NeuroAI and Beyond: Bridging Between Advances in Neuroscience and Artificial Intelligence

Anthony Zador[24,†], Jean-Marc Fellous[10,26,†], Terrence Sejnowski[10,22,†], Gina Adam[1], James B Aimone[2], Akwasi Akwaboah[3], Yiannis Aloimonos[4], Carmen Amo Alonso[5], Chiara Bartolozzi[6], Michael J. Bennington[7], Michael Berry[8], Bing W. Brunton[9], Gert Cauwenberghs[10], Hillel J. Chiel[11], Tobi Delbruck[12], John Doyle[13], Jason Eshraghian[14], Ralph Etienne-Cummings[3], Cornelia Fermüller[4], Matthew Jacobsen[4], Ali A. Minai[15], Barbara Oakley[16], Alexander G. Ororbia II[17], Joe Paton[18], Blake Richards[19], Yulia Sandamirskaya[20], Abhronil Sengupta[21], Shihab Shamma[4], Michael P. Stryker[23], Seong Jong Yoo[4], Steven W. Zucker[25]

Abstract—Neuroscience and Artificial Intelligence (AI) have made impressive progress in recent years but remain only loosely interconnected. Based on a workshop convened by the National Science Foundation in August 2025, we identify three fundamental capability gaps in current AI—the inability to interact with the physical world, inadequate learning that produces brittle systems, and unsustainable energy and data inefficiency—and describe the neuroscience principles that address each: co-design of body and controller, prediction through interaction, multi-scale learning with neuromodulatory control, hierarchical distributed architectures, and sparse event-driven computation. We present a research roadmap organized around these principles at near-, mid-, and long-term horizons. We argue that realizing this program requires a new generation of researchers trained across the boundary between neuroscience and engineering, and describe the institutional conditions—interdisciplinary training, hardware access, community standards, and ethics—needed to support them. We conclude that NeuroAI, neuroscience-informed artificial intelligence, has the potential to overcome limitations of current AI while deepening our understanding of biological neural computation.

[1]Dept. Electrical & Computer Engineering, George Washington University, Washington, DC, USA. [2]Neural Exploration & Research Laboratory, Sandia National Laboratories, Albuquerque, NM, USA. [3]Dept. Electrical & Computer Engineering, Johns Hopkins University, Baltimore, MD, USA. [4]Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA. [5]Dept. Aeronautics & Astronautics, Stanford University, Stanford, CA, USA. [6]Event-Driven Perception for Robotics, Italian Institute of Technology, Genoa, Italy. [7]Dept. Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. [8]Dept. Neuroscience, Princeton University, Princeton, NJ, USA. [9]Dept. Biology, University of Washington, Seattle, WA, USA. [10]Institute for Neural Computation, UC San Diego, La Jolla, CA, USA. [11]Depts. Biology, Neurosciences & Biomedical Engineering, Case Western Reserve University, Cleveland, OH, USA. [12]Inst. of Neuroinformatics, UZH-ETH Zurich, Zurich, Switzerland. [13]Dept. Control & Dynamical Systems, Caltech, Pasadena, CA, USA. [14]Dept. Electrical & Computer Engineering, UC Santa Cruz, Santa Cruz, CA, USA. [15]Dept. Electrical & Computer Engineering, University of Cincinnati, Cincinnati, OH, USA. [16]Dept. Engineering, Oakland University, Rochester, MI, USA. [17]Dept. Computer Science, Rochester Institute of Technology, Rochester, NY, USA. [18]Champalimaud Neuroscience Programme, Champalimaud Foundation, Lisbon, Portugal. [19]School of Computer Science + Montreal Neurological Institute, McGill University & Mila, The Quebec AI Institute. [20]Institute of Computational Life Sciences, Zurich University of Applied Sciences, Wädenswil, Switzerland. [21]School of Electrical Engineering & Computer Science, Penn State University, University Park, PA, USA. [22]Dept. Neurobiology, UC San Diego, La Jolla, CA, USA. [23]Dept. Physiology, UC San Francisco, San Francisco, CA, USA. [24]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [25]Dept. Computer Science and Biomedical Engineering, Yale University, New Haven, CT, USA. [26]Mathematics Department, Kalaheo High School, Kailua, HI, USA. [†]Corresponding authors: Jean-Marc Fellous (jmfellous@ucsd.edu); Terrence Sejnowski (terry@snl.salk.edu); Anthony Zador (zador@cshl.edu).

Authors emails: Gina Adam (ginaadam@email.gwu.edu); James B Aimone (jbaimon@sandia.gov); Akwasi Akwaboah (aakwabo1@jhu.edu); Yiannis Aloimonos (jyaloimo@umd.edu); Carmen Amo Alonso (camoalon@stanford.edu); Chiara Bartolozzi (chiara.bartolozzi@iit.it); Michael J. Bennington (mbenning@andrew.cmu.edu); Michael Berry (berry@princeton.edu); Bing W. Brunton (bbrunton@uw.edu); Gert Cauwenberghs (gcauwenberghs@ucsd.edu); Hillel J. Chiel (hjc@case.edu); Tobi Delbruck (tobi@ini.uzh.ch); John Doyle (doyle@caltech.edu); Jason Eshraghian (jsn@ucsc.edu); Ralph Etienne-Cummings (retienne@jhu.edu); Cornelia Fermüller (fermulcm@umd.edu); Matthew Jacobsen (mattjac@umd.edu); Ali A. Minai (ali.minai@uc.edu); Barbara Oakley (barbo8@gmail.com); Alexander G. Ororbia II (agovcs@rit.edu); Joe Paton (joe.paton@neuro.fchampalimaud.org); Blake Richards (blake.richards@mila.quebec); Yulia Sandamirskaya (sank@zhaw.ch); Abhronil Sengupta (sengupta@psu.edu); Shihab Shamma (sas@umd.edu); Michael Stryker (michael.stryker@ucsf.edu); Seong Jong Yoo (yoosj@umd.edu); Steven W. Zucker (steven.zucker@yale.edu).

