposes of this essay, is that both involve prefrontal cortex trafficking in reduced but behaviourally meaningful structure: representations that are more compact than episode lists yet more structured than generic summaries, preserving specifically those commonalities and hidden-state distinctions most useful for interpretation and control. The present essay treats this shared functional profile as the relevant level of description for the library framework, while acknowledging that the computational details may differ substantially between the two.

Section 4.2 predicts that stability must coexist with revisability. Persistent orbitofrontal representations of a prior schema can shape later learning, including learning when new demands conflict with earlier structure (Maor et al., 2026). This is broadly consistent with the idea that a stabilized abstraction can be an asset when the world conforms to it and can also bias how later evidence is incorporated. How the brain detects when accumulated mismatch warrants structural revision, rather than parameter updating, remains among the most important open questions.

Regarding causal depth, the evidence most clearly supports *control relevance* in the sense of Section 2.3. Few studies have tested whether the same prefrontal representations transfer

robustly across structurally similar but perceptually distinct domains. The most suggestive evidence comes from Samborska et al. (2022) and El-Gaby et al. (2024), but direct evidence for robustness under distributional shift remains sparse. Assessing the causal depth of prefrontal abstractions is one of the field's most important empirical priorities.

5.4 The loop, the format, and what the architecture adds

The three stages form a recurrent loop, not a pipeline. Stabilized abstractions shape future encoding (van Kesteren et al., 2012). This recurrence creates a genuine and unresolved problem that the tension framework must confront honestly rather than absorb as a minor failure mode. If encoding is shaped by prior schemas (as Sections 2.1 and 5.1 establish), and if the system relies on richly structured episodes to discover new abstractions (as Section 3.1 requires), then there is a circularity: schema-driven encoding can suppress exactly the anomalous details needed for schema revision. This is not merely the well-documented phenomenon of schema-consistent memory distortion (Bartlett, 1932; Ghosh and Gilboa, 2014). It is a structural tension within the framework itself: the same mechanism that makes encoding efficient (prior-driven selection) is the mechanism that makes it blind to the evidence needed for structural revision. A strong schema produces episodes already partially aligned with itself, which replay is then more likely to confirm than challenge, reinforcing the schema further. Breaking out of this confirmation loop requires encounters sufficiently anomalous to resist schema-driven encoding and survive into replay with enough unconforming detail to trigger structural revision. The Yang et al. (2024) finding that awake ripples selectively tag certain experiences is suggestive here: if the tagging mechanism is sensitive to novelty or prediction error rather than only to reward proximity, it could serve as a partial corrective.

A stronger candidate mechanism comes from the neuromodulatory literature. Acetylcholine levels in the hippocampus are elevated during novel or uncertain situations, and high cholinergic tone shifts hippocampal processing from a pattern-completion mode (which favours schema-driven encoding, filling in expected structure) toward a pattern-separation mode (which favours high-fidelity encoding of the current input, including details that deviate from schematic expectations) (Hasselmo, 2006; Hasselmo et al., 1995; Douchamps et al., 2013). This provides a biologically grounded gate: when prediction error is low and the schema fits, cholinergic tone is low and encoding is schema-dominated. When prediction error is high and the schema is failing, cholinergic tone rises and forces the hippocampus into a less compressed, more faithful encoding mode that captures the anomalous details needed for later structural revision. This does not fully resolve the encoding paradox, because moderate mismatches (large enough to matter but too small to trigger strong cholinergic responses) may still be suppressed by schema-driven encoding. But it provides a biologically plausible partial solution and generates a testable prediction: pharmacological manipulation of cholinergic tone during encoding should modulate the degree to which schema-inconsistent details survive into later replay and structural generalization.

A point that deserves emphasis concerns the *format* of the architecture's outputs. Section 3.3 argued that the central failure of deep learning in continual settings is a knowledge-

format problem, not a learning-algorithm problem. The three-stage architecture produces knowledge in a different format. Schemas, task-state models, and relational maps function as modular components with at least partially defined domains of applicability. A schema for restaurants does not govern behaviour at a swimming pool. Adding a new schema does not automatically corrupt existing ones. This modularity is what makes continual learning structurally possible: new learning extends the library without degrading it. The biological separation between fast hippocampal encoding and slow cortical stabilization prevents catastrophic interference not only through separated learning rates, but because the slow system produces modular, separately addressable components.

