ChatGPT

External Research Supporting Phase 1 Cognitive AI Architecture

To strengthen the Phase 1 Cognitive AI Architecture plan, we surveyed interdisciplinary research that aligns with its core concepts. Below we organize key findings from cognitive neuroscience, AI architecture design, and reinforcement learning that **complement or extend Phase 1**. These sources reinforce principles like multi-dimensional cognitive modeling, adaptive learning (neuroplasticity), graph-structured memory, hierarchical modular design, and neuro-symbolic integration.

Multidimensional Cognitive Parameters (“Thinking Vectors”)

Cognitive science and AI research both suggest that human-like intelligence requires multiple concurrent cognitive processes (attention, decision-making, intention, adaptation). For example, **Posner’s tri-network theory of attention** identifies separate neural networks for *alerting* (achieving readiness), *orienting* (shifting focus), and *executive control* (decision/conflict resolution) [1]. This implies that an AI’s “thinking vector” might need several dynamic dimensions. In a brain-inspired cognitive architecture by Ballard *et al.* (2013), the authors explicitly separate cognitive functions into a *Debug* level for high-level attention (monitoring and reprogramming modules), an *Operating System* level for executive selection of modules (task management), *Task* modules for goal-directed intentions, and low-level *Routine* modules for sensory-motor processing with learning [2] [3]. Such a hierarchy effectively implements distinct cognitive parameters: the Debug/attention layer **attends** to anomalies, the OS layer **decides** which processes to run, Task modules encapsulate **intentions** (goals), and Routine-level learning enables **adaptation** via reinforcement feedback (rewards) [4]. This shows that *multi-dimensional cognitive state vectors* (attending, deciding, intending, adapting) are grounded in both neuroscience and AI – architectures benefit from dedicating modules or parameters to each function, mirroring the human brain’s division of labor. Notably, brain network analyses find that having such **modular, multi-faceted organization** yields greater robustness and adaptivity [5]. In summary, research supports modeling cognition as a vector of interacting components, rather than a single monolithic state, to capture the rich dynamics of attention, decision, goals, and learning.

Neuroplasticity and Adaptive Reinforcement Learning

A key to human-level adaptability is **neuroplasticity** – the brain’s ability to rewire and adjust connections in response to experience. AI researchers are incorporating analogous plasticity mechanisms into reinforcement learning (RL) agents to enable continual adaptation. **Neuromodulated plasticity** is one approach: Miconi *et al.* (Uber AI Labs) showed that giving neural networks the ability to *self-modify* their weights in response to a differentiable “neuromodulator” signal (analogous to dopamine) greatly improves learning performance [6]. In their model, the network can “decide” when and where to alter its synapses based on incoming inputs and rewards, filtering out irrelevant activity and **reinforcing important connections in a reward-dependent manner** [7]. This differentiable plasticity allowed agents to learn tasks faster and retain skills longer, highlighting how *dynamic rewiring* yields continual learning akin to biological brains [8] [7]. Recent work on **“neuroplastic expansion” in deep RL** pushes this further by *growing* network capacity over time. For instance, Zhu *et al.* (2024) propose starting with a small network and gradually expanding it by adding neurons and connections as learning progresses [9] [10]. This growth, inspired by cortical expansion in neurodevelopment, helps maintain high plasticity and avoid convergence to a rigid policy. The agent can reactivate or prune neurons based on usage, and periodically **consolidate knowledge to balance stability vs. plasticity** [11] [12]. Together, these approaches demonstrate *adaptive RL algorithms* that mimic neuroplasticity – from synapse-level weight adjustments to structural network changes – enabling AI systems to continually update their "mental wiring" in response to new tasks or changing goals. Such mechanisms directly support the Phase 1 aim of an architecture that **learns and adapts dynamically** rather than remaining static.

