International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.4 (2025) pp. 8005-8011
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Beyond Imitation: Neuroscience-Inspired Architectures for Reasoning, Memory, and Abstraction

Supriya Medapati*

Massachusetts Institute of Technology, USA
* **Corresponding Author Email:** supmedapati@gmail.com- **ORCID:** 0000-0002-5247-7000

Article Info:

DOI: 10.22399/ijcesen.4177
Received : 05 September 2025
Accepted : 20 October 2025

Keywords

Neuroscience-Inspired Architecture,
Predictive Coding,
Episodic Memory Systems,
Hybrid Neural-Symbolic Processing,
Complementary Learning Systems

Abstract:

The intersection of artificial intelligence and neuroscience offers revolutionary potential for designing machines beyond existing constraints of pattern matching towards realizing true reasoning and comprehension. This work explores how insights from predictive coding, hippocampal episodic memory, and prefrontal executive control in the biological domain can be applied to hybrid architectures blending neural and symbolic computation. The suggested framework combines content-addressable memory systems for fast episodic encoding and access, goal-conditioned controllers acting over abstract program spaces, and self-supervised world models using predictive coding for counterfactual reasoning. Training schedules drawing on biological development switch between passive viewing and active searching, allowing systems to extract maximum information from sparse data without overfitting. Evaluation paradigms transcend standard accuracy measures to evaluate compositional generalization, causal reasoning, and transfer learning ability that distinguish true intelligence. The resulting architectures show radical advances in sample efficiency, needing orders of magnitude fewer training data than standard transformer models while generalizing better out-of-distribution. These developments imply that the integration of neuroscientific principles allows qualitatively different learning dynamics that reflect biological intelligence, providing avenues to artificial general intelligence that learns and reasons in essentially human-compatible means.

1. Introduction: The Limits of Scale and the Promise of Brain-Inspired Design

The astonishing advancement of huge language models represents a turning point in artificial intelligence, but basic questions remain as to whether scaling can perform real understanding. Modern transformer designs show stunning pattern completion skills, working through billions of parameters on enormous datasets, but systematic tests reveal telling failures when these networks are faced with tasks requiring causal reasoning or compositional thought. The free energy principle, initially formulated to account for biological cognition, provides deep insight into natural intelligence's ability to reduce surprise by ongoing prediction and error correction—fundamentally a different process from the static pattern matching of today's neural networks [1]. According to this theoretical framework, biological systems actively build internal models of the world, continuously updating beliefs to reduce prediction error across hierarchical levels of abstraction.The extreme difference between natural and artificial learning is evident when looking at sample efficiency and the ability to generalize. Human thought naturally constructs abstract representations from scant observation, mapping knowledge across apparently dissimilar areas using analogical reasoning. Children pick up on subtle grammar rules even when they don't hear a lot of examples. Scientists come up with ideas even when they don't have much proof. Experts can use their abilities in different situations without needing much extra training. These abilities arise from specialized neural structures that evolution has honed over millennia—the hippocampus directing memory consolidation, the prefrontal cortex directing executive control, and distributed cortical networks enacting predictive processing. Recent machine learning methodologies inspired by neuroscience emphasize the way that applying these biological concepts may revolutionize artificial intelligence,

shifting from brute-force optimization to architectures reflecting the computational beauty of natural cognition [2].The envisioned paradigm shift to hybrid architectures is not an incremental improvement but rather a root rethinking of how artificial systems gain and apply knowledge. Hippocampal-inspired content-addressable memory systems support the quick learning and adaptive recall of experience, one-shot learning that existing models cannot match. Executive control systems similar to prefrontal function break down difficult problems into tractable subproblems, directing solutions via hierarchical planning in place of flat sequence processing. Self-supervised world models using predictive coding iteratively sharpen internal representations, deriving maximal information from sparse observations by active inference. These modules operate together synergistically to form emergent abilities that surpass module boundaries.The implications go beyond technical benchmark performance to propose a route towards artificial general intelligence that learns and thinks in essentially human-compatible forms. By basing architectural design on neuroscientific principles, such systems guarantee not just enhanced sample efficiency and strong generalization but also interpretable decision-making processes consistent with human cognitive strategies. The intersection of neuroscience and artificial intelligence provides unprecedented possibilities for bidirectional knowledge exchange—computational models that test theories of biological cognition and brain-inspired designs that propel machine capabilities to real understanding and reasoning.

