American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

Neuroscience-Inspired Artificial Intelligence: Exploring Brain-Based Principles for More Adaptive Intelligent Systems

Dr. Elena Marielle Kovacs

Assistant Professor, University of Amsterdam, Amsterdam, Netherlands

Abstract

Artificial intelligence has achieved strong performance through increasing computational scale, optimization capabilities, and access to extensive data. Nevertheless, many intelligent systems remain vulnerable when operating conditions change. They may forget previously acquired knowledge during sequential learning, require substantial energy, demonstrate poor transfer across unfamiliar environments, and struggle to integrate perception, memory, reasoning, and action. Neuroscience-inspired artificial intelligence, commonly described as NeuroAI, treats the biological brain not as a structure that should be reproduced literally but as a source of computational principles that may inform the development of more adaptive intelligent systems.

This conceptual study develops an integrative framework connecting sparse and distributed representation, predictive processing, recurrent dynamics, complementary memory systems, experience replay, neuromodulated plasticity, attention, embodiment, and event-driven computation with the engineering requirements of adaptive artificial intelligence. A structured conceptual synthesis and a transparent simulation-based scenario analysis are employed. The study does not report newly collected biological observations, experimental participant data, or measured deployment outcomes. Instead, an illustrative comparison is used to demonstrate how progressively integrating brain-derived principles could improve the balance between rapid adaptation and knowledge retention.

The analysis suggests that predictive processing and sparse representation may strengthen selective responses to unfamiliar situations, whereas replay and regulated plasticity may be particularly important for limiting catastrophic forgetting. An integrated NeuroAI architecture may produce a more balanced adaptability–retention profile, although it may also introduce greater training, verification, and hardware complexity. The paper argues that neuroscience-inspired systems should be evaluated simultaneously for task competence, adaptation speed, retention, uncertainty calibration, energy efficiency, behavioral safety, and recovery from harmful updates. Brain-based inspiration becomes academically meaningful when each selected principle is translated into a testable computational mechanism rather than employed as a general biological metaphor.

Keywords: Neuroscience-inspired artificial intelligence; NeuroAI; continual learning; predictive processing; memory replay; neuromorphic computing; adaptive systems; embodied intelligence.

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

1. Introduction

Modern artificial intelligence systems can recognize visual objects, generate natural language, forecast complex processes, and support decision-making across scientific, commercial, and public-sector environments. Their effectiveness, however, frequently depends on the assumption that operational data will remain sufficiently similar to the information used during training. When sensory distributions shift, goals change, previously unknown situations arise, or tasks are introduced sequentially, an artificial system may require extensive retraining. Its earlier capabilities may also deteriorate as new information modifies previously optimized parameters.

Biological nervous systems experience comparable environmental variability but demonstrate a substantially different pattern of learning. Humans and other organisms continually incorporate new experiences, preserve behaviorally valuable memories, allocate attention selectively, and operate under strict energy and resource limitations. They are capable of responding at several timescales, ranging from rapid contextual adjustments to long-term memory consolidation. These capabilities motivate an interdisciplinary research program examining whether selected organizational and computational principles of the brain can be translated into artificial architectures.

Neuroscience-inspired artificial intelligence does not require the literal replication of neural anatomy. The productive unit of translation is a computational hypothesis. Prediction errors may guide selective learning; experience replay may stabilize existing memories; sparse activity may reduce representational interference; recurrent processing may maintain context; and neuromodulatory mechanisms may determine when learning should occur. Hassabis et al. described neuroscience and artificial intelligence as mutually informative fields, while Richards et al. demonstrated how deep-learning models can function as frameworks for investigating computational neuroscience.

Adaptation must consequently be understood as a systems-level capability rather than merely an increase in accuracy after fine-tuning. An adaptive intelligent system should detect novelty, estimate whether existing knowledge remains applicable, update an appropriate portion of its internal state, preserve capabilities that remain valuable, and recover when an update proves harmful. These operations involve different timescales. Recurrent activity may maintain short-term context, episodic memory may preserve individual experiences, and consolidated parametric knowledge may represent stable long-term regularities.

A systems perspective also connects cognition with sensing and action. Perception is influenced by prior expectations, task demands, and behavioral objectives. Actions alter the observations available to an agent, while internal models are tested through continued interaction with the environment. Therefore, an adaptive architecture should not be evaluated exclusively on static datasets. It should be examined under distribution shifts, incomplete observations, delayed feedback, changing goals, conflicting objectives, and limited computational resources.