Graph-Based Memory Structures with Weighted Relations

Both neuroscience and AI indicate that **knowledge is efficiently stored as a network** of interconnected concepts – a graph structure – where the *relationships* carry crucial weight. Cognitive neuroscience has long theorized that humans form "cognitive graphs" in memory: mental networks of entities and their interrelations that allow flexible reasoning and navigation of information [13]. Unlike a linear memory, a graph-based memory can link ideas via reinforced associations, much like the brain's semantic networks or spatial cognitive maps [14] [15]. Current AI research is embracing this idea through **knowledge graphs** and memory graphs. For example, Kim *et al.* (2024) built an RL agent that learns in a partially observable environment where the hidden state is represented as a dynamic **knowledge graph (KG)** [16]. The agent maintains both *episodic* and *semantic* memories by storing facts as graph triples augmented with context meta-data – e.g. adding a timestamp for episodic memory or a "strength" weight for the salience of a fact [17]. These weighted graph relations serve as a long-term memory: the agent can remember and retrieve relevant knowledge to answer questions or make decisions, and the weights are updated as the agent gains experience (similar to how human memories strengthen with reinforcement) [16] [17]. Another study introduced *AriGraphs* – memory graphs that integrate semantic knowledge with episodic traces – enabling a language-model agent to accumulate and use world knowledge over time [18] [19]. Critically, frameworks like **OpenCog** have shown how a cognitive architecture can implement a *graph memory* with continuous importance updating: in OpenCog's AtomSpace (a weighted knowledge network), an "attention allocation" process **assigns importance weights to each knowledge atom/link based on how useful or reinforced it has been in the past and present** [20]. This acts as a cognitive economy, concentrating activation on frequently-used or high-value associations and gradually fading out unimportant ones (a computational analog of synaptic strengthening/weakening). Practically, graph databases like Neo4j have been used to implement such dynamic memories in AI agents. For instance, Salvador (2021) describes a modular RL architecture using Neo4j as the agent's memory store, where **state perceptions are encoded as graph nodes/edges and queries (Cypher graph queries) are used to retrieve or update policies** [21] [22]. Their results confirmed that episodic data and learned policies can be effectively represented in a graph database, allowing the agent to *remember* and adapt within a unified graph memory. All these works underscore that a **flexible, graph-based memory with weighted relations** is not only biologically plausible but highly practical – it enables reinforcement-driven knowledge consolidation, fast recall of relevant info, and rich relational reasoning in cognitive AI systems.

Hierarchical and Modular Cognitive Structures (Meta-Regions, Clusters, Threads)

An overarching theme in cognitive architecture design is **hierarchical modularity** – breaking the mind into regions or modules that specialize, yet linking them in a layered control structure. This mirrors the brain's organization: neuroscience shows that the brain is organized as a hierarchy of modules (networks within networks), a design that yields greater robustness, adaptability, and evolvability [5]. In AI terms, this means structuring the architecture into tiers (meta-regions) composed of sub-modules (cognitive clusters), with parallel "threads" of processing for different tasks or contexts. The earlier example by Ballard *et al.* embodies this principle with its four-layer cognitive hierarchy (Debug, OS, Task, Routine), ensuring that processing is distributed and **each layer handles a distinct aspect of cognition** [2] [3]. Notably, they liken *working memory* to multiple **threads** of execution: each Task module runs as an independent thread maintaining its own state, enabling multitasking and context-switching without interference [23]. This thread-based view of working memory aligns with computer science concepts and helps explain phenomena like attention bottlenecks (only so many threads can be active in focus) [23]. Hierarchical modular architectures are also evident in modern brain-inspired AI projects. The EU's Human Brain Project, for instance, developed a cognitive framework that allows **flexible integration of heterogeneous modules (modeled after brain areas) into a coherent whole** [24]. Each module can be trained for a specialized function (vision, planning, motor control, etc.), and a top-level cognitive controller coordinates these modules in a *meta-regional* fashion. Such designs echo the idea of "meta-regions" overseeing collections of processing units (clusters) and spawning "threads" for concurrent tasks. The benefit is that higher-level reasoning or meta-cognitive units can monitor and adapt lower-level processes (as the Debug meta-region does in Ballard's architecture, reprogramming modules on the fly [25]). Research in network science further supports this approach: hierarchical **modules-within-modules** enable systems to handle complexity through divide-and-conquer, while inter-module links integrate the outcomes into unified behavior [26] [27]. In summary, a **hierarchical modular architecture** – with meta-regions guiding clusters of cognitive processes and threaded execution of tasks – is well-grounded in both brain science and AI studies. It offers a pathway to scalability (via parallelism and abstraction layers) and resilience (localized failures don't topple the whole system, and modules can adapt independently).