Current draft date: 2026-04-22

1 INTRODUCTION

The major leaps in artificial intelligence (AI) share a common origin: neuroscience. McCulloch and Pitts modeled the neuron as a logical gate; Rosenblatt turned that model into a learning machine. Hubel and Wiesel mapped the hierarchical organization of visual cortex; Fukushima and later LeCun translated it into convolutional networks. Schultz discovered that midbrain dopamine neurons encode reward prediction errors; Sutton and Barto had independently formalized the same signal in temporal difference learning. The attention mechanisms at the core of modern transformers draw on the cortical principle that processing resources must be allocated selectively to relevant inputs. Each of these, and many more, was a specific computational insight extracted from a specific neuroscience finding, then engineered into an algorithm [1, 2, 3, 4]. The deep learning revolution descends from this lineage.

Industry is now investing hundreds of billions of dollars per year in AI. Frontier models improve rapidly. Large language models pass professional examinations, write and debug functional code, and generate images and videos from natural language descriptions. Some leaders in the field predict artificial general intelligence within a few years. These achievements are substantial, and scaling current approaches may well continue to produce gains.

Nearly all of this investment, however, is concentrated on a single architectural paradigm: transformers trained on massive static datasets, running on dense synchronous digital GPU computation. The bet is also conceptually narrow: current systems route all cognition through language, which was a latecomer in brain evolution, and despite evidence that language and reasoning are subserved by anatomically distinct brain systems [5], suggesting that language-centric architectures may face limits that benchmarks within the language domain will not reveal. The major architectural ideas in AI's history—from perceptrons to convolutional networks to transformers—each brought capabilities the previous frameworks could not achieve, and none emerged from scaling what came before. Each of those ideas originated in academic research, and most were informed by neuroscience. Industry scaled them but did not generate them. The question facing the field now is whether the next generation of architectural ideas will be ready when the current paradigm reaches its limits.

Three fundamental capability gaps are already visible, and they are unlikely to close through scaling alone. First, AI cannot interact with the physical world: it passes the bar exam but cannot clear a dinner table and wash the dishes. Second, AI does not learn the way animals do—continuously, adaptively, with graceful fallback under novelty—and this makes it brittle. Third, AI is inefficient in energy and data to a degree that constrains who can build it and where it can be deployed. These gaps are architectural. They reflect deep mismatches between how current AI systems are designed and how biological intelligence works. They are also the problems that brains evolved to solve.

Neuroscience is now positioned to address them with a specificity that was not previously possible. The complete connectome of the fruit fly brain has been mapped at synaptic resolution. Large-scale electrophysiology using Neuropixels probes allows simultaneous recording from tens of thousands of neurons across multiple brain areas in behaving animals. Optogenetic tools enable causal manipulation of identified cell types. Comprehensive behavioral tracking links neural activity to behavior with a precision unavailable a generation ago. For the first time, we can extract specific algorithmic insights from biological computation about how neural circuits achieve robust, efficient, and adaptive performance.

We lay out three promising foundations for the next generation of AI. We describe the three capability gaps in detail and identify the neuroscience principles that address each one. We present a research roadmap for the coming decade, organized by concrete milestones. And we describe the institutional changes, in training, infrastructure, and community organization, needed to realize this program. The work we describe complements industry's current efforts, which are focused on optimizing the present paradigm. The research program we propose here develops the architectural alternatives that will be needed next, whether the current paradigm plateaus in three years or in fifteen.

This paper emerged from a workshop convened by the National Science Foundation in August 2025, bringing together about seventy researchers and program managers from neuroscience, artificial intelligence, robotics, neuromorphic engineering, and cognitive science [6]. The workshop was organized around five thematic areas—embodied cognition and computation, language and communication, robotics, learning in humans and machines, and neuromorphic AI engineering—but the discussions repeatedly converged on a smaller set of cross-cutting principles. We present those convergences here [7].

2 THREE PROBLEMS

We organize the limitations of current AI into three categories: the inability to interact with the physical world, a mode of learning that is both inflexible and fragile, and an inefficiency in energy and data that limits who can build AI and where it can operate. For each, we describe the capability gap, explain why scaling current approaches is unlikely to close it, and identify the neuroscience principles that point toward a solution.

2.1 AI cannot interact with the physical world

In 1988, Hans Moravec observed that the tasks humans find difficult—chess, calculus, logical reasoning—are easy for computers, while the tasks we find effortless—walking, catching a ball, recognizing a face across a room—are extraordinarily hard [8]. Nearly four decades later, the pattern holds in a new form. AI systems can now pass medical licensing exams, generate publishable prose, and write software. But decades into robotics research, the most capable robot in a typical household is a vacuum puck that bumps into furniture. No existing system can crack an egg, separate the yolk from the white, and make an omelet in an unfamiliar kitchen.

The reason is that physical interaction with the world poses a different kind of computational problem from the ones current AI is optimized for. Real-world action requires continuous control in real time, with noisy and partial sensor data, physical consequences for errors, and no option to reset. The agent must handle objects it has never seen, surfaces with unknown friction, forces it did not expect. These are the conditions under which biological intelligence evolved, and every animal on earth—from insects navigating turbulent air to octopuses manipulating objects with eight flexible arms—solves them routinely.