Whether this modularity is as clean in practice as the argument suggests is an open question. Prefrontal representations are not neatly compartmentalized, and the same voxels may participate in multiple schemas (Spalding et al., 2015). More fundamentally, the computational neuroscience of prefrontal cortex has emphasized non-linear mixed selectivity: individual neurons encode combinations of task variables in a high-dimensional, distributed format that supports flexible readout precisely because representations are entangled rather than modular (Rigotti et al., 2013; Fusi et al., 2016). This appears to contradict the "separately addressable library entries" claim, and it raises a pointed question: if the brain also stores knowledge in entangled, distributed neural weights, why does it avoid the catastrophic forgetting that plagues artificial networks using the same basic format?

The resolution is that the modularity the present analysis requires is *geometric*, not anatomical. Recent work on population coding has shown that the brain can maintain functionally independent representations within shared neural populations by routing different task contexts into near-orthogonal subspaces of population activity (Bernardi et al., 2020; Tang et al., 2020). When two schemas occupy sufficiently separated manifolds in the same neural population, activating one need not strongly interfere with the other, because the activity patterns are geometrically separated even though they use the same physical neurons. This provides, at least in principle, a concrete mechanism for "composability without interference" that is compatible with mixed selectivity: the same neurons participate in multiple representations, but the population-level geometry can keep those representations functionally distinct. The critical claim, then, is not that the brain avoids distributed representation (it does not) but that hippocampal–cortical consolidation may produce representations with this orthogonal geometric property more reliably than standard end-to-end backpropagation does. This gives the format hypothesis a testable empirical signature: the degree to which consolidated prefrontal representations occupy orthogonal subspaces should predict the degree to which the system resists cross-task interference.

An honest assessment of what this reading adds beyond existing frameworks. The three-stage structure is substantially indebted to CLS theory, and the essay should not obscure this. What is genuinely new is narrow but, this essay argues, consequential: (a) the tension framework, which identifies *why* the stages must have their specific functional properties and generates predictions (dissociations, thresholds, format-dependent continual learning) not derivable from CLS or schema theory alone; (b) the format hypothesis, which proposes

that the brain's continual-learning advantage depends not only on separated learning rates but on the geometric modularity of the resulting knowledge format, with a specific empirical signature (orthogonal population subspaces predicting resistance to cross-task interference) that could confirm or refute it; and (c) the identification of the encoding paradox as an open vulnerability, which shows where the architecture's own design principles work against each other. These are additions to an existing research programme, not a replacement for it. The testable predictions below provide one criterion for evaluating whether the additions are worth making.

5.5 Falsification conditions

The framework would be substantially undermined if any of the following turned out to be true.

First, and most fundamentally, if a single monolithic online learning system, without any separation between episodic encoding, offline comparison, and schematic stabilization, achieved equivalent library-like function: structural transfer across domains, continual learning without catastrophic forgetting, and flexible deployment of reusable abstractions. This subsumes a more specific version: that flexible generalization turned out to be fully achievable without any form of offline structural comparison, showing Stage 2 to be unnecessary. The strongest version of this test would be a system that learns continually from a stream of diverse tasks, transfers relational structure to novel domains, and does not degrade on old tasks, all without any form of replay, consolidation, or modular knowledge separation. If such a system existed and matched biological performance, the architectural claims of this essay would be unwarranted.

Second, if disruptions to different components produced unsystematic, undifferentiated impairments rather than the specific dissociations predicted in Section 4. The framework claims that different components serve different functions in navigating different tensions. If lesions produced only generic deficits, the functional specificity the tensions depend on would not be supported.

Third, if prefrontal representations that support schema deployment turned out to be entirely unable to transfer across structurally similar but perceptually distinct domains. The framework's strongest claim about Stage 3 is that stabilized abstractions are portable. If they were entirely bound to specific stimuli and contexts, the library label would be unwarranted.

