

Where Brains Meet Machines

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

A cross-cutting concept guide · The NeuroAI Handbook

How brain mechanisms map onto AI, where the mapping is tight, and where it quietly breaks.

One-sentence thesis. The book’s central throughline is a correspondence, not an equivalence: neuroscience and AI keep discovering the same computational solutions, but a resemblance is only trustworthy once you say which level of analysis it lives on and where it stops.

PART 1: The Caricature and Its Rules

A Brilliant Caricature

Begin with the object that started the whole dialogue: the artificial neuron. It is a deliberate simplification of the biological cell, one that keeps a single computational essence, integrate inputs, transform, emit an output, and throws away almost everything else: the ion channels, the dendritic nonlinearities, the metabolic cost, the spike. McCulloch and Pitts (1943) drew the first version as a binary threshold unit; Rosenblatt’s perceptron (1958) made it learn. Calling it a caricature is not an insult. A caricature exaggerates the one feature that matters and discards the rest, and that is exactly why the abstraction has been powerful enough to build modern AI and to serve as a testable theory of the brain.

The biological and artificial versions even differ in their governing equation. In symbols, a leaky integrate-and-fire neuron evolves a membrane voltage in continuous time, while an artificial unit is a single algebraic step:

$$\tau_m \frac{dV}{dt} = -(V - V_{rest}) + R_m I(t)$$

vs. $y = \max(0, \mathbf{w}^\top \mathbf{x} + b)$

Say Which Level You Mean

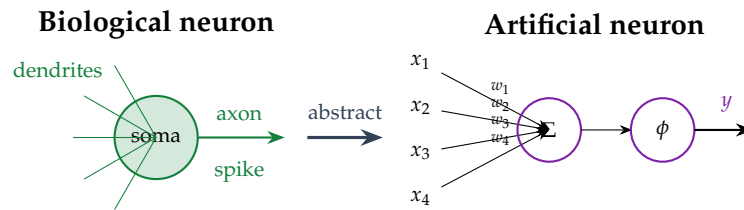
The discipline that keeps the analogy honest is David Marr’s three levels of analysis. The **computational**

level asks what problem is solved (recognize objects under varying conditions); the **algorithmic** level asks what representations and steps are used (hierarchical feature extraction); the **implementational** level asks how it is physically realized (neurons and synapses, or transistors and matrix multiplies). A brain and a network can share a computational goal and even an algorithm while differing completely in implementation. So the claim “this AI works like the brain” is empty until it names a level, and alignment at one level never guarantees alignment at the others.

Watch out. Depth is not homology. A deep stack of ReLU layers resembling cortex at the algorithmic level tells you nothing about mechanism at the implementational level. Respect the analogy, resist the mirage.

ANALOGY

An airplane does not fly by flapping its wings, yet it obeys the same aerodynamics as a bird. Same computational principle (lift), different implementation (fixed jet wing vs. flapping feathered one). Most brain-AI correspondences are exactly this kind: shared principle, divergent hardware.



The founding caricature: dendrites become weighted inputs, the soma's integration becomes a weighted sum, and the axon's spike becomes an activation ϕ . Everything biophysical is dropped on purpose.

Vision: A Hierarchy Rediscovered Twice

The clearest win begins with Hubel and Wiesel’s finding (1962) that neurons in primary visual cortex act as edge detectors, feeding higher areas that assemble more complex shapes, a hierarchy from V1 to V4 to inferotemporal (IT) cortex. Fukushima’s Neocognitron (1980) borrowed this directly, and LeCun’s convolutional networks turned it into engineering: local receptive fields and stacked feature detectors. Then the loop closed. Once AlexNet (2012) scaled, DiCarlo’s and Yamins’s labs found the layers of a CNN trained only on object classification predict firing in the ventral stream better than any hand-built neuroscience model, explaining roughly **60 percent** of the variance in IT. Early layers match V1, middle layers V4, late layers IT. A system optimized for one goal had independently rediscovered evolution’s solution.

ANALOGY

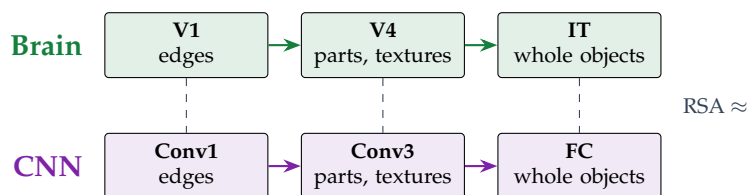
To ask whether a brain and a model “think alike,” do not compare neurons to units, compare their bookshelves. Give both the same movies: if Alice shelves by genre and Bob by director, their arrangements differ even with identical films. Representational Similarity Analysis compares these shelvings, the geometry of what-is-near-what, which is why it can align wetware with software at all.

Reward and Attention

Two more correspondences are almost eerily tight. In reinforcement learning, Sutton and Barto’s **temporal-difference** rule computes a prediction error; Schultz’s recordings (1997) found midbrain dopamine neurons fire in exactly this pattern, shifting from the reward to the cue that predicts it, and dipping when a promised reward is withheld. The abstract quantity and the biological signal obey the same equation:

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

Attention is the third. The parietal cortex (area LIP) builds priority maps that amplify relevant locations; the Transformer’s self-attention does the computational same thing, a soft, content-based lookup, $\text{softmax}(QK^\top / \sqrt{d_k}) V$. Both solved the problem of routing limited processing to what matters, and BrainScore now ranks architectures by how well they predict neural responses.