Current AI systems are trained on static datasets: text, images, video. They observe the world passively, through data collected and curated by others. Whatever internal representations large language models (LLMs) develop from this passive exposure, those representations were not built through the agent's own causal interaction with physical reality. A child learning the word "heavy" has spent months picking things up before encountering the word. The meaning is anchored in proprioceptive and tactile experience, and is augmented by auditory experience when the object falls. An LLM's representation of "heavy" is anchored in the co-occurrence statistics of text.

Neuroscience offers two relevant constructive principles. The first is *co-design of body and controller*. Biological brain architectures did not evolve in isolation. They co-evolved with the bodies they control, producing tight coupling between morphology and neural circuitry [9]. A fly's visual system is tuned to the optic flow statistics generated by its particular flight dynamics. An owl's auditory processing is calibrated to the interaural time differences produced by its head geometry. The computational problems the brain solves are shaped by the body it inhabits [10].

This co-evolution has concrete engineering consequences. Clever body design offloads computation from the neural controller. Tendon-driven limbs exploit passive dynamics—springs and dampers in the musculoskeletal system—to store and release energy during locomotion, reducing the demands on the nervous system. A human fingertip contains roughly 240 mechanoreceptors per square centimeter, providing dense information about texture, pressure, slip, and vibration. This massive peripheral sensing, integrated with proprioception and vision through tightly coupled neural circuits, is what makes human manipulation so adaptive. The retina illustrates the same principle at a different scale: rather than transmitting raw images, it performs edge detection, motion detection, contrast normalization, and prediction of upcoming stimuli, compressing information by orders of magnitude before it reaches the brain. Neuromorphic sensors, particularly dynamic vision sensors that encode luminance changes asynchronously rather than capturing static frames, represent an important engineered step toward this biological strategy. The next step is to combine (as in biology) these transients with the sustained information in frames to emulate the computations performed by the retina and early visual cortex, enabling efficient algorithms and, ultimately, hardware implementations that provide biologically grounded primitives on which higher-level processes can build.

The physical structure of the body shapes the control problems it must solve. The soft arms and suckers of the octopus are part of its solution for handling complex materials; more generally, embodied intelligence offloads part of the control problem onto the periphery. The size and speed of a behavior alter the relevant dynamics, and biological nervous systems adjust their control strategies accordingly [11].

Modern robots violate the co-design principle almost entirely. Most are rigid, over-actuated, and energy-inefficient, designed this way because rigid bodies are simpler to model mathematically. The control system and the physical hardware are typically designed by separate teams with separate objectives. The result is machines capable of precise repetitive motions in controlled environments but poor at adapting to novel objects, surfaces, or forces.

The second principle is *prediction through interaction*. The brain is organized around prediction at every level of the sensory hierarchy. Retinal ganglion cells respond most strongly when the visual scene violates their expectations, transmitting prediction errors rather than raw data. Predictive coding models [12, 13] propose that each level of the cortical hierarchy maintains a generative model of its inputs, passing only the discrepancies upward. If this framework is approximately correct, the brain's primary computational strategy is to predict the next state of the world conditioned on its own actions, and to learn by correcting those predictions.

This has direct implications for how AI systems should be trained. Infants learn in a developmental sequence: first building representations through passive observation, then learning cause and effect by incorporating their own actions as motor control matures, then constructing causal models through active experimentation. This progression—from passive prediction to action-conditioned prediction to causal reasoning—suggests a natural curriculum for training embodied AI agents, and aligns with current proposals like LeCun's Joint Embedding Predictive Architecture [14], which learns by predicting future states in abstract representation space rather than in pixel or token space.

Physical robots are essential to this program. They are not just consumers of NeuroAI principles; they generate the questions that sharpen them. Building a robot forces precise definitions of otherwise vague concepts. When a neural architecture designed for one body plan is applied to a robot with a fundamentally different morphology, the resulting failures expose assumptions about brain-body coupling that purely computational work leaves untested. Robots reproduce the actual operating conditions of biological intelligence—partial data, continuous control, sparse reward, real consequences—in ways that curated benchmarks cannot.

Interacting with the physical world includes interacting with others, be they humans or machines. Current AI LLMs can emulate empathy, but do not yet possess it, resulting in sometimes disastrous consequences. Humans may be fooled into thinking the chatbot cares for them, when in fact it does not. AI systems do not have a genuine sense of values, or of right and wrong. They currently do not have artificial emotions, the functional equivalent of human or animal emotions. Yet such emotions serve many computational purposes including providing feedback that drives learning [15]. A dog trained by a scolding tone modifies its behavior; a presenter corrected in public feels embarrassment and avoids the same mistake. Current LLMs apologize graciously when told they are wrong, but cannot use the emotional consequences of their errors as a learning signal to improve.

2.2 AI learning is limited

AI systems cannot learn from experience after deployment. A physician improves with every patient seen. A factory worker notices when a machine starts sounding different and adjusts. A self-driving car that encounters an unusual intersection does not learn from it. A medical imaging classifier trained at one hospital degrades when deployed at another, where the scanners, patient demographics, and imaging protocols differ, and it cannot self-correct. The model is generally frozen the moment training ends. Whatever it knew at that point is all it will ever know, until a successor model is retrained.

A related failure is that current systems do not represent what they do not know. Hallucinations—the confident generation of false or fabricated information—are the most visible symptom. When an LLM's training data is sparse or contradictory on a topic, the system produces fluent confabulation with no internal signal that anything is wrong. A human who has never done plumbing knows they don't know plumbing. Current AI has no comparable self-model of its own competence boundaries, or of its confidence level in its answers, because it has no episodic record of what it has and has not experienced and no grounded world model against which to check its outputs.

