

Bridging Biology and AI

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 17 of "The NeuroAI Handbook"

How brains and machines take turns teaching each other, and what the brain still has left to teach.

One-sentence thesis. Neuroscience hands AI its best blueprints, and the AI those blueprints produce becomes neuroscience's best instrument for reading the very brain that inspired it.

PART 1: The Virtuous Cycle

Two Fields, One Loop

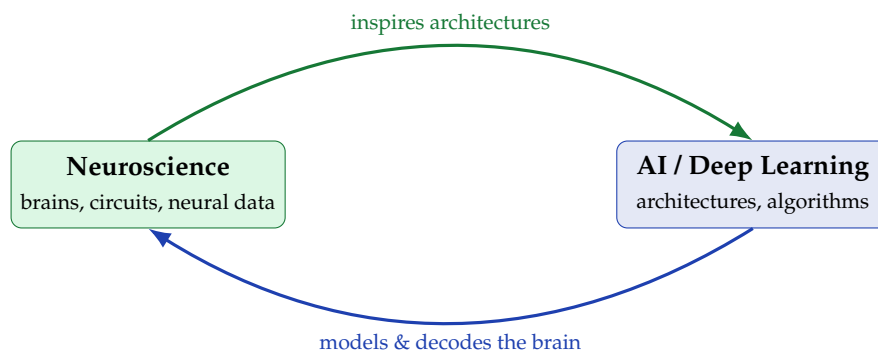
For most of their histories, neuroscience and AI borrowed from each other one direction at a time: a fact about the brain would suggest a machine, or a machine would offer a metaphor for the brain. NeuroAI's central claim is that the borrowing has closed into a loop. A discovery about how the brain is wired suggests an architecture; once that architecture is scaled up on real data and real compute, it turns out to be the most accurate model we have of the exact brain region that inspired it, and that model exposes the next discovery.

So the cycle runs in two directions at once. **Brain as**

inspiration for AI: neuroscience supplies blueprints and organizing principles for new learning systems. **AI as a tool for neuroscience:** deep networks model neural data, test theories of brain function, and accelerate discovery. The load-bearing idea of this chapter is that these are not two separate stories but one feedback loop, and the loop is speeding up.

ANALOGY

A blacksmith who forges a better hammer, then uses that hammer to forge a still-better one. Each turn of brain-into-AI yields a sharper instrument for turning brains back into science.



The virtuous cycle: neither field is the source; the source is the loop between them.

Vision: from Hubel and Wiesel to CNNs

Start with the oldest question in perception: how does a sheet of neurons turn light into “cat”? In the 1950s and 60s, David Hubel and Torsten Wiesel recorded from cat visual cortex and found that neurons in the primary visual area (V1) fire only for edges at a particular orientation, and that simple features are combined into complex ones as you ascend the hierarchy. Two principles fell out: **local receptive fields** (each unit sees a small patch) and **hierarchical processing** (edges to textures to parts to objects). Kunihiro Fukushima wired both into the Neocognitron (1980), and Yann LeCun refined it into the first modern CNNs in the 1990s.

Then the loop closed. Once AlexNet (2012), trained on ImageNet with GPUs, hit superhuman image accuracy, DiCarlo’s lab at MIT used deep CNNs as *in silico* models of the primate ventral visual stream. Layer-by-layer activations matched neural firing across visual cortex, and a trained CNN could predict a real neuron’s firing rate in area IT, explaining roughly 60% of the variance, beating every hand-built neuroscience model. **Watch out.** A good match is not identity: CNNs need millions of labels where a child needs a handful, collapse under adversarial pixels invisible to us, and lack the recurrence and attentional modulation of biological vision.

Reward: TD learning meets dopamine

Now a harder puzzle: how does an animal learn from a reward that arrives seconds after the action that earned it? Sutton and Barto’s answer in the 1980s was the **prediction error**, learning from the gap between the reward you expected and the one you got. Their **temporal-difference (TD)** rule made this precise. In symbols, the error is

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

with reward r_t , discount γ , and value estimate $V(s)$. Then Wolfram Schultz recorded dopamine neurons in monkeys and found they behaved exactly as δ_t

predicts: firing for an unexpected juice reward, then, once the animal learns, shifting to fire for the *cue* that predicts the juice, and dipping below baseline when a predicted reward is withheld. A pure algorithm had a measurable neural signature, and the **basal ganglia** became the brain’s known RL system. The loop kept turning: RL powered DeepMind’s AlphaGo, while giving neuroscience a language for addiction, a drug-driven, unnaturally large positive prediction error.

Attention: priority maps and Transformers

The intuition here is triage. The brain cannot process everything at full depth, so it must choose what to amplify. The parietal cortex, especially the lateral intraparietal area (LIP), builds **priority maps** that flag important locations, set by both bottom-up salience (a bright, moving object) and top-down goals (finding a friend in a crowd). The attention mechanism of Transformers (Vaswani et al., 2017) implements the same selective routing. Formally, queries, keys, and values combine as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

a soft, content-based lookup that decides which inputs matter now. This one principle became the backbone of BERT, GPT, and Vision Transformers.

Memory: replay, awake and asleep

Why does sleep help you learn? During rest and slow-wave sleep, the hippocampus **replays** recent experience: place cells that fired as a rat ran a maze re-fire the same sequence, often hundreds of times faster than real time, consolidating the memory into cortex. That biology inspired **experience replay** in deep RL: an agent stores past transitions in a buffer and resamples them, which breaks the temporal correlations that destabilize training and lets rare events be reused. **Prioritized experience replay** resamples high-TD-error (surprising) transitions more often, mirroring the brain’s bias toward replaying novel or rewarding events, and cut the frames needed to reach expert Atari play by 2–3x.

