

Future Directions

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 19 of "The NeuroAI Handbook"

Three frontiers, hardware, learning, and architecture, where the next decade of NeuroAI will be won.

One-sentence thesis. The future of NeuroAI is not just better algorithms but a rethinking of computation itself along three fronts, building hardware like the brain, learning like the brain, and integrating like the brain, with each brain-inspired advance looping back as a tool to understand the brain.

PART 1: Frontier 1, Hardware Like the Brain

The energy problem

Modern AI runs on hardware that is spectacularly inefficient for intelligence. A GPU is a general-purpose factory that can build anything but burns a city's worth of power to make a single car; the brain is a hyper-specialized factory that rivals a supercomputer on **20 watts**. The root cause is the **von Neumann bottleneck**: conventional computers keep processing and memory in separate units and spend most of their energy shuttling data across the bus between them. The brain has no such split, computation and memory are co-located in the same synaptic connections, and that is the secret of its efficiency. **Neuromorphic computing** is the project of rebuilding hardware in the brain's image.

Spikes and in-memory computing

Two brain-borrowed ideas attack the bottleneck. **Spiking neural networks (SNNs)** communicate with discrete, all-or-nothing spikes; neurons are silent by default and draw power only when they fire, an **event-driven** design instead of a fixed clock grinding through every tensor. **In-memory computing**

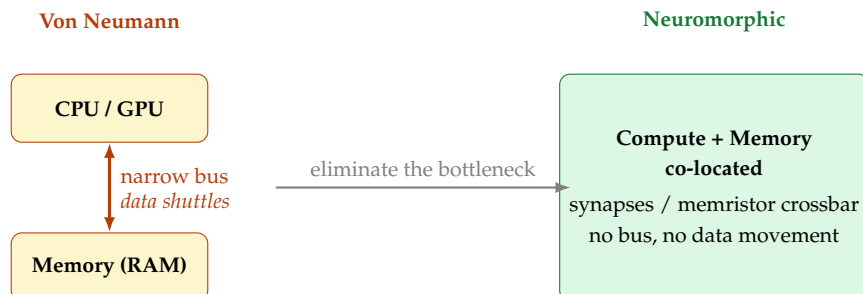
with **memristors**, whose resistance stores a synaptic weight, performs the multiply *inside* the memory. Arrange memristors in a crossbar, apply input voltages, and by Ohm's law and Kirchhoff's current law the output current on each column *is* a matrix-vector product, computed in one physical step:

$$I_j = \sum_{i=1}^n V_i G_{ij} = \sum_{i=1}^n V_i W_{ij}.$$

No transistor switching, no data movement, just physics. The leading device families are resistive RAM (RRAM) and phase-change memory (PCM).

ANALOGY

The von Neumann bottleneck is a chef whose ingredients sit in a pantry across the building: most of the "cooking" is walking back and forth. Neuromorphic hardware keeps the pantry on the counter, so computation happens where the data already lives.



Conventional machines burn energy moving data to compute; neuromorphic chips compute where the data already sits.

Systems already running

This is no longer theory. Intel's **Loihi 2** (1 million spiking neurons, 128 million synapses) ran real-time object detection at **0.3 mJ** per inference against 45 mJ on a GPU, a 150x saving, at comparable 89% accuracy. IBM's **TrueNorth**, paired with event-based (dynamic vision) cameras, hit 95% gesture recognition

on just 63 mW. Manchester's **SpiNNaker**, a million-core machine, simulates up to a billion neurons in biological real time for computational neuroscience. The catch: training algorithms for SNNs still lag those for standard nets, which is the field's main bottleneck to adoption, but as inference moves to milliwatt edge devices, the advantage grows compelling.

Platform	Energy / inference	Latency	Accuracy	Active power
NVIDIA V100 GPU	42 mJ	4.8 ms	76.5%	300 W
Intel Loihi 2 (SNN)	0.28 mJ (150x)	18 ms	74.2%	3 W
IBM TrueNorth	0.09 mJ (467x)	25 ms	71.8%	0.5 W
Human V1 (estimate)	~0.001 mJ	50–100 ms	high	~20 W (whole brain)

The forgetting problem

Today's models are data-hungry and brittle: train one on a new task and it tends to overwrite the old, a failure called **catastrophic forgetting**. Underneath sits the **stability-plasticity dilemma**, the system must be plastic enough to absorb new information yet stable enough to preserve old knowledge. The brain solves it with **complementary learning systems**: a fast, plastic **hippocampus** that encodes new episodes like a scratchpad, and a slow, stable **neocortex** that gradually integrates them into its long-term model, largely during sleep through **memory replay**. Two speeds, so the new never clobbers the old.

ANALOGY

The hippocampus writes on a whiteboard, fast and erasable; the neocortex carves in stone, slow and permanent. Sleep is the nightly work of copying the day's whiteboard onto stone without cracking what is already there.

Making it algorithmic

Three families translate the biology. **Regularization**. **Elastic Weight Consolidation** first measures how important each weight was to the old task using the Fisher information F_i , then, when learning task B, penalizes changes to the important ones:

$$\mathcal{L}_{\text{EWC}}(\theta) = \mathcal{L}_B(\theta) + \frac{\lambda}{2} \sum_i F_i (\theta_i - \theta_i^*)^2,$$

so high-Fisher weights become stiff (stable) while the rest stay plastic. **Synaptic Intelligence** tracks importance online along the training path. **Replay**. **Experience replay** interleaves stored past samples with new data, echoing hippocampal replay. **Architecture**. **Progressive neural networks** grow a fresh column of weights per task and freeze the old ones, adding lateral connections so new tasks can reuse old features without overwriting them.