The most discriminating single test is the offline–online dissociation prediction from Section 4.3: disrupting hippocampal ripple-linked replay should impair the ability to extract shared relational structure across episodes, even when simple item memory is relatively spared. This dissociation between structural generalization and item retention is the critical element, and it is important to state why it is not already predicted by existing frameworks. Standard systems consolidation models predict that ripple disruption impairs *overall* memory strength; they do not predict a differential effect on structural generalization versus item retention. The Tolman–Eichenbaum Machine (Whittington et al., 2020) provides a computational account of how relational structure might be factored, but does not make specific predictions about ripple disruption effects. Schema theory

(Tse et al., 2007) predicts that schemas accelerate encoding, but does not predict that disrupting replay would specifically impair future structural transfer while sparing simple item consolidation. The present framework makes this specific dissociation prediction because it assigns replay a particular functional role (structural comparison and alignment, not merely trace strengthening) that the other frameworks do not.

6 Relation to Existing Frameworks

This section makes the relationship to four existing frameworks explicit, stating honestly where the present analysis adds something and where it may be subsumed.

Complementary Learning Systems. CLS theory (McClelland et al., 1995) is the most direct precursor to the architecture proposed in Section 5. It established the foundational insight that fast hippocampal encoding and slow cortical integration must be separated to prevent catastrophic interference, and provided a computational account of why this separation is necessary. The present analysis is deeply indebted to CLS and should be understood as building on it rather than replacing it.

The present analysis adds specificity on two points CLS left open. First, the *criterion* for what gets consolidated: Section 4.2 predicts sensitivity to the degree of mismatch between new episodes and existing structure, with small mismatches triggering parameter-level assimilation and large mismatches triggering structural revision. CLS explains that replay-mediated interleaving gradually integrates new traces into cortical representations, but it does not specify what determines which traces are prioritized or how the system decides when to assimilate versus when to create new structure. The tension framework provides a candidate answer. Second, the *format* of consolidated knowledge: the present analysis adds the claim that the slow system produces modular library entries, not merely a slow-learned distributed representation, and that this format (implemented via orthogonal population geometry, as discussed in Section 5.4) is what makes extension without interference structurally possible. CLS explains why two learning rates are needed; the present analysis adds that what the slow system *builds* matters as much as how fast it builds it.

Modern descendants of CLS have narrowed the distance between the original theory and the present analysis considerably. CLS+ (Kumaran et al., 2016) incorporates replay-based generalization and structure learning, moving beyond the original focus on avoiding interference toward a more active role for consolidation in extracting regularities. The Tolman–Eichenbaum Machine (Whittington et al., 2020) provides a specific computational model of how hippocampal representations can factorize sensory content from relational structure, supporting structural transfer across domains. These developments address some of the same concerns as the present essay, and the present analysis is best understood as a further development within the CLS lineage rather than an alternative to it. What it specifically adds is the tension framework as an organizing vocabulary, the format hypothesis as a specific and testable proposal about why modularity matters, and the encoding paradox as an identified vulnerability that CLS and its descendants do not foreground.

Probabilistic programs and library learning. The programme of treating cognition as inference over probabilistic programs (Goodman et al., 2008; Lake et al., 2015) provides the most explicit computational account of what it means to learn with a compositional, reusable representational language. DreamCoder (Ellis et al., 2020) extends this into a library-growth mechanism that most directly parallels the present essay's concerns: it not only solves tasks but expands its own primitive vocabulary through compression of recurring structure. Theory-based RL (Tsividis et al., 2021) brings these ideas into sequential decision-making, showing that agents equipped with rich intuitive theories—essentially executable generative programs encoding object physics and goal-directed reasoning—can learn to play novel video games from very few examples by leveraging causal structure for planning and exploration. This demonstrates that compositional program-like representations can support not just classification but active, goal-directed control in novel environments. The present analysis agrees on the centrality of compositional library growth but diverges on three points. First, raw material: probabilistic program learners work with symbolic solution traces or pre-structured game descriptions; the biological system works with noisy relational episodes from continuous, partially observable experience. Second, the present analysis adds the constraint that causal depth is limited by the behavioural repertoire (Section 4.4), which does not arise when tasks and their causal structure are given rather than discovered. Third, the biological system faces the encoding paradox (Section 5.4) in a way that program induction systems do not, because the biological system's prior knowledge actively shapes what episodes are available for future abstraction. A shared gap is the grounding problem: program induction produces explicit compositional structure but struggles to ground it in high-dimensional sensory experience; neural world models do the opposite.