Within the domain of language, frontier LLMs have become impressively robust. They handle paraphrased questions, ambiguous prompts, and unusual formatting with a fluency that would have seemed impossible five years ago. The same is becoming true of image or audio generation. Outside these domains, the picture is different. Self-driving systems still fail at edge cases with fatal consequences. Robotic manipulation breaks when objects differ modestly from training examples. Industrial monitoring systems degrade when manufacturing conditions drift. These are domains where the system encounters the world as it actually is, rather than as a curated dataset represents it.

The field has recognized the memory problem and responded with engineering workarounds. Retrieval-augmented generation (RAG) gives models access to external documents at inference time. Context windows have grown from thousands to millions of tokens. Tool use lets models query databases and APIs. Fine-tuning adapts models to new domains. These help, but they are workarounds for an architectural gap, not solutions to it. RAG retrieves documents but does not change what the model knows in the long run. A longer context window is still a fixed buffer, wiped between sessions. Fine-tuning degrades previously learned knowledge and must be performed offline. None of these approaches give the system the ability to update its internal model of the world in response to ongoing experience. They simulate memory without implementing learning.

These failures, the inability to learn after deployment, the absence of calibrated uncertainty, the brittleness in physical domains, the silent degradation, share a common architectural root: current AI has no mechanism for continual, multi-timescale learning and no hierarchical organization that separates fast safe responses from slow flexible ones. Neuroscience offers specific principles for addressing both [16].

Multiple memory systems operating at different timescales. Biological learning spans a continuous spectrum: sensory habituation in seconds, motor adaptation over minutes to hours, episodic and semantic memory consolidation over days to years, evolutionary shaping of neural architecture over millennia. Each timescale serves a distinct function, and the interaction between them is what produces both adaptability and stability. Long-term working memory used during cognitive processing lasts for hours but only salient and attended events are retained in long-term memory [17]. The hippocampus learns rapidly, forming sparse, pattern-separated representations of individual episodes within one or a few exposures. The neocortex learns slowly, integrating across many episodes to extract statistical regularities. Sleep-based replay of hippocampal activity transfers memories to cortical representations gradually, enabling fast learning and stable long-term memory to coexist because they are implemented in separate but communicating structures. Sleep allows for selective forgetting of likely useless information and for the consolidation of likely important memory episodes and facts. Both are important components of continuous lifelong learning. This is the core insight of complementary learning systems theory, and it provides an architectural solution to the stability-plasticity dilemma that plagues all current approaches to continual learning in AI.

Neuromodulatory systems add context-sensitive control over this architecture. Dopaminergic neurons broadcast reward prediction error signals that gate plasticity in the striatum and prefrontal cortex. Noradrenergic neurons in the locus coeruleus signal arousal and novelty, shifting the balance between exploiting known strategies and exploring alternatives. Acetylcholine levels control the timing and gating of synaptic plasticity, in addition to enabling muscle contraction and movement, crucial aspects of learning and memory in many tasks where behavioral actions are used to measure learning. These neuromodulatory signals interact in ways that we are only beginning to understand [18]. These signals specify whether to learn and how much, operating differentially across brain regions and synaptic populations, with a nuance that current optimization algorithms do not approach [19]. The spectrum of synaptic timescales provides a further dimension: some synapses change in seconds through short-term facilitation and depression, others require minutes of sustained activity, still others change only after hours of protein-synthesis-dependent consolidation. Fast synapses capture immediate statistical structure; slow synapses accumulate evidence and resist transient fluctuations. "Fast weights," transient synaptic modifications that store recent context in associative memory, provide a neurally plausible mechanism for algorithmic features such as the context window in a transformer [17], but one that emerges from synaptic dynamics and could be implemented in neuromorphic hardware. Beyond weight changes, structural plasticity, apoptosis, and adult neurogenesis modify the network's topology over time, expanding representational capacity in a way that has no analogue in current deep learning, where architecture is fixed at initialization.

Hierarchical fallback architecture. Robustness in biological systems arises from layered, distributed control. The brain's motor system is a stack of specialized layers operating at different timescales: the spinal cord mediates reflexes in milliseconds, the cerebellum calibrates movements over seconds to minutes, the basal ganglia select among competing action plans, and the motor cortex and prefrontal cortex handle planning over longer horizons. Each layer occupies a distinct neural structure with its own learning rules and network dynamics.

The vestibulo-ocular reflex (VOR) illustrates how these layers interact. The VOR stabilizes gaze during head movements with a latency of about 10 milliseconds, far too fast for cortical involvement. When the reflex gain becomes incorrect (as when someone starts wearing new glasses), the cerebellum detects the resulting retinal slip and slowly recalibrates over hours to days; once the new gain is learned, it is transferred to the vestibular nucleus for fast execution, freeing the cerebellum for new calibration tasks. The general motif is slow adaptation followed by fast deployment: invest time in learning, then execute what has been learned with minimal latency.

John Doyle's framework of layered control architectures provides mathematical grounding for why this pattern produces robustness [20]. Architectures with hard constraints at lower levels—what Doyle calls "constraints that deconstrain"—are the established path to reliable performance under uncertainty. Spinal reflexes enforce safety constraints (withdrawal from pain, postural correction) that free the cortex to plan without attending to moment-by-moment stability. Distributed control extends this principle throughout the motor system. Sensorimotor feedback loops operate at every level of the hierarchy, with local controllers handling tasks like joint compliance and reflex coordination that centralized robotic systems must compute explicitly. Higher levels coordinate and set goals rather than micromanaging execution. Recent advances in system-level synthesis have made it possible to formally analyze such distributed controllers with the sparse signals, time delays, and local feedback characteristic of biological motor control, providing quantitative tools for designing systems that are simultaneously distributed, adaptive, and provably safe.