The surprise at scale

A twist: very large foundation models show *emergent* resistance to forgetting, without explicit EWC or replay. Parameter-efficient tuning like **LoRA** adapts only 0.1–1% of weights, leaving a frozen base intact; **in-context learning** absorbs a new task from the prompt with no weight update at all, a working-memory analogue; and curriculum training yields stable early layers with plastic late ones, mirroring complementary learning systems. One 540B-parameter model retained over **85%** on early tasks across 1,000 sequential tasks with no explicit continual-learning machinery, hinting that scale itself can partly dissolve the dilemma.

The integration problem

The brain is not one giant transformer. It is a society of specialized modules, vision, language, memory, planning, motor control, each with distinct wiring, tightly integrated and orchestrated by a central executive. The architectural frontier is to graduate from monolithic, single-purpose models to **cognitive architectures** that assemble such modules into a flexible whole.

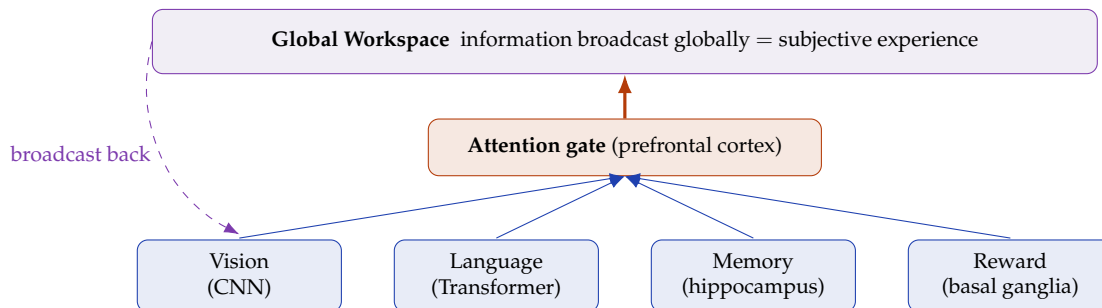
Global Workspace Theory

A leading account of how the brain integrates its modules is **Global Workspace Theory** (Baars). Specialized, largely unconscious modules compete to place their output on a central **workspace**, which then broadcasts that information globally to the rest of the brain; **attention** is the gatekeeper deciding what gets in, and the broadcast itself is proposed to be what we experience as consciousness. The blueprint for AI is direct: a vision system (CNN),

a language system (Transformer), an episodic memory (hippocampus-inspired), and a reward system (basal-ganglia-inspired), coordinated by an attentional controller that selects and broadcasts to a shared workspace for context-aware reasoning.

Symbols, and defining intelligence

Two threads point the way. **Neurosymbolic systems** fuse the pattern recognition of deep nets with the logical reasoning of symbolic AI; on multi-step visual question answering, splitting perception (neural) from reasoning (symbolic) lifted accuracy from **65%** (end-to-end neural) to **92%**, echoing the prefrontal cortex's blend of retrieval and rule-manipulation. And what would count as success? Legg (2008) proposed a **universal intelligence measure**, competence across a broad range of environments weighted by their simplicity, which mirrors the psychometric **g-factor** in humans: both say true intelligence is domain-general, measured across many tasks, not mastery of one.



A cognitive architecture: specialized modules compete through an attentional gate for a global workspace that broadcasts back to all of them.

The neuromorphic foundation model

The real prize is the fusion of all three frontiers: neuromorphic hardware running foundation-scale cognitive architectures with continual learning built in. Prototypes exist. A 2024 spiking transformer on Loihi 2 hit roughly **1/1000th** the energy per token of a GPU-based model at competitive quality; IBM's **North-Pole** reached about 4000x GPU energy efficiency on ResNet-50; and Australia's **DeepSouth** (2024) targets brain-scale simulation at 228 trillion synaptic operations per second in real time. The stated aim is whole-brain emulation of small mammals by around 2030, first as science, then as a substrate for cognitive systems.

Closing the loop, and its price

The cycle never ends with the brain merely inspiring AI. These same models are now neuroscience's

sharpest tools, decoding mental states from fMRI and EEG, powering large-scale circuit simulations, and, through representational similarity analysis, letting researchers compare a network's internal code directly against neural activity to test hypotheses. **But the stakes rise with the capability.** Ultra-low-power neuromorphic chips could enable pervasive, always-on surveillance; lifelong continual-learning agents raise control and alignment questions over long horizons; and whole-brain architectures force the hardest question of all, whether such a system could be conscious, and what moral status it would then hold. This is exactly where the **neurorights** of the previous chapter, mental privacy, personal identity, free will, become the central battleground.

Quick Self-Test

1. What is the von Neumann bottleneck, and how do neuromorphic systems avoid it?
2. In the crossbar equation $I_j = \sum_i V_i G_{ij}$, what physical quantity encodes a network weight, and why is the multiply "free"?
3. State the stability-plasticity dilemma and the brain's two-system solution.
4. In the EWC loss, what does the Fisher information F_i do to a weight that mattered for the old task?
5. What does Global Workspace Theory propose, and how does it map onto a cognitive architecture?

If You Remember One Thing

The next era of NeuroAI is defined by three brain-borrowed moves, compute where the data lives (neuromorphic), learn without overwriting (continual learning), and integrate specialized modules through a shared workspace (cognitive architecture). Their convergence promises AI that runs on milliwatts, learns for a lifetime, and reasons with human-like flexibility, which is precisely why the ethical guardrails must be ready before it arrives.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552