Active inference. The active inference framework (Friston, 2010; Da Costa et al., 2020; Parr et al., 2022) provides a comprehensive account unifying perception, learning, and action under variational free energy minimization within hierarchical generative models. It supplies several ingredients this analysis draws on: multi-scale hierarchical structure in which higher levels represent slower, more abstract aspects of the environment; the coupling of inference and control through expected free energy, which naturally balances epistemic exploration and pragmatic exploitation; and structure-learning mechanisms through Bayesian model reduction, which provides a principled criterion for when to simplify the model by pruning unnecessary components (Friston et al., 2018). The object-centric model of Heins et al. (2025) and the Renormalising Generative Model are direct applications of these principles to the abstraction-growth problem (Heins et al., 2025; Friston et al., 2024), providing concrete demonstrations of how object-centric ontologies and multi-scale hierarchies can expand and contract in response to evidence.

The relationship requires transparency. Active inference is arguably more general than the present framework. It provides a unified mathematical language for inference, learning, action, and structure revision, whereas the present analysis identifies specific tensions and proposes a biological reading without committing to a single formalism. It is possible that the present analysis is subsumable within active inference, in the sense that the four

tensions could in principle be derived as consequences of free energy minimization within an appropriately structured hierarchical model. The timescale-separated inference framework discussed earlier, which distinguishes state, parameter, and structure inference, comes most explicitly from recent work within this tradition (Costa et al., 2025), while remaining continuous with the broader active-inference treatment in Da Costa et al. (2020). The epistemic–pragmatic decomposition of expected free energy, which this essay invokes in discussing Section 4.4, is likewise native to active inference (Friston et al., 2015).

What the present analysis may add, if it adds anything beyond what active inference already provides, is the explicit naming of functional tensions that the active inference formalism does not foreground. Active inference provides the objective (minimize free energy) and the computational architecture (hierarchical generative model), but its standard formulations do not typically state the conflicts between specificity and reusability, between stability and revisability, between commitment and uncertainty, or between causal depth and epistemic cost as explicit architectural constraints that generate specific empirical predictions. The predictions derived from the tensions (threshold effects in assimilation vs. accommodation, double dissociations between online deployment and offline revision, format-dependent continual learning, active-exploration-dependent transfer robustness) are not standard outputs of active inference models. Whether they could be derived from active inference with appropriate model specification is an open formal question that would require work beyond this essay's scope, and pursuing that derivation would itself be a valuable contribution.

Causal abstraction theory. The formal causal abstraction literature (Beckers and Halpern, 2019; Rubenstein et al., 2017; Zennaro, 2022) provides the mathematical target for what a causally faithful macro-level model should look like. It specifies the conditions under which a higher-level model can be said to stand in a principled relation to a lower-level one, preserving interventionally relevant structure across levels. Causal representation learning (Schölkopf et al., 2021; Peters et al., 2017) translates this into the language of machine learning, arguing that learned variables should align with independent causal mechanisms so that they support transfer under distributional shift.

The relationship between the present analysis and causal abstraction theory is the most straightforward of the four comparisons. Causal abstraction theory defines the formal target; the present analysis proposes a developmental perspective on how a bounded, embodied agent might approach that target over the course of experience. Causal abstraction theory specifies what a causally faithful representation *is* but does not address how one is *built* from embodied interaction. The present analysis proposes a biological architecture for building causally graded representations and identifies the specific constraint (Section 4.4) that determines how far toward the formal target any given abstraction can get: the agent's behavioural repertoire limits its access to interventional evidence, bounding the causal depth that is epistemically achievable.