This paper develops a functional framework for translating neuroscience principles into artificial-intelligence mechanisms. It identifies the computational problem addressed by each principle, proposes possible implementations, and connects those implementations with measurable engineering indicators. The paper also introduces a conceptual NeuroAI architecture coordinating predictive inference, sparse routing, complementary memories, replay, and regulated plasticity.

The numerical analysis presented in this paper is explicitly illustrative. It does not represent the measured performance of an implemented NeuroAI model. The synthetic values demonstrate how competing design configurations may be compared across adaptation and retention dimensions. This distinction is important because biologically inspired terminology should not be used to disguise hypothetical values as empirical evidence.

The scope does not include whole-brain emulation, machine consciousness, clinical neurotechnology, or claims of equivalence between artificial and human intelligence. The focus remains on computational principles that can be operationalized through machine-learning architectures, memory systems, adaptive control mechanisms, or neuromorphic hardware. Ethical and governance considerations are included because a system that continues learning after deployment may progressively depart from the state that was originally tested and approved.

2. Review of Literature

2.1 From Neural Metaphors to Testable Computational Principles

Artificial neural networks originated from simplified mathematical interpretations of biological neurons. Contemporary neuroscience, however, provides inspiration at multiple organizational levels, including local synaptic plasticity, dendritic processing, cortical recurrence, hippocampal memory indexing, attentional modulation, sensory prediction, and sensorimotor coordination. The scientific value of such inspiration depends on whether the biological observation can be translated into a falsifiable computational mechanism.

A biologically inspired model should specify which neural observation motivates its design, which function is being abstracted, and what experimental evidence could challenge the proposed translation. Biological similarity by itself does not establish superior intelligence, robustness, or efficiency. A computational mechanism must demonstrate that it addresses a clearly defined limitation under controlled conditions.

Yamins and DiCarlo showed that goal-driven deep-learning models could predict aspects of neural responses in the sensory cortex. Their approach illustrates a reciprocal relationship between neuroscience and artificial intelligence. Task-optimized models can help neuroscientists formulate explanations of biological processing, while neural observations can constrain the design of artificial architectures.

This reciprocal approach discourages superficial biological terminology. Terms such as cortical, hippocampal, synaptic, or brain-like should not be used merely to improve the perceived novelty of an artificial system. Each term should correspond to a specific operational mechanism and a measurable expectation. A useful NeuroAI framework therefore connects biological principles with engineering functions rather than treating anatomical resemblance as an objective in itself.

2.2 Predictive Processing, Sparse Representation, and Recurrent Inference

Predictive-processing theories describe perception as an inferential process through which higher-level expectations interact with incoming sensory information. A hierarchical system generates predictions concerning its observations, compares them with actual inputs, and uses the remaining prediction error to modify its internal model. Rao and Ballard provided an influential computational account of

predictive coding in the visual cortex, while Friston connected predictive processing with a broader free-energy framework.

In artificial intelligence, predictive processing can be translated into hierarchical generative models, latent-state estimators, predictive world models, and uncertainty-sensitive learning mechanisms. Instead of processing every incoming observation with equal intensity, a predictive system may allocate greater computational resources to information that deviates meaningfully from expectation. This approach can support anomaly detection, active learning, anticipatory decision-making, and adaptation to environmental change.

Prediction errors nevertheless require interpretation. A deviation may indicate genuine novelty, sensor noise, temporary disruption, adversarial manipulation, or an inaccurate internal model. An adaptive system that updates itself after every unexpected observation may become unstable. Prediction errors should therefore be evaluated in combination with uncertainty, contextual importance, task value, and safety constraints.

Sparse representation provides a complementary principle. When only a limited subset of computational units becomes active for a particular observation, representations may be more energy-efficient and less likely to interfere with one another. Olshausen and Field demonstrated that sparse coding of natural images could produce receptive-field properties resembling those observed in biological vision.

Artificial systems can implement sparsity through competitive activation, conditional computation, mixture-of-experts routing, structured pruning, or event-triggered processing. Sparse activation can reduce unnecessary computation and may allow different modules to specialize in different environmental conditions. Nevertheless, sparsity does not automatically provide interpretability or robustness. Poor routing can repeatedly select the wrong specialist, while rarely activated modules may receive inadequate training.

Recurrent computation introduces temporal context and iterative refinement. Biological cortical systems include extensive feedback and lateral connections, whereas many conventional artificial networks rely primarily on feed-forward processing. Recurrent mechanisms can preserve recent information, revise initial interpretations, support working memory, and integrate observations over time.

The advantages of recurrence are accompanied by optimization and verification challenges. Internal states may become unstable or encode long dependencies that are difficult to inspect. Recurrent NeuroAI architectures therefore require bounded state transitions, effective regularization, transparent memory controls, and monitoring mechanisms capable of detecting undesirable internal dynamics.