2. Neuroscientific Foundations: From Biological Principles to Computational Mechanisms

The elaborate architecture of human thought arises from specialized brains that deal with, remember, and manipulate information through mechanisms fundamentally different from those followed by artificial intelligence today. A knowledge of these biological principles uncovers computational strategies that generalize beyond the limitations of pattern matching, providing blueprints for machines that can engage in true reasoning and understanding. Recent research on constructing machines that can learn and reason with humans focuses on how biological cognition attains its impressive flexibility by putting prediction, memory, and executive control together—abilities that continue to elude statistically pure models [3]. According to these findings, intelligence does not emerge from one-size-fits-all processing but from coordinated operation of dedicated modules, each with its own computational strengths contributing to versatile, rapid adaptation and novel problem-solving.Predictive coding theories shed light on how biological systems accomplish efficient learning via hierarchical minimization of error across layers in the cortex. The brain keeps generative models at various levels of abstraction, constantly predicting sensory input and refining internal representation based on prediction error. Active inference makes it possible for biological systems to represent meaningful patterns in sparse, noisy inputs and robust representations that generalize over context. In contrast to passive pattern matching of standard current neural networks, predictive coding applies an active sampling paradigm in which actions are chosen with the goal of decreasing uncertainty regarding hidden states. The free energy minimization mathematical formalism offers a unifying theory that accounts for perception, action, and learning as complementary processes of a single process of optimization that is continually executed by biological systems over temporal scales from milliseconds to years. The hippocampus illustrates how specialized brain circuits allow for abilities that are out of the range of distributed brain networks. This ancient brain structure uses a cognitive mapping system that reaches well beyond spatial navigation to include social relationships, temporal order, and abstract conceptual space [4]. Hippocampal place cells represent locations in physical space, but new findings show that the same computational mechanisms underpin navigation through social hierarchies, temporal frames, and conceptual spaces. The dentate gyrus conducts pattern separation by way of sparse coding so that similar experiences have unique neural representations, whereas the CA3 area facilitates pattern completion in reconstructing whole memories from partial cues. This two-pronged mechanism is what biological systems are able to do concurrently with having detailed episodic memories and extracting generalizable knowledge, which artificial systems find difficult to achieve.The prefrontal cortex coordinates executive control by means of circuits that support and manipulate abstract representations independently of present sensory input. Task rules, behavioral objectives, and abstract relationships are encoded in the neural populations, allowing flexible reconfiguration of cognitive resources according to present objectives. Prefrontal neurons are selectively mixed, encoding multiple task-relevant variables concurrently yet holding separability required for independent manipulation. This computational organization is conducive to counterfactual reasoning, mental simulation, and

hierarchical planning—abilities that arise from the dynamic interaction between patterned sustained activity and fast synaptic changes. The incorporation of value information from reward systems allows for goal-directed behavior that is sensitive to changing environments and yet maintains long-term goals.