This complementarity also reveals a gap that neither tradition has fully bridged. Causal abstraction theory works with fully specified structural causal models at both the micro and

macro levels and asks whether a formal mapping between them preserves interventional structure. The biological system does not have access to fully specified models at either level. It must discover both the micro-level regularities and the macro-level abstractions simultaneously, from a stream of partially observed, action-dependent experience. How to extend the formal causal abstraction framework to handle this online, incremental, agent-driven setting is an important open problem that connects the present essay's concerns to the foundations of causal inference.

7 Open Questions

7.1 How causal are biological abstractions?

The evidence most clearly supports control relevance: prefrontal representations guide decisions under partial observability, and disrupting them impairs inference about hidden structure (Schuck et al., 2016; Schiereck et al., 2025). What remains largely untested is interventional stability: whether the same representations remain valid when the task changes or distributional context shifts.

Testing this requires designs that independently manipulate surface features and relational structure. Subjects would learn an abstract task structure in one perceptual domain and be tested on a structurally isomorphic task in a perceptually different domain, with the critical measure being whether the same prefrontal representation is deployed or a new one must be learned. Section 4.4 adds the specific prediction that subjects who actively explored the first domain should show more robust transfer than those who passively observed equivalent experience, because active exploration provides the interventional evidence needed for causal depth.

7.2 The vocabulary problem

This is the deepest unsolved question. Every structure-learning method presupposes a vocabulary of representational primitives; what grows is the number or composition of components, never the type. Developmental psychology documents qualitative shifts in representational capacity (Carey, 2009), but whether these reflect genuine vocabulary expansion or sophisticated recombination remains debated.

The most promising direction may be bridging neural world models, which learn grounded but opaque representations (Hafner et al., 2023; LeCun, 2022), with program induction and structured Bayesian models, which produce explicit compositional structure but struggle with grounding (Ellis et al., 2020; Heins et al., 2025). One concrete avenue is amortized structure-learning via Generative Flow Networks (GFlowNets) (Bengio et al., 2021; Deleu et al., 2022) searching over structured model spaces using reward signals from a grounded world model. Whether this can scale to genuine vocabulary expansion remains to be seen.

7.3 Fine-graining

Nearly all formalism in this area addresses coarse-graining. But agents must sometimes expand the resolution of a previously coarse abstraction when it proves insufficient. A schema for "vehicles" may be adequate until the agent needs to repair an engine. This operation is undertheorized: coarse-graining has rate–distortion theory, bisimulation, and the renormalization group; fine-graining has almost none. Yet from the tension framework's perspective, fine-graining is essential for navigating Section 4.2: when a coarse abstraction fails, the appropriate response may be to elaborate it locally rather than discard it entirely. The neural signature might be increased hippocampal engagement in a previously schema-dominated domain, producing a shift from prefrontally dominated back to hippocampally dominated processing in the subspace where the abstraction has failed.

7.4 Further open problems

Two additional problems deserve brief mention. First, *nested causal structure*: most causal discovery work assumes a flat set of variables at a single level, but physical systems have causal structure at molecular, material, object, and ecological levels simultaneously. The hippocampal–prefrontal system appears to represent information at multiple levels (Samborska et al., 2022; Courellis et al., 2024), but how multi-scale representations are coordinated and how new levels of description are introduced remain fundamental open questions.

Second, *implementation*: moving from computational analysis to working systems requires bridging a gap. The most promising direction combines a grounded world model for fast state inference, a structure-learning module for slow model revision, and an intrinsic motivation signal driving exploration toward uncertain parts of the environment. Whether the right approach will resemble the methods discussed here is uncertain. Sutton's bitter lesson (Sutton, 2019) applies: methods that scale with computation have historically outperformed methods embedding structural assumptions. Whether implicit structure discovery can support the modular, separately addressable library entries the present analysis argues are needed for continual learning remains one of the most important empirical questions in the field.

8 Conclusion

This essay began with a puzzle: how do specific, unrepeatable episodes give rise to internal structure that is general, reusable, and compositional? The analysis identified five functional constraints, showed they generate four inherent tensions, and argued that the hippocampal–prefrontal system can be read as an architecture shaped by these tensions.