The engineering prescription is a hybrid: lower levels governed by control theory with formal safety guarantees, higher levels governed by learned policies that adapt to new tasks and environments. Current robotic systems tend toward one extreme or the other, hand-designed hierarchies that are reliable but inflexible, or end-to-end learned controllers that are flexible but offer no safety guarantees. Biological systems combine both. Getting this combination right is among the most consequential open problems in NeuroAI.

Starting from a better place. One further principle addresses the inefficiency of current training and supports all of the above. Current practice initializes deep networks with random weights and trains to convergence: the equivalent of building an organism from a random genome. Biological development works differently: the genome specifies a compact set of rules and a neural architecture with multiple brain areas and an overall pattern of connectivity, the result of uncountable "training" trials over evolutionary history [21]. Within this framework the brain self-organizes through developmental processes that interact with early sensory experience. The resulting networks also contain useful priors, such as edge detectors in visual cortex and tonotopic maps in auditory cortex, before any task-specific learning begins. This is called inductive bias in machine learning. Replacing random initialization with structured developmental programs that generate useful starting points could bypass the need for exhaustive retraining and immense data availability, and, more importantly, make continual adaptation more natural from the outset [22].

2.3 AI is inefficient

Training a frontier language model consumes on the order of 50 GWh of electricity and costs hundreds of millions of dollars. A human brain performs vastly more diverse computation on roughly 20 W [23]. A child learns to recognize dogs from a handful of examples; LLMs require effectively the entire internet. These disparities determine where AI can be deployed, who can afford to build it, and whether the technology's growth is sustainable. They also concentrate AI capability in a small number of private organizations with access to massive compute and data.

The root cause is architectural. Modern AI runs on von Neumann hardware, in which memory and processing are physically separated. Moving data between them consumes far more energy than the arithmetic itself. The brain has no such bottleneck: synapses and neurons together store information and perform computation, eliminating the cost of data transport. The computational style also differs fundamentally. Transformers perform dense matrix multiplications across billions of parameters at every step regardless of input. Biological circuits are sparse: only about 0.1% of neurons are active at any moment, communication occurs through discrete spikes, and the precise millisecond timing of those spikes carries information that rate-based codes discard. Sparsity in the brain is a computational advantage, yielding codes that are more energy-efficient, more expressive per unit of activity, and more robust to noise. And where current AI is predominantly feedforward, the brain is massively recurrent: feedback connections from higher cortical areas to lower sensory areas are at least as numerous as feedforward connections. This feedback enables top-down prediction and iterative refinement; current AI compensates for its absence by stacking more layers spatially rather than reprocessing through fewer layers multiple times [24], wasting both memory and energy.

Neuromorphic computing implements these biological principles in silicon. Chips like Intel's Loihi 2 run spiking neural networks in which computation occurs only when a spike arrives; the hardware is otherwise quiescent. The energy savings for sparse workloads are substantial, but broader adoption requires algorithms designed natively for event-driven processing rather than dense models retroactively pruned [25, 26]. Bridging strategies are emerging: the "spike as packet" concept extends binary spikes to carry multibit payloads while preserving sparse asynchronous communication, providing a path for translating high-precision AI models onto neuromorphic substrates. Rank-order coding and oscillation-based phase coding offer additional schemes beyond simple rate codes.

Co-designing hardware and algorithms around sparsity is central to the NeuroAI efficiency agenda [27]. GPUs excel at dense synchronous computation and are poorly matched to sparse, asynchronous, recurrent workloads. Three-dimensional chip architectures that co-locate memory and processing offer a further path forward: sparse computation generates less heat per unit volume, potentially resolving the thermal problems that prevent dense 3D integration. An immediate practical opportunity also exists in commercial hardware. The neural processing units (NPUs) in modern smartphones, optimized by fierce market competition for cost and power efficiency, represent the state-of-the-art in battery-powered AI processing. Opening these NPUs to applications beyond the smartphone ecosystem, including robotics, prosthetics, and wearable health monitoring, would give NeuroAI researchers access to powerful, inexpensive platforms without waiting for custom neuromorphic chips to reach production scale [28].

3 A RESEARCH ROADMAP

The principles described above are grounded in well-characterized neuroscience, and several already have engineering implementations in early stages. What has been missing is a coordinated program connecting them. Here we outline research goals organized by the three problems identified above, at three time horizons. We note that five-year projections in fast-moving fields tend to be optimistic, while ten- to twenty-year horizons tend to be conservative. The Human Genome Project and the BRAIN Initiative both exceeded their long-term goals ahead of schedule through sustained collaboration between scientists and engineers.

Connectome-based embodied digital twins serve as a cross-cutting platform for all three problems, providing a testbed for embodied control strategies, learning architectures, and efficient neuromorphic computation simultaneously. The fruit fly connectome, roughly 140,000 neurons and 50 million synapses, has been mapped and initial simulations have produced biologically plausible motor patterns. Realizing the potential of these twins will require pairing the neural connectome with an accurate body model: a connectome alone cannot produce behavior without the biomechanical interface through which the nervous system acts on the world. Within five years, a coarse-grained neuromorphic twin of the mouse brain ($\sim$70 million neurons) is feasible, providing a platform for testing how connectomic structure supports learning and behavior. At the ten-year horizon, a primate connectome twin would bring the field closer to a brain that shares key architectural features with humans, including a six-layered neocortex, a developed prefrontal cortex, and subcortical areas, with abstracted circuit-level versions potentially usable for clinical applications in disorders like Parkinson's disease and epilepsy. At the twenty-year horizon, a functional digital twin of the human brain including the peripheral nervous system would constitute both a landmark scientific achievement and a source of architectural principles that cannot currently be derived from first principles.