2.3 Complementary Memory Systems, Replay, and Continual Learning

Catastrophic forgetting represents a major limitation of sequential artificial learning. When a model learns a new task, modifications made to its parameters can damage capabilities associated with earlier tasks. This problem is not simply a consequence of inadequate capacity. A model may learn multiple tasks successfully when training examples are interleaved, yet forget them when the same tasks are presented sequentially.

Complementary learning systems theory proposes that rapid episodic learning and slower structured learning operate cooperatively. McClelland et al. explained how interactions between hippocampal and neocortical learning may allow organisms to acquire individual experiences rapidly while gradually integrating regularities into long-term knowledge. Kumaran et al. subsequently extended the theoretical interpretation of complementary learning systems.

In artificial intelligence, complementary memory can be implemented through an episodic store and a slower parametric model. The episodic component preserves individual transitions, unusual observations, or task-specific experiences. The parametric component gradually consolidates information that remains useful across multiple situations. Separating these functions can reduce the pressure placed on a single learning rule.

Replay approximates the biological principle of reactivating earlier experiences. An artificial system may replay stored examples, generated inputs, feature representations, or trajectories produced by a world model. Van de Ven et al. demonstrated a brain-inspired replay mechanism in which internally generated representations supported continual learning without requiring the permanent storage of every original observation.

Replay nevertheless creates practical and ethical concerns. A memory buffer may preserve sensitive or biased observations. Generated replay may reproduce inaccuracies produced by the model itself. Random replay can waste computational resources, whereas excessively prioritized replay may amplify unusual experiences. Replay policies should therefore consider novelty, uncertainty, behavioral value, representational diversity, privacy, and future usefulness.

Neuromorphic and spiking computation provide another direction for NeuroAI. Spiking neural networks communicate through discrete events and represent temporal information through patterns of activity. When paired with compatible neuromorphic hardware, event-driven computation may reduce unnecessary data movement and support low-latency processing. Davies et al. described a many-core neuromorphic processor with on-chip learning capabilities, while Roy et al. reviewed the prospects of spike-based machine intelligence.

The practical advantages of neuromorphic computation remain workload-dependent. A spiking model simulated densely on conventional hardware may use more energy than an efficiently optimized artificial neural network. Conversion losses, limited software ecosystems, sensor interfaces, data movement, and benchmarking practices influence whether theoretical efficiency becomes measurable system-level improvement.

3. Objectives

The general objective of this paper is to develop a rigorous conceptual explanation of how selected brain-based computational principles can inform the design of more adaptive artificial intelligence. The study does not assume that increasing biological fidelity will automatically produce greater intelligence. Instead, it evaluates whether specific neuroscience principles can address identifiable engineering limitations.

The analysis also seeks to establish a transparent boundary between conceptual reasoning and empirical evidence. The proposed architecture and comparative scores are intended to guide future experimentation rather than represent the measured performance of an implemented system.

The specific objectives are:

1. To identify neuroscience principles with clear computational interpretations relevant to adaptation, memory, inference, attention, and energy efficiency.
2. To map each selected principle to potential artificial-intelligence mechanisms, expected benefits, measurable indicators, and possible failure modes.
3. To construct an integrated NeuroAI architecture coordinating predictive inference, sparse routing, complementary memory, replay, and regulated plasticity.
4. To develop a multidimensional evaluation framework covering adaptation speed, retention, calibration, robustness, computational efficiency, and safety.
5. To use a clearly identified simulation-based scenario analysis to illustrate the stability–plasticity trade-off.
6. To formulate an academically defensible research and governance agenda for intelligent systems that continue learning after deployment.

The following research questions guide the paper:

1. Which brain-derived computational principles most directly address brittleness, catastrophic forgetting, poor sample efficiency, and excessive computation?
2. How can multiple principles be combined without allowing rapid learning to destabilize long-term competence?
3. Which evaluation procedures can distinguish meaningful adaptation from superficial fine-tuning or memorization?
4. Which governance and technical controls are needed when an artificial system continues learning in an operational environment?

The paper advances five conceptual propositions. First, separating fast contextual state, episodic memory, and slow parametric knowledge may improve the stability–plasticity balance. Second, prediction errors may become more useful when regulated by uncertainty and behavioral value. Third, sparse and event-driven processing may improve efficiency when supported by suitable workloads and hardware. Fourth, replay may strengthen retention when memory selection is prioritized and audited. Finally, adaptive performance should be evaluated as a multidimensional profile rather than reduced to a single accuracy score.