Two claims go beyond what any single reviewed framework provides. First, the tensions between functional requirements are as important as the requirements themselves. They generate predictions: dissociations, thresholds, format-dependent continual learning, that would not follow from stating requirements in isolation. Second, the format hypothesis: if the brain's slow consolidation system produces representations with geometric modularity

(orthogonal population subspaces supporting independent readout), then continual learning follows from the knowledge format itself, not merely from the separation of learning rates. This connects the neuroscience of hippocampal–prefrontal memory to the artificial intelligence (AI) problem of catastrophic forgetting at a level deeper than analogy: both fields are grappling with how knowledge should be organized so that learning new things does not destroy old ones. The hypothesis is testable through the empirical signature proposed in Section 5.4, and its confirmation or refutation would have implications for both fields.

The analysis has clear limits. The three-stage architecture is an interpretive framework, not a mechanistic model. Evidence for control-relevant abstractions is strong; evidence for interventionally stable abstractions in the formal sense is sparse. The vocabulary problem remains unsolved. The deepest implication is that growing causal abstractions is not a subproblem of memory, reinforcement learning, causal inference, or program induction. It is the problem all of these fields are, in different vocabularies, trying to solve.

Acknowledgements

I am especially grateful to Dr. Lancelot Da Costa for his faith in me in pursuing such an ambitious topic, and for the intellectual freedom he gave me to explore it seriously. His close supervision, thoughtful guidance, and generous feedback were invaluable throughout this rotation. I also thank Prof. Dr. Bernhard Schölkopf for welcoming me into the Department of Empirical Inference and for fostering such a stimulating and intellectually generous research environment.

References

Abel, D. (2020). *A Theory of Abstraction in Reinforcement Learning*. PhD thesis, Brown University. 3 5 7

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429. 8 12

Audrain, S. and McAndrews, M. P. (2022). Schemas provide a scaffold for neocortical integration of new memories over time. *Nature Communications*, 13:5795. 3 20

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press. 22

Battaglia, P. W., Hamrick, J. B., Bapst, V., et al. (2018). Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261. 9

Beckers, S. and Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 2678–2685*. 3 7 27

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., and Kurth-Nelson, Z. (2018). What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509. 3 4

Bengio, Y., Malkin, N., Jain, M., et al. (2021). Flow network based generative models for non-iterative diverse candidate generation. arXiv:2106.04399. 28

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., and Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell, 183(4), 954–967.e21*. 23

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., and Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, 64(4):215–234. 13

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press. 28

Clune, J., Mouret, J.-B., and Lipson, H. (2013). The evolutionary origins of modularity. *Proceedings of the Royal Society B*, 280(1755):20122863. 11

Correa, C. G., Ho, M. K., Callaway, F., Daw, N. D., and Griffiths, T. L. (2023). Humans decompose tasks by trading off utility and computational cost. *PLoS Computational Biology*, 19(6):e1011087. 11

Costa, L. D., Gavenčiak, T., Hyland, D., Samiei, M., Dragos-Manta, C., Pattisapu, C., Razi, A., and Friston, K. (2025). Possible principles for aligned structure learning agents. arXiv. 10 15 27

Courellis, H. S., Minxha, J., Cardenas, A. R., et al. (2024). Abstract representations emerge in human hippocampal neurons during inference. *Nature*, 632:678–686. 5 19 29

Da Costa, L., Parr, T., Sengupta, B., and Friston, K. (2020). Active inference on discrete state-spaces: a synthesis. *Journal of Mathematical Psychology*, 99:102447. 3 10 15 17 26 27

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624. 3 10

Deleu, T., Bouchard-Côté, A., and Bengio, Y. (2022). Bayesian structure learning with generative flow networks. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*. 28

Diuk, C., Cohen, A., and Littman, M. L. (2008). An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*. 10

Douchamps, V., Jeewajee, A., Blundell, P., Burgess, N., and Lever, C. (2013). Evidence for encoding versus retrieval scheduling in the hippocampus by theta phase and acetylcholine. *Journal of Neuroscience*, 33(20):8689–8704. 22

Eichenbaum, H. (2017). On the integration of space, time, and memory. *Neuron*, 95(5):1007–1018. 5 19