Interacting with the physical world. In the near term (0–5 years), priorities include biomechanically realistic simulated embodied agents in physics engines; robots with continual fine-tuning for manipulation, terrain adaptation, and load management; event-based neuromorphic sensing for real-time low-power robotic perception, with foveated projection and event-based augmented reality as early applications; and protective sensory channels—force, pressure, temperature—that trigger withdrawal reflexes, giving robots the equivalent of a pain response that prevents damage to themselves and others. In the mid-term (5–10 years), the emphasis shifts to world models built from action-conditioned prediction and validated on physical robots rather than in simulation alone; fleet learning systems in which many robots share experience and converge on solutions faster than any individual; and a standardized layered operating system for robots, defining communication protocols between sensors, actuators, local controllers, and higher-level planners. Foundational research on the interaction of body and nervous system should proceed in parallel: groups studying neuromechanics and motor control in animals should collaborate closely with roboticists, so that robots serve as models that sharpen biological questions and biological insights feed back into peripheral design and neuroAI controllers. In the long term (10–20 years), the goal is to build robots with adaptive autonomy in unstructured human environments—systems that handle novelty, collaborate with people, and learn continuously from interaction—supported by human-level tactile and olfactory sensing with low-latency feedback and intrinsic ability to recognize and learn the ethical and safety features of their actions. In animals and humans, autonomy emerges from internal goals that shape perception and action. Toddlers learn to avoid hot stoves and sharp corners, seek food and comfort, and discover that inflicting pain on others brings consequences. Integrating analogous internal goal states into artificial controllers may be essential for exploration and learning in complex environments without exhaustive training data.

Robust learning. Near-term goals include AI architectures with multiple memory systems operating at different timescales and different information resolution, testable as modules within existing frameworks; incorporation of biological features such as neuromodulatory gating, synaptic timescale diversity, and dendritic computation into hybrid models; and community benchmarks for continual learning that test graceful degradation under distributional shift, including but going beyond resistance to catastrophic forgetting. In the mid-term, priorities are hybrid hierarchical control architectures combining formal safety guarantees at lower levels with learned flexibility at higher levels; fully probabilistic AI systems that represent and propagate uncertainty natively, essential for safety-critical domains such as autonomous driving and surgical robotics; and developmental initialization methods that replace random weights with structured starting points generated through evolutionary or developmental search over compact rule sets. Long-term goals include systems that self-organize from genomic-like programs and already contain useful priors before task-specific training begins; lifelong agents and fleets that accumulate knowledge over years without degradation; and closed-loop neuromorphic neural interfaces that sense, compute, and stimulate in real time, the ultimate test of adaptive hierarchical control.

Efficiency. Near-term priorities include community chip development projects for hardware implementations of recurrent neural networks and state-space models, architectures naturally suited to neuromorphic substrates; repurposing smartphone NPUs as platforms for edge NeuroAI applications in robotics, prosthetics, and wearable health monitoring; neuromorphic wearable devices running spike-based inference at ultra-low power; and software tools that evaluate algorithms against hardware-aware efficiency metrics rather than accuracy alone. In the mid-term, the agenda advances to heterogeneous 3D chip architectures integrating distributed sensors, memory, and compute while exploiting spatial, temporal, and precision-based sparsity; neuromorphic substrates running models translated from high-precision AI via spike-as-packet and related coding schemes; and bio-compatible neuromorphic interfaces moving from laboratory demonstration to clinical prototype. The long-term benchmark is a sub-kilowatt AI supercomputer: a system performing at the level of today's large-scale clusters while consuming less power than a household appliance. Achieving it will require every principle described in this paper—sparse computation, event-driven processing, co-located memory and compute, 3D integration, and algorithms designed for efficiency from the ground up—along with neuromorphic chips manufactured at production scale.

4 ENABLING CONDITIONS

The roadmap above describes what should be built. Whether it gets built depends on who does the work, what tools they have access to, and how the research community is organized.

Training the next generation. NeuroAI requires researchers who can move fluently between neuroscience and engineering, and there are not enough of them. Most AI researchers are trained in computer science departments with little exposure to neuroscience, cognitive science, or control theory. Most neuroscientists cannot formalize the computational principles they discover in ways that engineers can implement. The result is a translation gap: important biological insights lie dormant for years because no one in a position to implement them understands their significance, and engineers draw on neuroscience at the level of loose metaphor rather than specific mechanism.

Closing this gap requires structural changes to training at every career stage, and the common principle is embedding: placing researchers in an unfamiliar discipline long enough to acquire working competence, not just exposure. Exposure to the brain, its computations and how it may inspire Artificial Intelligence can and should start in high school with specifically designed AI-enabled devices and hands-on curricula suitable for this age-range. For graduate students, this means interdisciplinary doctoral programs designed around cross-training, including coursework in neuroscience, cognitive science, dynamical systems, and control theory that goes beyond a survey level, combined with research rotations in laboratories outside the student's home discipline. For postdoctoral researchers, it means fellowship programs that embed AI-trained scientists in neuroscience laboratories for one to three years, with a structured progression from applying AI tools to neuroscience problems (which builds familiarity with the data and the questions) toward carrying neuroscience principles back into AI architecture and algorithm design. For established researchers, it means funded exchanges of weeks to months in a collaborator's laboratory, with the explicit goal of building a shared language and identifying joint problems. At every level, the key is critical mass: isolated individuals placed in unfamiliar departments rarely thrive, but cohorts of fellows at the same institution form a community that sustains itself.