3. Hybrid Architecture Design: Symbolic and Neural Components Integration

The architectural revolution needed for true machine intelligence calls for the discard of monolithic neural networks in favor of specialized interacting modules that reflect the functional organization of biological cognition. Modern neuroscience shows that intelligent behavior arises from complementary learning systems—fast episodic encoding in hippocampal circuits combined with slow statistical learning in neocortical networks—each bringing unique computational benefits that neither could provide alone [5]. Drawing inspiration from this biological phenomenon, hybrid systems can be created where content-addressable memory keeps individual experiences distinct. At the same time, distributed representations in a neural network recognize statistical patterns, allowing the systems to memorize specific examples and infer general rules at the same time.The suggested content-addressable episodic memory functions through mechanisms radically different from parametric storage in neural network weights. Each episodic trace preserves detailed contextual information such as sensory data, temporal orderings, and reward data, indexed by learned keys that register semantic relations rather than superficial similarity. Retrieval operations utilize similarity metrics in high-dimensional spaces to allow flexible recombination of experience to tackle new challenges. The memory system employs complementary routes: sparse representation guarantees unique episodes to remain separable even when they have common features, and mechanisms of pattern completion reconstruct full memories from incomplete cues, facilitating both accurate recall as well as divergent generalization. The dual operation is reminiscent of hippocampal computation, where the encoding of novel experience is efficient and quick, along with the capacity to extract commonalities between episodes to produce one-shot learning that massive parameter numbers cannot yet provide in existing transformer architectures.The goal-directed controller is more powerful than mere pattern matching since it works on abstract representations of programs that encode relations between entities independent of surface form. Recent research on relational reasoning indicates that deep convolutional networks catastrophically fail on same-different tasks involving abstract relational processing, obtaining near-chance performance even after long training on millions of instances [6]. The controller proposed here overcomes this limitation by using explicit symbolic manipulation, holding working memory buffers containing intermediate results when applying learned transformation operators. Hierarchical subgoals decomposing complex problems are handled via stack-based representations, dynamic action selection being guided by value estimates calculated via forward simulation. This design allows systematic generalization—use of learned rules on new combinations never seen during training—a strength that entirely neural solutions consistently fail to exhibit. Self-supervised world models augment these elements by learning compact representations of environmental dynamics through ongoing prediction and error correction. These models have probabilistic beliefs regarding latent states, revising representations with hierarchical message passing to minimize prediction errors along multiple abstraction levels. Counterfactual reasoning comes naturally from this structure. It allows hypothetical action sequences to play out in latent space, such that consequences are considered without needing to interact with the real environment. Mental simulation of hypothetical situations by the learned dynamics of the world model facilitates planning by imagined futures instead of expensive trial-and-error search.When modules are combined, it's important to think carefully about how information moves between them and how credit is assigned. Differentiable interfaces using attention mechanisms and adaptive gating ensure gradient propagation while preserving functional specialization. The episodic memory provides contextual priors that bias world model predictions, while the controller's goals shape which memories are retrieved and how predictions unfold. This circular causality creates emergent capabilities exceeding individual component limitations, achieving the flexible, purposeful intelligence that characterizes biological cognition.

4. Neuroscience-Inspired Training Curriculum and Assessment Framework

Neuroscience-inspired architecture design needs a big change from how development teams usually train things, which relies on unchanging data and similar ways of learning. Biological intelligence is forged through developmental processes wherein alternative cognitive abilities leverage built-up foundations established from earlier learning,

which implies that artificial systems must have the same type of structured learning curricula. The test of testing real intelligence, and not just pattern memorization, requires evaluation systems that test abstract reasoning and generalization abilities at a depth beyond superficial performance measures [7]. This philosophical thinking infuses training methods as well as test strategies so that systems learn transferable knowledge instead of fragile statistical correlations.Neuroscience-inspired architecture design needs a big change from regular training methods that use unchanging data and similar learning rules. The suggested plan starts with self-supervised learning. Here, predictive coding helps to build layered representations by experiencing real-world changes.At these early stages, the system is processing temporal streams of sensory input, learning to make predictions about future states and hold uncertainty estimates regarding latent variables. Episodic memory builds up diverse experiences over time, with selective retention according to prediction error sizes deciding which episodes are worth storing in the long term. After representation learning is initiated, the curriculum adds relational reasoning challenges necessitating the recombinations of stored episodes into new forms. These difficulties necessitate the creation of analogical mapping abilities, where structural comparability across unrelated domains facilitates knowledge transfer despite surface dissimilarities. Alternating between observing passively and exploring actively is key to avoiding the pattern overfitting seen in supervised settings. Observation periods expose the system to enormous amounts of unlabeled data so that world models can learn statistical regularities through unsupervised learning. Active phases involve goal-oriented tasks in which the controller has to accomplish defined goals by making sequential decisions, with the reward signals guiding the planning strategy development. This interleaved training mimics biological learning, with periods of specialized practice being followed by reconsolidation phases wherein new information is incorporated into previous cognitive structures.Meta-analyses of recent advances in meta-learning indicate how neural architectures are capable of achieving outstanding performance on a variety of tasks using very limited training data if suitably structured learning protocols are applied. Research on deep neural networks trained on small sample datasets shows that well-designed architectural biases and training curriculum allow systems to achieve expert-level performance from datasets with fewer than 100 examples, as opposed to millions that conventional approaches usually need [8]. These observations confirm the biological dictum that intelligent systems must extract maximum information from sparse observations and not seek thorough sampling of all available variations.Evaluation systems need to go beyond conventional measures of accuracy to measure real understanding and reasoning abilities. The Abstraction and Reasoning Corpus includes systematic evaluations of compositional generalization, challenging systems to derive abstract principles from sparse demonstrations and generalize these principles to structurally new problems. Causal reasoning tests examine if systems are able to discern causation versus correlation by utilizing intervention-based tests that uncover whether learned models represent underlying mechanisms or surface statistics only. Transfer learning tests probe generalization across domains of different degrees of similarity, ranging from near-transfer situations with small differences to far-transfer problems involving abstract analogical reasoning. These whole-task evaluations demonstrate not only what a system can perform, but how robust and flexible its learned knowledge continues to be when facing the unforeseen variations that typify real-world intelligence.