Layer-by-layer correspondence of the ventral stream and a task-trained CNN. The match is real and quantitative, but it is an algorithmic-level resemblance, not proof of shared mechanism.

Replay and Two Speeds of Learning

During sleep and quiet rest the hippocampus replays recent experience at high speed, and place cells that fired along a maze re-fire in sequence, consolidating memory into cortex. This inspired **experience replay** in deep reinforcement learning: agents store past transitions in a buffer and resample them, which breaks the harmful correlations between consecutive frames and lets rare events be learned more than once. The parallel even extends to what gets replayed, **prioritized experience replay** favors high-error, surprising transitions, mirroring the hippocampus's preference for novel or rewarding episodes, and cut the frames needed to reach expert Atari play by **2 to 3 times**.

Behind this sits the **complementary learning systems** idea (McClelland et al., 1995): a fast hippocampus that grabs single episodes and a slow neocortex that distills statistics over many exposures. AI recapitulated it in episodic control (MFEC, NEC), pairing an instance-based memory with a gradient-trained

network so a single rewarding experience can be exploited at once.

Place, Grids, and Embeddings

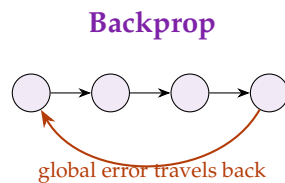
The brain's spatial code is the book's most vivid map. O'Keefe's **place cells** (1971) each fire at one location, a sparse "you are here" pin; the Mosers' **grid cells** (2005, a Nobel discovery) tile space in a hexagonal lattice at multiple scales, a universal coordinate system reused across environments. Their AI counterpart is the **learned embedding**: a distributed vector code where nearby points in the world map to nearby vectors. Framing the cognitive map as a graph, places as nodes and paths as edges, connects it to graph neural networks.

Watch out. Here the correspondence is looser than in vision. A learned embedding can carry map-like structure, but nothing forces it into crisp, single-peaked place fields or hexagonal grids; the similarity is functional (relational geometry), not a claim that the network grows grid cells.

PART 4: Where the Analogy Breaks

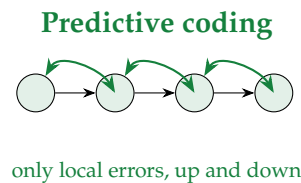
Credit Assignment: Backprop vs. the Brain

Both systems must solve the same problem, deciding which of millions of connections to blame for an error. Backpropagation is a mathematically exact answer, and it is the engine of deep learning, but it is widely regarded as biologically implausible for three concrete reasons: it needs the backward weights to mirror the forward ones, it requires storing every forward activation, and its error signals are non-local, traveling backward along paths real synapses do not have. The brain, if it approximates gradients at all, must do so differently. The book's candidates are **predictive coding**, where each level only ever computes local prediction errors (Rao & Ballard, 1999), plus feed-back alignment, target propagation, and three-factor Hebbian rules gated by neuromodulators. Hebbian learning is purely local; backprop is global; the gap between them is a live research frontier, not a solved mapping.



Codes, Signs, and Fragility

Three deeper divergences remain. First, **sparse spiking vs. dense rates**: biological neurons are event-driven, only about **5 percent** active at once, spending energy only on a spike, which is a large part of how the brain runs on **20 watts**; most artificial nets are dense and clock-driven, computing every unit every cycle. Spiking networks and neuromorphic hardware try to close this gap. Second, **Dale's principle**: a biological neuron releases the same transmitter at all its terminals and so is either excitatory or inhibitory, never both, whereas an artificial weight flips sign freely, an unremarked but real structural mismatch. Third, **robustness**: CNNs are fooled by adversarial perturbations invisible to people, need millions of labels where a child needs a few, and suffer catastrophic forgetting, none of which afflict the brain, whose recurrence, predictive feedback, and complementary memory the field is still trying to import.



Two answers to credit assignment. Backprop sends one global error signal backward along mirror-image weights; predictive coding trades it for local prediction errors exchanged between adjacent levels.

Brain mechanism	AI counterpart	Correspondence	Where it breaks
Ventral visual hierarchy	CNN layers (V1→IT)	Tight (algorithmic)	Adversarial, label-hungry, no recurrence
Dopamine prediction error	TD learning δ_t	Very tight	Brain uses many value signals
Parietal priority maps	Transformer attention	Tight	Softmax is not a circuit
Hippocampal replay	Experience replay	Functional	Buffer is not sleep
Place / grid cells	Learned embeddings	Loose	No forced grid structure
Cortical credit assignment	Backpropagation	Implausible	Non-local, weight-symmetric
Excit./inhib. (Dale)	Signed weights	Mismatch	Weights flip sign freely

Quick Self-Test

1. Why is the artificial neuron called a “caricature” rather than a “model” of the biological neuron, and what does that word buy you?

2. Someone says a deep network “works just like the brain.” Using Marr’s levels, what one question must you ask before the claim means anything?
3. What did Schultz’s dopamine recordings share with the TD learning rule, and why is this correspondence considered unusually tight?
4. Prioritized experience replay favors high-error transitions. What hippocampal phenomenon does that mirror, and is the match mechanistic or functional?
5. Give two concrete reasons backpropagation is considered biologically implausible, and name one biological alternative the book offers.
6. State Dale’s principle in one line and explain why an artificial weight violates it.

If You Remember One Thing

Brains and machines keep converging on the same computational solutions, hierarchical vision, prediction-error learning, attention, replay, because those solutions are good, not because silicon is wet. The value of the correspondence is entirely in its precision: name the level, cite the evidence, and mark where it stops. A correspondence you cannot break is not science; it is a mirage.

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