Synthesis paper

The Forgetting Paradox

How the brain's memory system is built to forget, and what that might mean for AI

The Central Contradiction

The same mechanism the brain uses to build memories is also the mechanism it uses to erase them, under the right conditions. That's the central argument of this paper, not a flaw in the design.

Adult hippocampal neurogenesis (AHN) is usually described in purely positive terms: new neurons born in the dentate gyrus improve pattern separation, the hippocampus's ability to tell similar experiences apart. But a large body of rodent research complicates that picture. Boost neurogenesis after a memory has already formed, and the memory doesn't get stronger. It gets less stable. The hippocampus forgets it faster (Akers et al., 2014).

The same process that sharpens new memories can erase old ones. That's closer to the point than a side effect.

This paradox touches three ideas: neurogenesis, neuroplasticity, and neuroprotection. It also maps directly, if imperfectly, onto two dominant frameworks in AI: backpropagation, the standard rule for updating a neural network from its errors, and predictive coding, a theory of the brain as a machine for minimizing prediction error. What follows tries to trace those connections honestly: where the evidence is strong, where it's contested, and where the biology just doesn't match the AI story.

Why Neurogenesis Causes Forgetting

In the dentate gyrus, newborn granule cells go through a weeks-long critical window where they're unusually excitable and synaptically plastic. During that window they get preferentially recruited into new memory traces, or engrams. Because these young cells respond to a wide range of inputs, they build representations that decorrelate overlapping experiences, a process called pattern separation, which cuts down interference between similar memories.

The destabilization runs on the same mechanism in reverse. When a fresh wave of neurons integrates into a circuit that already holds an established engram, they compete for synaptic resources and reshape the connectivity the original trace relied on. Existing representations get overwritten or degraded, something that looks a lot like what machine learning calls catastrophic forgetting.

The rodent evidence here is solid: suppressing neurogenesis genetically or pharmacologically after learning stabilizes existing hippocampal memories, while boosting it speeds up forgetting (Akers et al., 2014). This helps explain infantile amnesia, the near-universal fact that adults can't recall their early childhood. Neurogenesis peaks in infancy, so memories form and then the next wave of neurons washes them out. Whether the same mechanism drives forgetting in humans

the way it does in rodents is a separate, unresolved question. Human AHN is much harder to measure directly, and the debate over how much of it even occurs in adults is taken up later in this paper.

There's a functional logic to this. The hippocampus only has so much representational capacity, and clearing out old traces makes room for new discrimination while pushing consolidated material toward long-term cortical storage. Seen this way, forgetting looks less like a system failure and more like routine maintenance.

The Three Concepts and Their Relationship

Neurogenesis, neuroplasticity, and neuroprotection aren't independent processes. They're three angles on the same architectural problem: how do you build a system that keeps learning without wiping out what it already knows?

Neurogenesis adds new units. It's what expands the hippocampus's representational capacity and supplies the young, hyperplastic neurons behind the forgetting effect described above. How much of this actually happens in adult humans is still genuinely unsettled. Careful post-mortem studies disagree, and a lot of that disagreement comes down to methodology: how the tissue was fixed, how long after death it was processed, and so on.

Neuroplasticity changes the weights. It's the umbrella term for LTP and LTD at synapses, structural spine remodeling, shifts in intrinsic excitability, and glial and myelin plasticity, basically the machinery underneath all learning and memory. Without it, new neurons from neurogenesis would integrate and just sit there doing nothing.

Neuroprotection is what keeps the other two running. BDNF/TrkB signaling, Nrf2-driven antioxidant pathways, mitochondrial upkeep, and the vascular benefits of aerobic exercise all work to protect synapses and neurons against stress. It's the stability half of what's usually called the stability-plasticity dilemma.

Put simply: neurogenesis adds units, neuroplasticity adjusts the weights, and neuroprotection keeps both from degrading. Together they manage the basic tension every learning system faces: enough plasticity to take in new information, enough stability not to lose the old.

Mapping to AI: Where the Analogy Holds and Where It Breaks

With neurogenesis, neuroplasticity, and neuroprotection laid out as a system, the next question is how well that system maps onto ideas already in use in AI. The table below is the centerpiece of this paper. It's an attempt to be precise about where the biological and computational concepts genuinely line up, and where the analogy is flattering the biology or the AI, depending on which direction you read it from.

Biological concept	What it does in the brain	Closest AI analogy	Where they diverge

Neurogenesis	Adds units with critical-period hyperplasticity; drives pattern separation and adaptive forgetting.	Dynamic architectures, curriculum learning, regularization that favors forgetting.	Mainstream ANNs have fixed architectures and no critical-period rules; human AHN magnitude is still debated.
Neuroplasticity	Adjusts synapses, excitability, and myelin, the substrate of learning and memory.	Parameter updates (gradients), batch norm, attention as precision control.	Backprop is global, symmetric, and clocked; biology uses asynchronous, local, neuromodulated rules.
Neuroprotection	Preserves neurons and synapses via Nrf2, BDNF, mitochondrial quality, vascular support.	Regularization, noise injection, robustness training, model averaging.	AI has no direct metabolic or inflammatory equivalent. 'Protection' is abstract, not biophysical.

