INSTITUTE FOR PHYSICAL AI @ BMI
The Charlot Lab

Technical Report TR-2026-21
Survey / Position · v0.1
4 August 2026

LEARNING WITHOUT FORGETTING

The Adaptive Unit

Why a machine that keeps learning needs a richer neuron, what dendrites buy, and whether the better primitive can win on the hardware we already have.

Institute for Physical AI @ Bailey Military Institute

Charlot Lab · Adaptive Unit research topic

Drafted with AI assistance from public sources. Quantitative figures are labelled measured, reported or modelled, and every result is attributed to the paper that reported it rather than restated as our own.

Abstract. Almost every deployed neural network is trained once and then frozen, because training it on something new tends to destroy what it already knew. This is catastrophic forgetting, and it is the single largest obstacle to machines that work in the physical world, where the floor changes, the tool changes, and the failure is new. This report argues that the problem is being attacked at the wrong level. The usual remedies operate on the training procedure; the difference that matters is in the unit. A cortical neuron is not a weighted sum followed by a nonlinearity: its dendrites each apply their own nonlinearity, and matching one real neuron's input-output behaviour takes an artificial network five to eight layers deep. That same structure buys something the point neuron cannot offer, which is that different contexts can recruit different dendrites, routing each task to its own sparse sub-network so that learning a new skill barely touches the weights carrying the old one. We set out the mechanism, the measured evidence that it works on robot-manipulation benchmarks, and the objection that matters most: that a primitive better suited to the problem can still lose to one better suited to the hardware. We give the reasons to think this case is not the usual one, and state plainly what would show us wrong.

1. The capability embodied machines most lack

A language model can be trained once and shipped, because the distribution it serves is roughly the distribution it was trained on. A machine with a body cannot make that assumption. It meets a floor with a different friction, a gripper that has worn, a part presented at an angle nobody photographed. The useful response is to learn the new thing, and the constraint is that learning it must not cost the machine what it already knew.

Standard networks fail exactly here. Train one on a second task and performance on the first collapses, often toward chance, because both tasks are written into the same dense set of weights and the second overwrites the first. The phenomenon has been understood since the late 1980s and the field's usual response is avoidance: train once, freeze, and redeploy the whole model when the world moves. For a robot that response is not available, or rather it is available only as a return to the factory, which is a statement that the machine cannot adapt rather than a solution.

Most remedies work on the procedure. Rehearse old data alongside new, penalise movement in weights judged important, or allocate fresh capacity per task. Each helps and each carries a cost: storage that grows, a penalty that eventually freezes the network, or a parameter count that grows with the task list. What follows takes a different line, which is that the point neuron is an impoverished unit and some of the difficulty is downstream of that choice.

2. One neuron is already a network

The artificial neuron in ordinary use computes a weighted sum of its inputs and passes it through one nonlinearity. It is a deliberate simplification of a cell that does considerably more. A pyramidal neuron's dendrites are not passive wires summing what arrives; individual branches perform local nonlinear operations, and NMDA receptors in particular allow a branch to produce a regenerative response when enough of its own inputs are active together. Each dendritic branch is closer to a small unit in its own right than to a wire.

The size of the gap was measured directly. Beniaguev, Segev and London trained artificial networks to reproduce the input-output mapping of a detailed biophysical model of a layer-5 cortical pyramidal neuron, and asked how deep a network had to be before it could match.[1] The answer was a temporal convolutional network five to eight layers deep. The result that matters most for this report is the ablation: with NMDA-dependent dendritic nonlinearity removed, a far shallower network sufficed. The depth is not incidental to the cell's complexity, it is produced by the dendritic nonlinearity specifically.

> **The claim in one line.** Replacing a real neuron takes roughly a small network. The thing that makes it take a small network is the same thing that lets different contexts use different parts of the cell.

That coincidence is the whole argument. The dendritic nonlinearity is not only a source of representational capacity; it is a gate. A branch that fires only when its own inputs coincide is a branch that can be selectively recruited, and selective recruitment is the mechanism the next section needs.

3. What dendrites buy: routing, and therefore non-interference

Give an artificial unit a set of dendritic segments, each with its own weights, and let a context signal choose which segment gates the unit's output. The unit computes its ordinary weighted sum, and the winning segment modulates it. Nothing about this is exotic; the addition is a second input pathway carrying context, and a selection rule.