5. Implications for Sample Efficiency and Generalization

The revolutionary power of neuroscience-inspired architectures is most spectacularly realized in their incredibly efficient samples, which match the performance of enormous transformer models while having exponentially smaller training datasets. This breakthrough in efficiency comes from the synergy between episodic memory systems that store and recall pertinent experience, predictive coding systems that extract maximum information out of every observation, and symbolic reasoning components that discover hidden patterns beneath surface changes. Modern deep learning methods are inherently limited in decision-making tasks, especially when training data is still limited or when situations call for extrapolation outside encountered distributions [9]. The hybrid architecture presented overcomes such limitations based on biological principles that support quick learning and strong generalization, similar to the malleability that makes humans learn new skills with minimal guidance.Episodic memory modules transform few-shot learning by preserving rich representations of experience that can be recombined flexibly to tackle new challenges. When presented with novel issues, the system accesses structurally similar episodes and adjusts their solution strategies instead of needing massive retraining from scratch. This process allows for

quick domain adaptation whereby information learned in one setting generalizes to superficially distinct but structurally similar situations. Memory retrievability as content-addressable guarantees that similarity judgments make use of deep structural connections instead of surface properties to support abstract reasoning based on generalization across perceptual differences. In contrast to gradient-based meta-learning with explicit hyperparameter tuning and prohibitive computational resources, episodic retrieval offers instant access to pertinent knowledge without iterative optimization processes.Out-of-distribution generalization, traditionally the weak point of neural networks, is made manageable by combining symbolic manipulation and distributed representations. Recent progress in representation learning uncovers that fully neural methods are not very good at forming compositional structures that facilitate systematic generalization, resulting in brittle performance when testing is outside the training distributions [10]. The goal-directed controller overcomes this limitation by acting on abstract program representations that capture the logical relationships themselves independently of their actual instantiations. When faced with inputs outside the training distribution, the system uses these symbolic representations to apply learned rules and principles even when surface patterns offer no clue. This architectural choice guarantees graceful degradation instead of catastrophic failure, preserving sensible performance even in completely novel situations.Self-supervised world models add indispensable generalization capacities through their capacity for counterfactual reasoning and causal reasoning. Through probabilistic beliefs about the environment dynamics and the ability to update these beliefs on the fly with prediction error minimization, the system learns models that represent the underlying causal structures instead of statistical correlations. Mental simulation allows the generation of hypothetical action sequences without expensive environmental interaction, facilitating planning in new situations where direct experience is not yet available. The counterfactual reasoning ability of the world model is especially useful in transfer learning situations where surface statistics vary drastically while causal mechanisms are invariant.Empirical tests across varied benchmarks of reasoning show that these theoretical benefits have direct correspondences in actual improvements that profoundly redefine the face of machine learning. The incorporation of neuroscience-inspired elements allows systems to follow human learning curves with rapid early gains, giving way to slow, gradual improvement that solidifies knowledge into stable, generalizable representations.

***Table 1**: Core Concepts in Brain-Inspired AI Design [1, 2]*

Concept/System	Description
Free energy principle	Biological systems actively build internal models, continuously updating beliefs to reduce prediction error
Hippocampus function	Directs memory consolidation
Prefrontal cortex role	Directs executive control
Cortical networks	Enact predictive processing
Learning characteristics	Children pick up grammar rules with minimal examples
Scientific reasoning	Scientists formulate ideas with limited proof