Programs like the Telluride, CapoCaccia, and Bangalore neuromorphic workshops, which for decades have brought scientists and technologists together on shared projects with support from NSF and other agencies, provide a proven model for shorter-duration immersion. Neuromatch and NeuroPAC offer complementary approaches at different scales. Expanding and sustaining programs like these across the full range of career stages is among the most cost-effective investments the field can make. And because the most important problems on this roadmap require teams, including experimentalists, theorists, engineers, and roboticists, funding mechanisms should support multi-laboratory collaborations rather than individual investigators working in isolation.

Hardware and infrastructure access. Academic researchers in neuromorphic computing face a specific bottleneck: the memory interface and fabrication infrastructure needed to test their ideas at realistic scale is proprietary, expensive, or unavailable. Open-source DRAM interface IP, developed through joint effort between funding agencies and industry, would enable academic groups to build neuromorphic chips at modest cost relative to its impact. A successor to the MOSIS program—which from the 1980s through the 2010s gave university researchers affordable access to chip fabrication and enabled an entire generation of neuromorphic engineers—would have comparable leverage if updated for modern process nodes and 3D integration. More broadly, as industry retires GPU clusters in favor of newer hardware, making that compute available to academic researchers through national partnerships would provide substantial capacity at low marginal cost for NeuroAI workloads that are smaller than frontier LLM training but larger than typical academic jobs.

Standards and shared platforms. Progress in computing has always depended on common infrastructure. The x86 and ARM instruction sets, the Linux and ROS operating systems, and the USB protocol each enabled diverse innovation by providing stable layers that researchers could build upon. NeuroAI lacks equivalents. A standardized layered architecture for robots, defining communication protocols between sensors, actuators, local controllers, and planning modules, would let researchers swap components without redesigning entire systems. Shared data repositories and standardized model formats, following the examples of FASTA in genomics and Neurodata Without Borders in neuroscience, would reduce duplication of effort that currently results from incompatible simulation and modeling frameworks.

Ethics and safety. NeuroAI systems that interact with the physical world, including robotic assistants, prosthetics, and brain-computer interfaces, introduce safety and liability considerations that go beyond those posed by purely digital AI. Systems designed to model emotional processing or social cognition raise questions about the potential for manipulation of human users. Systems that adapt their behavior through continual learning pose novel challenges for certification and regulation. Those that can physically act in the real world pose complementary challenges. These questions require attention now, while the community is small enough and the technology early enough for norms and safeguards to be established by the researchers building the systems. Integrating AI ethics into training programs and engaging proactively with policymakers are necessary and should not be deferred.

The competitive landscape. Industry AI laboratories command resources, including compute, data, and engineering talent, that academic groups cannot match. Industry's incentives, understandably, favor optimizing the current paradigm over exploring alternatives. The international dimension adds urgency: substantial government investments in AI, robotics, and neuroscience are being made worldwide, with implications for scientific leadership and national competitiveness.

Every major architectural transition in AI originated in research environments where scientists had the freedom to explore alternatives to the dominant paradigm of their day. Industry built on these ideas at scale and with resources that academia could not deploy, but the ideas themselves came from researchers pursuing fundamental questions. If the next architectural transition follows this pattern, and there is no reason to think it will not, then sustaining the academic research ecosystem that produces foundational ideas is a strategic investment. NeuroAI is where the next set of ideas is most likely to originate. Supporting it is supporting the future of the field.

5 CONCLUSION

Artificial intelligence has explored a narrow region of the space of possible architectures, one built on dense, feedforward, energy-intensive computation that separates training from deployment and ignores the physical world. Neuroscience has only recently acquired the tools to characterize biological computation at the resolution needed to extract engineering principles rather than loose analogies. The convergence of these two trajectories defines the present opportunity.

We have identified three capability gaps in current AI—inability to interact with the physical world, inadequate learning that produces brittle systems, and unsustainable inefficiency—and described the neuroscience principles that address each one: co-design of body and controller, prediction through interaction, multi-scale learning with neuromodulatory control, hierarchical distributed architectures, and sparse event-driven computation. Each corresponds to biological mechanisms whose computational logic is increasingly understood, and each points toward implementations that are feasible within the coming decade. The research roadmap we have presented translates these principles into concrete milestones, from neuromorphic digital twins and community hardware platforms in the near term to sub-kilowatt AI supercomputers and developmental self-organizing architectures in the longer term.

Realizing this program requires training researchers who can work across the boundary between neuroscience and engineering, hardware infrastructure accessible to academic groups, community standards for neuromorphic and robotic platforms, and sustained investment at a time when the dominance of a single AI paradigm and the concentration of resources in industry create strong headwinds for alternative approaches.

The industrial revolution enhanced our physical capabilities. A NeuroAI-informed transformation of artificial intelligence could substantially enhance our cognitive capabilities while deepening our understanding of our own minds. The researchers entering this field now will inherit experimental tools and computational resources that previous generations could not have imagined. The highest-leverage commitment that funders, universities, and governments can make is to invest in their training, in the infrastructure they need, and in the fundamental research that connects brains to machines. NeuroAI is the continuation of the intellectual tradition that produced modern AI. The next chapter depends on whether we choose to sustain it.

ACKNOWLEDGEMENTS

This manuscript is adapted from an NSF Workshop report entitled "NeuroAI and Beyond" held on August 27–29, 2025 in Arlington, Virginia (Organizers: Cauwenberghs, Fellous, Fermüller, Sandamirskaya, and Sejnowski). We are grateful to H. Chiel, B. Oakley, C. Bartolozzi, B. Richards, E. Fedorenko, and T. Delbruck for leading working-group discussions. We thank Drs. S.-S. Lim, E. Chua (NSF/SBE), and G. Hwang (NINDS/BRAIN) for helpful discussions.