El-Gaby, M., Lopes-dos Santos, V., Chen, X., et al. (2024). A cellular basis for mapping behavioural structure. *Nature*. 6 19 22

Ellis, K., Wong, C., Nye, M., et al. (2020). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. arXiv:2006.08381. 3 8 12 26 28

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2008). An hdp-hmm for systems with state persistence. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*. 12

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135. 3 11

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138. 3 26

Friston, K., Heins, C., Verbelen, T., Da Costa, L., Salvatori, T., Markovic, D., Tschantz, A., Koudahl, M., Buckley, C., and Parr, T. (2024). From pixels to planning: scale-free active inference. arXiv:2407.20292. 12 26

Friston, K., Parr, T., and Zeidman, P. (2018). Bayesian model reduction. arXiv:1805.07092. 8 26

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214. 17 27

Fusi, S., Miller, E. K., and Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37:66–74. 23

Gershman, S. J. (2017). On the blessing of abstraction. *Quarterly Journal of Experimental Psychology*, 70(3):361–365. 9

Ghosh, V. E. and Gilboa, A. (2014). What is a memory schema? a historical perspective on current neuroscience literature. *Neuropsychologia*, 53:104–114. 4 5 22

Gilboa, A. and Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory. *Trends in Cognitive Sciences*, 21(8):618–631. 4 21

Goodman, N. D., Tenenbaum, J. B., Feldman, J., and Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154. 12 26

Gopnik, A. (2012). Scientific thinking in young children: theoretical advances, empirical research, and policy implications. *Science*, 337(6102):1623–1627. 13

Griffiths, T. L. and Ghahramani, Z. (2011). The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224. 8 12

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, 87(1):1–51. 15

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. (2023). Mastering diverse domains through world models. Nature, 656? / or arXiv preprint for DreamerV3. 13 28

Hasselmo, M. E. (2006). The role of acetylcholine in learning and memory. *Current Opinion in Neurobiology*, 16(6):710–715. 22

Hasselmo, M. E., Schnell, E., and Barkai, E. (1995). Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region ca3. *Journal of Neuroscience*, 15(7 Pt 2):5249–5262. 22

Heins, C., Van de Maele, T., Tschantz, A., Linander, H., Markovic, D., Salvatori, T., Pezzato, C., Catal, O., Wei, R., Koudahl, M., Perin, M., Friston, K., Verbelen, T., and Buckley, C. (2025). Axiom: Learning to play games in minutes with expanding object-centric models. arXiv:2505.24784. 3 8 12 26 28

Howard, M. W. and Eichenbaum, H. (2015). Time and space in the hippocampus. *Brain Research*, 1621:345–354. 19

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134. 10

Kemp, C. and Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692. 3 9

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). Empowerment: A universal agent-centric measure of control. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2005)*. 13 17

Kumaran, D., Hassabis, D., and McClelland, J. L. (2016). What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7):512–534. 25

Lake, B. M. and Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, 623:115–121. 9

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338. 12 26

LeCun, Y. (2022). A path towards autonomous machine intelligence. OpenReview / position paper. 13 28

Lieder, F. and Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1. 11 17

Maor, I., Atwell, J., Ascher, I., Zhao, Y., Takahashi, Y. K., Hart, E., Pereira, F., et al. (2026). Persistent representation of a prior schema in the orbitofrontal cortex facilitates learning of a conflicting schema. *Nature Communications*. article in press / online 2026. 15 21

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (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(3):419–457. 3 11 14 19 25

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In Bower, G. H., editor, *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press. 3 5 11

McCourt, T., Fiete, I. R., and Chuang, I. L. (2023). Noisy dynamical systems evolve error correcting codes and modularity. arXiv:2303.14448. 11

Moneta, N., Grossman, S., and Schuck, N. W. (2024). Representational spaces in orbitofrontal and ventromedial prefrontal cortex: task states, values, and beyond. *Trends in Neurosciences*, 47(12):1055–1069. 21

Ni, T., Eysenbach, B., and Salakhutdinov, R. (2022). Recurrent model-free rl can be a strong baseline for many pomdps. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*. 10