Domain	Brain finding	AI it inspired	How the loop closed
Vision	V1 edge detectors, hierarchy (Hubel & Wiesel)	Neocognitron → CNNs	CNNs became the best predictive model of area IT
Reward	Dopamine = TD error (Schultz)	TD learning, AlphaGo	RL now models addiction and decision-making
Attention	Parietal priority maps (LIP)	Transformer attention	Attention weights compared back to neural data
Memory	Hippocampal replay in sleep	Experience replay, PER	Prioritized replay mirrors novelty-biased replay

Lesson 1: Spend energy only when it matters

Begin with a humbling number. The human brain rivals a supercomputer on **20 watts**, less than a light-bulb, while training one large model can burn a small town's worth of power. Two brain principles explain the gap. **Sparsity**: at any instant only about **5%** of neurons in a region are active, whereas most artificial nets do dense, all-to-all math with nearly everything switched on. **Event-driven processing**: biological neurons spend energy only when they spike, while conventional nets grind through whole tensors on a fixed clock whether the data changed or not. Spiking neural networks and neuromorphic hardware are the direct attempt to import both.

Lesson 2: Learn fast, and do not forget

A child learns “cat” from a few examples and never has to unlearn “dog” to do it. Deep nets need millions of images, and worse, they suffer **catastrophic forgetting**: train on a new task and the old one is overwritten. The brain avoids this two ways. It is not a blank slate, its architecture carries evolved **inductive biases** that give learning a head start. And it runs **complementary learning systems**: the hippocampus encodes new episodes fast while the neocortex slowly integrates them into its world model, largely during sleep, so new facts do not clobber old ones. **The AI echo**. Elastic Weight Consolidation protects the

weights most important to old tasks (weighted by Fisher information), and meta-learning methods like MAML learn how to learn from few examples.

Lesson 3: Build modules, then integrate

The brain is not one monolithic processor. It is a society of specialized modules, vision, hearing, language, motor control, each with its own cell types and wiring, tied together by long-range connections and orchestrated by prefrontal control. That buys both specialization and flexibility. AI has trended toward giant end-to-end models, but is rediscovering modularity through **Mixture-of-Experts**, where a gating network routes each input to a few specialist sub-networks, much as different brain regions are recruited for different tasks.

An aside on learning rules

Backpropagation, the workhorse of deep learning, is biologically implausible, real neurons do not ship error signals backward. That gap motivates more brain-like rules: **predictive coding** (minimize prediction errors passed between hierarchical layers), **target propagation** (local targets instead of a global gradient), and **spike-timing-dependent plasticity** (synapses change based on the precise timing of pre- and post-synaptic spikes). None has displaced back-prop, but each hints that the brain reaches similar solutions by more local, distributed means.

PART 4: The Gaps That Remain

Where the brain is still ahead

The honest way to read the case studies is as a list of what AI still cannot do. Each gap is also a research direction, and a reason the loop is not finished.

Robustness. A few crafted stickers turn a stop sign into a “speed limit” sign for a deep net, while human vision shrugs the perturbation off. The brain’s robustness may come from **recurrent processing** (feedback refines feedforward guesses) and **predictive coding** (top-down predictions filter inconsistent inputs), neither of which a plain feedforward CNN has.

Compositionality. Having never seen a “purple elephant,” you can picture one instantly; nets trained on “red triangle” and “blue circle” often fail on “blue triangle.” The brain leans on compositional codes, **grid cells** that combine spatial scales, and even **concept cells** that respond to an abstract idea rather than a specific image.

Continual and causal reasoning. Humans transfer knowledge across tasks and build **causal** models that support “what if” simulation; current nets mostly capture correlation and forget. One vivid bridge is the **world model** (Ha & Schmidhuber, 2018): an agent that learns the dynamics of its environment can train a policy *inside its own dream* and transfer it back to reality, echoing the hypothesized role of offline replay and imagination in the brain.

The road ahead

The tools are converging from both sides. Neuropixels probes and higher-resolution fMRI are flooding neuroscience with data that only deep models can decode, and deep nets already stand as the best predictive models of sensory cortex. Meanwhile the brain remains the reference design for everything AI does badly. Brain-computer interfaces sit at the very center of the loop, since they demand solving a neuroscience problem (decode the signal) and an AI problem (adapt to non-stationary neural data) at the same time.

Quick Self-Test

1. State the two directions of the NeuroAI virtuous cycle, and give one concrete example of each.
2. Deep CNNs explain about 60% of the variance in area IT responses. Name two ways CNNs still differ from biological vision.
3. In the TD rule $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$, what happens to a dopamine neuron’s firing as an animal learns a cue-reward association?
4. The brain runs on 20 watts. Which two principles buy that efficiency, and what AI ideas import them?
5. What is catastrophic forgetting, and how do complementary learning systems (and their AI echo, EWC) address it?

If You Remember One Thing

The relationship between brains and machines is a loop, not a one-way borrowing. Every principle AI absorbs, hierarchy, prediction error, attention, replay, comes back as a sharper instrument for reading the brain, which exposes the next principle. The next breakthrough is more likely to come from the loop than from either field alone.

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