AUTHOR CONTRIBUTIONS

J.-M.F., A.Z., and T.S. drafted and prepared the manuscript with assistance from Claude (Anthropic). All authors edited, reviewed, and approved the final text.

COMPETING INTERESTS

The authors declare no competing interests.

References

[1] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

[2] Celia Ford, Eva Childers, and Sheena M Posey Norris. Exploring the bidirectional relationship between artificial intelligence and neuroscience. *National Academies of Sciences, Engineering, and Medicine*, 2024.

[3] Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Catalyzing next-generation artificial intelligence through neuroai. *Nature communications*, 14(1):1597, 2023.

[4] Adam H. Marblestone, Greg Wayne, and Konrad P. Kording. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 10:94, 2016.

[5] Evelina Fedorenko, Anna A Ivanova, and Tamar I Regev. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312, 2024.

[6] Jean-Marc Fellous, Gert Cauwenberghs, Cornelia Fermüller, Yulia Sandamirskaya, and Terrence J. Sejnowski. NeuroAI and beyond. Preprint at https://arxiv.org/abs/2601.19955, 2026.

[7] Terrence J. Sejnowski. *ChatGPT and the Future of AI.* MIT Press, 2024.

[8] Hans P. Moravec. *Mind Children: The Future of Robot and Human Intelligence.* Harvard University Press, 1988.

[9] Hillel J. Chiel and Randall D. Beer. The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in Neurosciences*, 20(12):553–557, 1997.

[10] Gianfranco D'Angelo, Jens E. Pedersen, Tarek Hassan, et al. A benchmarking framework for embodied neuromorphic agents. *Nature Machine Intelligence*, 8:300–312, 2026.

[11] Gregory P Sutton, Nicholas S Szczecinski, Roger D Quinn, and Hillel J Chiel. Phase shift between joint rotation and actuation reflects dominant forces and predicts muscle activation patterns. *PNAS nexus*, 2(10):pgad298, 2023.

[12] Tommaso Salvatori, A. Mali, Christopher L. Buckley, Thomas Lukasiewicz, Rajesh P. N. Rao, Karl J. Friston, and Alexander Ororbia. A survey on neuromimetic deep learning via predictive coding. *Neural networks*, 195:108161, 2025.

[13] Skye Gunasekaran, Assel Kembay, Hugo Ladret, Rui-Jie Zhu, Laurent Perrinet, Omid Kavehei, and Jason Eshraghian. A predictive approach to enhance time-series forecasting. *Nature Communications*, 16:8645, 2025.

[14] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629, 2023.

[15] Michael A Arbib and Jean-Marc Fellous. Emotions: from brain to robot. *Trends in cognitive sciences*, 8(12):554–561, 2004.

[16] MacKenzie W. Mathis. Leveraging insights from neuroscience to build adaptive artificial intelligence. *Nature Neuroscience*, 29:13–24, 2025.

[17] Terrence J. Sejnowski. Dynamical mechanisms for coordinating long-term working memory based on the precision of spike-timing in cortical neurons. Preprint at https://arxiv.org/abs/2512.15891, 2025.

[18] Hee Jae Jang, Royall McMahon Ward, Carla EM Golden, and Christine M Constantinople. Acetylcholine demixes heterogeneous dopamine signals for learning and moving. *Nature Neuroscience*, 29:840–850, 2026.

[19] Samuel Schmidgall, Rojin Ziaei, Jascha Achterberg, Thomas Miconi, Louis Kirsch, S. Pardis Hajiseyedrazi, and Jason Eshraghian. Brain-inspired learning in artificial neural networks: A review. *APL Machine Learning*, 2(2):021501, 2024.

[20] Nikolai Matni, Aaron D. Ames, and John C. Doyle. A quantitative framework for layered multirate control: Toward a theory of control architecture. *IEEE Control Systems*, 44:52–94, 2024.

[21] Anthony M. Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(1):3770, 2019.

[22] Ali A. Minai. Deep intelligence: what AI should learn from nature's imagination. *Cognitive Computation*, 16:2389–2404, 2024.

[23] Hillel J Chiel, Jay S Coggan, Gourav Datta, Jean-Marc Fellous, William RP Nourse, Roger D Quinn, and Peter J Thomas. Brain-inspired energy efficient

technologies for next-generation artificial intelligence. *Biological cybernetics*, 120(2):5, 2026.

[24] Rui-Jie Zhu, Zixuan Wang, Kai Hua, et al. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025.

[25] Jason K. Eshraghian, Max Ward, Emre O. Neftci, Xinxin Wang, Gregor Lenz, Girish Dwivedi, Mohammed Bennamoun, Doo Seok Jeong, and Wei D. Lu. Training spiking neural networks using lessons from deep learning. *Proceedings of the IEEE*, 111(9):1016–1054, 2023.

[26] Timo C Wunderlich and Christian Pehle. Event-based backpropagation can compute exact gradients for spiking neural networks. *Scientific Reports*, 11(1):12829, 2021.

[27] Shuo Zhou, Chang Gao, Tobi Delbruck, and Shih-Chii Liu. Exploiting neuro-inspired dynamic sparsity for energy-efficient intelligent perception. *Nature Communications*, 16:9928, 2025.

[28] Dhireesha Kudithipudi, Catherine D. Schuman, Craig M. Vineyard, et al. Neuromorphic computing at scale. *Nature*, 637:801–812, 2025.