The consequence is that the network becomes a family of overlapping sub-networks indexed by context. Task A drives one sparse set of units through one set of segments; task B, arriving with a different context, drives a substantially different set. Because the two sub-networks share few active parameters, gradient updates from task B land largely where task A's competence is not stored. Forgetting falls, not because the optimiser has been restrained, but because the two skills were never written on top of each other.

Sparsity is the control that makes this trade explicit. Very sparse activation means minimal overlap and therefore minimal interference, but also less capacity available to any one task. Denser activation gives each task more of the network at the price of collisions with its neighbours. Neither end is right in general, and the setting is a real hyperparameter rather than a detail, which is why the Institute's simulation of this material exposes it as a slider rather than describing it in prose.[5]

Two things are worth stating plainly so the mechanism is not oversold. First, this is not memory in the biological sense and nothing here is recalled; it is interference avoidance through structure. Second, it requires a context signal. Where the task identity is known, this is trivially available; where it is not, it has to be inferred, and inferring it is a real open problem rather than a detail of implementation.

4. The measured evidence

The mechanism is attractive on paper, so the relevant question is what has actually been measured, by whom, and on what.

The strongest result for our purposes comes from Numenta. Iyer and colleagues added active dendritic segments with context-dependent gating to standard networks and evaluated them on two settings.[2] On a continual-learning benchmark of one hundred sequential tasks, the dendritic network reached roughly 81 per cent accuracy where a comparable standard network degrades severely. On a multi-task reinforcement-learning benchmark of ten simulated robot-manipulation tasks drawn from Meta-World, it reached roughly 88 per cent success. The second is the one this laboratory weights most heavily, because it is manipulation rather than a permuted-image benchmark, which is to say it is the setting the argument is actually about.

Chavlis and Poirazi report a complementary result from the representational side rather than the continual-learning side, finding that dendritic structure in artificial networks buys accuracy and robustness at markedly lower parameter counts than the fully connected equivalent.[3] Read together with the Numenta work, the two say that

the richer unit is not merely a device for avoiding forgetting; it is a more efficient use of parameters in its own right.

These are reported results from their respective groups. To our knowledge neither has been the subject of a large independent replication effort, and the honest grade for both is therefore *reported* rather than *verified*. That is not a criticism of the work; it is the ordinary state of a young literature, and it is recorded here because a reader deciding how much weight to place on the argument is entitled to know.

5. The objection that matters: the hardware lottery

A better idea does not automatically win. Hooker's argument, which we take seriously because the history supports it, is that research directions succeed in part because they happen to suit the hardware and software of their moment, and that ideas requiring different machinery can lose for reasons that have nothing to do with their merit.[4] The canonical case is the one that made the present era: backpropagation on dense matrices was not obviously the best approach, but it was the approach that matched hardware built for dense matrix multiply, and the match compounded for a decade.

Dendritic computation is exactly the kind of idea that argument should worry about. It is sparse where the hardware likes dense, conditional where the hardware likes uniform, and structured per unit where the hardware likes homogeneous blocks. A sparse network that touches ten per cent of its units does not automatically run ten times faster on a machine whose throughput comes from doing the same operation on everything at once.

Three things nonetheless argue this case is not simply the lottery repeating.

The gating is cheap and the arithmetic is unchanged. Active dendrites add a second small linear map and a selection; the bulk of the computation remains the dense operation the hardware is built for. This is a modification of the unit, not a replacement of the numerical kernel, which is what makes it implementable now rather than after a hardware generation.

Sparsity has become the industry's own direction. Mixture-of-experts routing is conditional computation under a different name and is now standard in frontier models. The infrastructure argument against conditional execution has weakened considerably, because the largest deployed systems already do it.

The sparse, local, brain-shaped primitive has now been shown to scale. The strongest recent evidence is Kosowski and colleagues' Dragon Hatchling architecture, a scale-free network of locally interacting neuron particles whose activation vectors are sparse and positive and whose working memory relies on Hebbian synaptic plasticity at inference rather than on gradient updates.[6] They report performance rivalling GPT-2 at the ten-million to one-billion parameter range, and, notably for a different reason, that interpretability of state is a property of the architecture rather than something recovered afterwards. We cite this as external work we situate against

rather than as our own result, and at the bar its authors set rather than the one its publicity sets. Its significance here is narrow and important: it is evidence that a primitive of this family can be made to scale on ordinary hardware, which is precisely the claim the lottery objection denies.