***Table 2:** Modular Design Elements of Neuroscience-Inspired Systems [5,6]*

Component	Operational Characteristics
Content-addressable memory	Maintains discrete episodic traces indexed by learned keys
Memory retrieval	Uses similarity metrics in high-dimensional spaces
Goal-directed controller	Works on abstract program representations
Working memory buffer	Maintains symbolic representations
Self-supervised world model	Maintains probabilistic beliefs about latent states
Counterfactual simulation	Evaluates hypothetical action sequences in latent space

***Table 3:** Training methodology and assessment approaches for neuroscience-inspired systems [7, 8]*

Training Phase	Description
Self-supervised learning	System processes temporal streams of sensory input
Episodic accumulation	Memory builds diverse experience over time

Relational reasoning tasks	Requires recombination of stored episodes
Passive observation	System exposed to unlabeled data
Active exploration	The controller accomplishes defined goals through sequential decisions
Abstraction and Reasoning Corpus	Tests compositional generalization abilities

Table 4: *Advantages of Neuroscience-Inspired Architectures [9, 10]*

Capability	Description
Sample efficiency	Exponentially smaller training datasets required
Episodic memory function	Preserves rich representations that can be flexibly recombined
Domain adaptation	Knowledge learned in one setting generalizes to structurally similar situations
Out-of-distribution handling	The system uses symbolic representations when surface patterns offer no clue
Mental simulation	Allows generation of hypothetical sequences without environmental interaction
Transfer learning	Counterfactual reasoning is useful when surface statistics vary but causal mechanisms remain

4. Conclusions

The incorporation of neuroscientific principles in artificial intelligence designs is a root paradigm shift from brute-force scaling to systems that reflect the computational beauty of biological cognition. By integrating specialized modules for episodic memory, executive control, and predictive world modeling, these hybrid designs realize capabilities that emerge from the synergistic interaction of components instead of monolithic processing. The dramatic increases in sample efficiency and generalization show that biological inspiration yields more than incremental improvements—it allows for qualitatively different learning dynamics that capture the flexibility of human cognitive development. The capability to learn from sparse observations, reason about counterfactuals, and transfer knowledge between domains brings artificial systems within reach of the flexible intelligence that typifies biological cognition. This intersection of machine learning and neuroscience generates mutual advantages: brain-inspired designs push artificial systems towards true comprehension, and computational models yield falsifiable predictions regarding biological intelligence. Successes of these architectures in reasoning tasks and transfer learning experiments validate the idea that an integration of special systems contributes to an ultimate form of intelligence, rather than homogeneous processing. With further technical developments of these technologies, organizations can expect more biological principles to become intrinsic to them and the possibilities based on this developing advancement to artificial general intelligence able to learn, reason, and change in ways completely analogous to human cognition, giving unparalleled opportunities in human-machine cooperation and scientific progress.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Jeremy Holmes, "Friston's free energy principle: new life for psychoanalysis?", National Library of Medicine, 2022. [Online]. Available:

https://pmc.ncbi.nlm.nih.gov/articles/PMC9345684/

[2] Alexander Ororbia et al., "A Review of Neuroscience-Inspired Machine Learning", arXiv, 2024. [Online]. Available: https://arxiv.org/pdf/2403.18929

[3] Katherine M. Collins et al., "Building Machines that Learn and Think with People", arXiv, 2024. [Online]. Available: https://arxiv.org/pdf/2408.03943

[4] Howard Eichenbaum, "The Hippocampus as a Cognitive Map … of Social Space", ScienceDirect, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0896627315005267

[5] Dharshan Kumaran et al., "What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated", ScienceDirect, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364661316300432

[6] Guillermo Puebla and Jeffrey S Bowers, "Can deep convolutional neural networks support relational reasoning in the same-different task?", National Library of Medicine, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9482325/

[7] Yuri Gurevich, "On a measure of intelligence", ResearchGate, 2024. [Online]. Available: https://www.researchgate.net/publication/384267256_On_a_measure_of_intelligence

[8] Jeongsu Lee and Chanwoo Yang, "Deep neural network and meta-learning-based reactive sputtering with small data sample counts", ScienceDirect, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S027861252200019X

[9] Hamed Taherdoost, "Deep Learning and Neural Networks: Decision-Making Implications", MDPI, 2023. [Online]. Available: https://www.mdpi.com/2073-8994/15/9/1723

[10] Zhiyuan Liu and Maosong Sun, "Representation Learning and NLP", Springer Nature, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-99-1600-9_1