Backpropagation

Backpropagation is the workhorse of deep learning: compute the gradients of a global loss function and update all the weights accordingly. It's not biologically plausible. Real neurons can't implement the weight transport exact backprop requires (symmetric forward and backward connections), the precise phase separation, or the nonlocal error signals.

A few approximations close part of the gap. Feedback alignment swaps out exact transpose weights for random feedback matrices, and networks still learn, suggesting the cortex might not need perfect symmetry after all. Dendritic error models propose that apical dendrites carry top-down predictions while basal synapses respond to local prediction errors, a microcircuit setup that approximates gradient descent without needing global coordination. Equilibrium propagation gets at something similar through energy-based relaxation.

The connection to neurogenesis is indirect but real. AHN provides something like capacity expansion and dynamic sparsity in the dentate gyrus, adding specialized units that cut down representational overlap before downstream areas do their classifying. That's structurally similar to adding neurons to reduce interference in a network, but backprop itself doesn't explain why a biological system would add neurons instead of just adjusting existing weights. It's a different strategy for managing interference, one backprop doesn't predict.

Predictive Coding

Predictive coding is a closer architectural match. In the standard version, the brain works as a hierarchical generative model: higher areas send predictions downward, lower areas send prediction errors upward, and weights update to minimize free energy (weighted prediction error). Precision weighting, which modulates the gain on error signals, plays something like the role attention plays in transformer architectures.

There's a formal link to backprop worth noting: certain predictive coding formulations provably approximate backpropagation's gradients using only local Hebbian updates. Each layer holds prediction units and error units, and weight updates are driven entirely by local activity. That

makes predictive coding both more biologically plausible than backprop and mathematically compatible with it. It's less an alternative than a local-rule way of implementing it.

The link to AHN is looser but still resonant. If the hippocampus is a generative model trying to minimize prediction conflict between similar inputs, then AHN is a structural way of reshaping internal representations so similar inputs generate more separated internal causes, which produces smaller prediction errors downstream. Seen this way, AHN-driven forgetting isn't loss so much as revision: the generative model updating its structure to match new statistical patterns in the environment.

What Remains Unsettled

I don't have a tidy conclusion to offer here, because the underlying science doesn't have one yet.

The human AHN debate is genuinely unresolved. Two of the most-cited papers on the question, both using careful tissue-handling protocols, land on opposite conclusions about whether immature neurons are even detectable in a healthy adult human hippocampus. The current, cautious consensus is that AHN probably persists at low levels and declines with age and disease, but the actual magnitude, and so its functional significance, is still contested. Everything above about pattern separation and adaptive forgetting is well-established in rodents. Whether it translates to humans is plausible, not proven.

The predictive coding link to backprop is mathematically elegant, but it rests on specific network formulations that may not reflect how cortical microcircuits actually work. The Free Energy Principle, the broader theory behind all this, has also been criticized for being hard to falsify in its strongest form.

And the AI analogies in the table above are abstractions. Neuroprotection doesn't have a real computational equivalent. Neurogenesis in artificial networks exists as a research direction, but it's nowhere near standard practice. These analogies are useful, maybe even generative, but they shouldn't be mistaken for mechanistic claims.

The most honest conclusion I can offer is this: the forgetting paradox is real, the stability-plasticity tradeoff is the organizing problem underneath it, and how brains and AI systems ought to solve that problem is still an open question.

Key References

Akers, K. G., Martinez-Canabal, A., Restivo, L., Yiu, A. P., De Cristofaro, A., Hsiang, H-L., Wheeler, A. L., Guskjolen, A., Niibori, Y., Shoji, H., Ohira, K., Richards, B. A., Miyakawa, T., Josselyn, S. A., & Frankland, P. W. (2014). Hippocampal neurogenesis regulates forgetting during adulthood and infancy. Science, 344(6184), 598-602.

Sorrells et al. (2018). Human hippocampal neurogenesis drops sharply in children to undetectable levels in adults. Nature, 555, 377-381.

Duque & Spector (2019). A balanced evaluation of the evidence for adult neurogenesis in humans: implication for neuropsychiatric disorders. Brain Structure and Function, 224(7), 2281-2295.

Takeuchi, Duszkiewicz & Morris (2014). The synaptic plasticity and memory hypothesis. Philosophical Transactions B, 369, 20130288.

Whittington & Bogacz (2019). Theories of error back-propagation in the brain. Trends in Cognitive Sciences, 23, 235-250.

Scellier & Bengio (2017). Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. Frontiers in Computational Neuroscience, 11, 24.

Millidge, Tschantz & Buckley (2022). Predictive coding approximates backprop along arbitrary computation graphs. Neural Computation, 34(6), 1329-1368.

Fang, Aronov, Abbott & Mackevicius (2023). Neural learning rules for generating flexible predictions and computing the successor representation. eLife, 12, e80680.