Neuro-Symbolic Integration of Reasoning and Learning

Integrating symbolic reasoning (manipulation of concepts, logic, language) with data-driven learning (neural networks, reinforcement learning) is crucial for a flexible AI that can both *think* and *learn*. A growing body of **neuro-symbolic AI** research directly supports this hybrid approach. Recent surveys describe neuro-symbolic AI as combining the strengths of neural networks (high-dimensional pattern learning) with those of symbolic systems (explicit logic, interpretable knowledge) to create advanced cognitive systems [28]. The goal is an AI that can, for example, perceive patterns in raw data *and* perform abstract reasoning on acquired knowledge – bridging low-level learning and high-level cognition. One prominent example is the **OpenCog** framework's new incarnation as *OpenCog Hyperon*, which is designed as a hybrid architecture mixing neural and symbolic subsystems. In a prototype described by Goertzel *et al.*, a *logical reasoning engine* (operating over a graph knowledge base of symbols) is tightly coupled with a *hierarchical neural network*; remarkably, **each side influences the other's inner workings** in real time [29]. They report that *symbolic inference* steps can be guided by neural activations, while conversely the neural network's attractor dynamics are influenced by on-the-fly symbolic logic results [29]. This kind of deep integration – "neurons and symbols in a **combined dynamical system**" – goes beyond mere pipelines (where one module's output feeds another) to a genuinely interactive cognitive loop. Other neuro-symbolic approaches, as reviewed by Nawaz *et al.* (2025), include techniques like **Logic Tensor Networks, differentiable logic programming, and neural theorem provers**, all of which allow symbolic structures (e.g. logical rules or knowledge graphs) to be embedded in neural computation [30]. These methods enable learning algorithms to respect symbolic constraints or factual knowledge, and conversely let logical reasoning benefit from learned neural representations. A key trend is making such systems *adaptive and real-time*: future neuro-symbolic architectures are expected to be **adaptive, dynamic, and responsive**, maintaining interpretability *without* sacrificing the flexibility of learning [31] [32]. In practical terms, this could mean an AI that uses neural nets to interpret sensory inputs and propose candidate solutions, then uses a symbolic planner to verify constraints or long-term coherence, all within an integrated loop. The Phase 1 architecture's call for *hybrid reasoning* is strongly backed by these developments – from academic proposals to industry prototypes – indicating that uniting symbolic reasoning with subsymbolic learning is a viable path to human-like intelligence. Neuro-symbolic systems deliver the “best of both worlds” [33] : the adaptability and pattern-recognition power of neural networks, coupled with the compositional, explainable problem-solving of symbolic AI, yielding **cognitive architectures that learn, reason, and explain** in tandem.

Sources: The references cited (【2】–【38】 above) include peer-reviewed research articles, conference papers, and technical reports spanning cognitive neuroscience (e.g. studies of cognitive maps [13] and attention networks [1]), AI reinforcement learning innovations [6] [10] , cognitive architecture designs [2] [20] , and neuro-symbolic integration frameworks [29] [28] . Each source was selected for its relevance to the corresponding concept and to provide theoretical as well as practical guidance (for instance, using graph databases for memory [21] or adopting hybrid learning/reasoning algorithms). This interdisciplinary literature underpins the Phase 1 implementation plan, offering both inspiration and validated patterns for building a cognitively robust AI system.

---

[1] [2] [3] [4] [23] [25] msr2414.dvi
https://www.cs.utexas.edu/~dana/MSR2414.pdf

[5] [26] [27] Modular and hierarchically modular organization of brain networks - PubMed
https://pubmed.ncbi.nlm.nih.gov/21151783/

[6] [7] [8] openreview.net
https://openreview.net/pdf?id=r1lrAiA5Ym

[9] [10] [11] [12] Neuroplastic Expansion in Deep Reinforcement Learning
https://arxiv.org/html/2410.07994v1

[13] [14] [15] Structuring Knowledge with Cognitive Maps and Cognitive Graphs - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC7746605/

[16] [17] Leveraging Knowledge Graph-Based Human-Like Memory Systems to Solve Partially Observable Markov Decision Processes
https://arxiv.org/html/2408.05861v1

[18] [19] AriGraph: Learning Knowledge Graph World Models with Episodic Memory for LLM Agents
https://www.ijcai.org/proceedings/2025/0002.pdf

[20] Hands On with Attention Allocation (2017 Archive) - OpenCog
https://wiki.opencog.org/w/Hands_On_with_Attention_Allocation_(2017_Archive)

[21] [22] (PDF) REINFORCEMENT LEARNING WITH GRAPH DATABASES: my Human Brain Project
https://www.researchgate.net/publication/348077190_REINFORCEMENT_LEARNING_WITH_GRAPH_DATABASES_my_Human_Brain_Project

[24] Brain Inspired Cognitive Architectures
https://www.humanbrainproject.eu/en/science-development/focus-areas/brain-inspired-cognitive-architectures/

[28] [30] [31] [32] A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems - ScienceDirect
https://www.sciencedirect.com/science/article/pii/S2667305325000675

[29] [33] OpenCog NS: A Deeply-Interactive Hybrid Neural-Symbolic Cognitive Architecture Designed for Global/Local Memory Synergy
https://cdn.aaai.org/ocs/871/871-4244-1-PB.pdf