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, 22(10):1544–1553. 3 4 10

O'Keefe, J. and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford University Press. 4 19

Pan, C., Yang, X., Li, Y., Wei, W., Li, T., An, B., and Liang, J. (2025). A survey of continual reinforcement learning. arXiv:2506.21872. 11

Parr, T., Pezzulo, G., and Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press. 3 10 26

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference (2nd ed.*). Cambridge University Press. 13

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books. 13

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press. 13 27

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I., and Battaglia, F. P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature Neuroscience*, 12(7):919–926. 16 20

Piaget, J. (1952). *The Origins of Intelligence in Children*. International Universities Press. 15

Preston, A. R. and Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, 23(17):R764–R773. 3

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497:585–590. 23

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. (2017). Causal consistency of structural equation models. arXiv:1707.00819. 27

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. (2016). Progressive neural networks. arXiv:1606.04671. 11

Samborska, V., Butler, J. L., Walton, M. E., Behrens, T. E. J., and Akam, T. (2022). Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nature Neuroscience*, 25(10):1314–1326. 19 22 29

Schiereck, S. S., Pérez-Rivera, D. T., Mah, A., DeMaegd, M. L., et al. (2025). The orbitofrontal cortex updates beliefs for state inference. *Neuron*, page 114(3) / online 2025. 21 28

Schuck, N. W., Cai, M. B., Wilson, R. C., and Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, 91(6):1402–1412. 3 4 6 21 28

Schwartenbeck, P., Baram, A. B., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., and Behrens, T. (2023). Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. *Cell, 186(22), 4885–4897.e14*. 20

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Towards causal representation learning. *Proceedings of the IEEE*, 109(5):612–634. 3 7 13 27

Scoville, W. B. and Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology*, pages Neurosurgery, and Psychiatry, 20(1), 11–21. 19

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR 2017*. 11

Spalding, K. N., Jones, S. H., Duff, M. C., Tranel, D., and Warren, D. E. (2015). Investigating the neural correlates of schemas: Ventromedial prefrontal cortex is necessary for normal schematic influence on memory. *Journal of Neuroscience*, 35(47):15746–15751. 21 23

Sutton, R. S. (2019). The bitter lesson. Essay. 29

Sutton, R. S., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211. 3 5 10

Tang, C., Herikstad, R., Parthasarathy, A., et al. (2020). Minimally dependent activity subspaces for working memory and motor preparation in the lateral prefrontal cortex. *eLife*, 9:e58154. 23

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285. 3 9

Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing / arXiv:physics/0004057. 7

Tompary, A. and Davachi, L. (2017). Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron, 96(1), 228–241.e5*. 3 20

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P., Morris, R. G. M., and Bothwell, M. (2007). Schemas and memory consolidation. *Science*, 316(5821):76–82. 19 25

Tsividis, P. A., Loula, J., Burga, J., Foss, N., Campero, A., Pouncy, T., Gershman, S. J., and Tenenbaum, J. B. (2021). Human-level reinforcement learning through theory-based modeling, exploration, and planning. arXiv:2107.12544. 12 26

van Kesteren, M. T. R., Fernández, G., Norris, D. G., and Hermans, E. J. (2010). Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proceedings of the National Academy of Sciences*, 107(16):7550–7555. 21

van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., and Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, 35(4):211–219. 21 22

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8:279–292. 7

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. E. J. (2020). The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell, 183(5), 1249–1263.e23*. 24 25

Wilson, M. A. and McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172):676–679. 16 20

Xiao, Z., Wang, X., Zhang, J., Ou, J., He, L., Qu, Y., Hu, X., Behrens, T., Liu, Y., et al. (2025). Human hippocampal ripples align new experiences with a grid-like schema. *Neuron*, 113(21):3661–3672.e4. 20

Yang, W., Sun, C., Huszár, R., Hainmueller, T., Kiselev, K., and Buzsáki, G. (2024). Selection of experience for memory by hippocampal sharp wave ripples. *Science*, 383(6690):1478–1483. 16 20 22

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2):273–293. 5

Zennaro, F. M. (2022). Abstraction between structural causal models. OpenReview / causal abstraction review preprint. 27