6. What is open

Several things this argument depends on are unsettled, and they are listed so it can be attacked where it is weak rather than where it is strong.

Context inference. Nearly every result above supplies task identity to the network. An embodied machine is not told which task it is in. Whether context can be inferred online, from the observation stream, well enough to drive routing without a supervisor is the gap between this literature and a working robot, and it is the largest one.

Independent replication. The headline continual-learning and manipulation numbers come from the groups that proposed the methods. Independent reproduction on the same benchmarks, published whichever way it falls, would do more for this area than another architecture.

Measured cost, not modelled cost. We are not aware of a careful published accounting of the wall-clock and joules-per-task cost of dendritic gating against a matched dense baseline on the same device. Without it, the hardware-lottery question stays a matter of argument when it could be a matter of measurement. That measurement is tractable for anyone with a device and a fortnight, and the Institute would rather see it exist than be the only party to hold it.

How far the biology carries. That a cortical neuron requires a deep network to imitate does not by itself establish that artificial units should be dendritic. Biology optimises under constraints artificial systems do not share. The measured continual-learning results are what carry the argument; the neuroscience explains why the mechanism was worth trying, and should not be asked to do more work than that.

7. Why this is taught rather than only published

Catastrophic forgetting is a phenomenon nobody believes at the right depth until they watch it happen. The Institute therefore teaches this material as something to run rather than read: a retention matrix in which a point-neuron network's early tasks visibly redden toward chance while a dendritic-gated network's stay bright, a sparsity control that makes the capacity-versus-interference trade tangible, and a toggle that collapses a single neuron's equivalent depth when the NMDA nonlinearity is removed.[5] A companion visualisation draws a real network in full, every sphere a neuron and every line a weight, and lets a reader switch task context and watch a different sparse sub-network light up for the same input. The full treatment is PAI-135.[7]

The reason for insisting on the runnable version is narrow. The claim of this report is about which parts of a network change when it learns something new. That is a claim about a picture, and a picture is better shown than described.

Notes and sources

1. Beniaguev, D., Segev, I., London, M. Single cortical neurons as deep artificial neural networks. *Neuron* 109(17) (2021). A temporal convolutional network five to eight layers deep is required to match an L5 pyramidal neuron's input-output mapping; the ablation shows NMDA-dependent dendritic nonlinearity is what produces the required depth.
   `verified: peer-reviewed, widely cited`
2. Iyer, A., Grewal, K., Velu, A., Souza, L. O., Forest, J., Ahmad, S. Avoiding catastrophe: active dendrites enable multi-task learning in dynamic environments. *Frontiers in Neurorobotics* 16 (2022). Numenta. Approximately 81% on 100 sequential permuted-MNIST tasks; approximately 88% success on 10 Meta-World robot-manipulation tasks.
   `reported: peer-reviewed, not independently replicated at scale to our knowledge`
3. Chavlis, S., Poirazi, P. Dendritic structure in artificial neural networks yields accurate, robust and parameter-efficient learning. *Nature Communications* (2025).
   `reported: peer-reviewed; complementary to [2] on the representational rather than continual-learning axis`
4. Hooker, S. The hardware lottery. *Communications of the ACM* 64(12) (2021); arXiv:2009.06489.
   `verified: peer-reviewed, the standard reference for this argument`
5. Institute for Physical AI @ BMI. *The Adaptive Unit* research topic and its on-device simulations: the retention matrix, the NMDA depth toggle, and "The Network, Seen". Runnable in the browser, no installation.
   `primary: Institute teaching material`
6. Kosowski, A., et al. The Dragon Hatchling: the missing link between the transformer and models of the brain. arXiv:2509.26507 (2025); code at pathwaycom/bdh. Cited at the paper's own stated claims.
   `reported: preprint; cited as external work this report situates against, not as a result of ours`
7. Institute for Physical AI @ BMI. PAI-135, *Dendritic Computation: Learning Without Forgetting*. Three modules, six lessons, runnable benches.
   `primary: Institute teaching material`

Technical Report TR-2026-21 · Institute for Physical AI @ Bailey Military Institute · Charlot Lab. Companion material: the research topic *The adaptive unit*, the retention-matrix and network visualisations, and PAI-135. Results attributed above belong to the groups that reported them; where a figure is reported rather than independently verified, the reference says so. Corrections are welcome and will be recorded rather than quietly applied.