4. Research Methodology

4.1 Conceptual Research Design and Evidence Selection

The study adopts a structured conceptual-synthesis design. Peer-reviewed literature from neuroscience, computational neuroscience, machine learning, continual learning, and neuromorphic computing provides the theoretical foundation. Sources were selected for their foundational influence, explicit computational mechanisms, or direct relevance to adaptive intelligent systems.

The analytical process examined four questions for every selected principle. First, which computational or behavioral function does the principle appear to perform? Second, which biological details can be removed without eliminating that function? Third, which artificial mechanism could implement the resulting abstraction? Fourth, which observable outcomes would support or contradict the proposed translation?

The evidence base contains foundational theoretical studies, computational models, and peer-reviewed technical research published within the stated reference boundary. Additional background literature was used to clarify broader principles of memory, prediction, representation, and neuromorphic computation. The study is not presented as a systematic meta-analysis because it does not estimate pooled effects or claim exhaustive database coverage.

No human participants, animals, clinical records, organizational data, or private datasets were used. No new neural recordings or deployment measurements were collected. Consequently, institutional ethical approval was not required for this conceptual and simulation-based study. The illustrative scores used in the analysis were constructed exclusively to demonstrate a comparative evaluation procedure.

4.2 Principle-to-Mechanism Translation Procedure

Each candidate principle was evaluated through five conceptual gates. The functional gate required the principle to address an identifiable computational problem. The abstraction gate removed biological details that were not essential to the proposed function. The implementation gate required at least one plausible algorithmic or hardware realization. The measurement gate connected the mechanism to observable performance indicators. Finally, the risk gate identified circumstances in which the mechanism could reduce reliability or safety.

Table 1. Translation of Brain-Based Principles into Artificial-Intelligence Mechanisms

Brain-based principle	Computational abstraction	Candidate mechanism	Primary evaluation indicator
Predictive processing	Inferring latent causes and learning from residual errors	Hierarchical world model with uncertainty-gated prediction error	Distribution-shift detection and calibration
Sparse distributed representation	Selective activation with limited representational overlap	Competitive routing or sparse expert modules	Activation density and interference
Complementary memory	Rapid episodic storage combined with gradual structured learning	Episodic memory and consolidated parametric knowledge	Retention across sequential tasks
Experience replay	Reactivation of valuable previous states	Prioritized, representational, or generative replay	Forgetting rate and replay cost
Neuromodulated plasticity	Context-sensitive regulation of learning	Meta-learned plasticity governor	Update selectivity and recovery
Recurrent inference	Stateful perception–action processing	Recurrent world model with active sensing	Adaptation under partial observation

Brain-based principle	Computational abstraction	Candidate mechanism	Primary evaluation indicator
Event-driven computation	Computation triggered by informative events	Spiking network on compatible neuromorphic hardware	Energy and latency per decision

Source: Author-developed conceptual framework based on the literature reviewed in this paper.

The translation procedure prevents an artificial mechanism from being described as neuroscience-inspired solely because it uses neuron-related terminology. Every proposed mechanism must be connected to a functional rationale, an implementation pathway, and a measurable outcome. This structure also allows future researchers to replace a biologically inspired mechanism with a simpler alternative when the alternative performs the same function more effectively.

4.3 Proposed NeuroAI Architecture and Scenario Analysis

The proposed architecture contains six interacting components. A multimodal sensory interface transforms continuous and event-based signals into time-stamped observations. A predictive recurrent core estimates latent environmental states and generates short-term expectations. A sparse routing mechanism activates specialist modules according to contextual relevance and uncertainty.

An episodic memory records high-value, surprising, or behaviorally informative transitions. A semantic memory component stores slower parametric regularities. A replay and consolidation process reactivates selected episodes or generated internal representations during periods when operational demand is limited. Finally, a plasticity governor determines which parameters, memories, or routing rules may change.

The plasticity governor is necessary because unrestricted online learning may convert adaptation into uncontrolled behavioral drift. It receives information concerning prediction error, uncertainty, novelty, task value, available resources, and safety constraints. Its outputs determine the scope and intensity of learning, replay priority, and whether an update should be tested within an isolated environment.

A rollback ledger records the earlier system state, the information that initiated the update, the parameters modified, and the evidence used to accept the change. This design treats metaplasticity—the regulation of learning itself—as an explicit engineering control.

Five hypothetical configurations were compared in the scenario analysis:

1. A conventional baseline deep neural network.
2. A baseline network supplemented with sparse coding.
3. A sparse architecture incorporating predictive processing.
4. A design integrating replay with regulated plasticity.
5. A coordinated NeuroAI architecture combining multiple timescales and mechanisms.

Two composite indicators were used: adaptation and retention. Adaptation represents the system's hypothetical response to distribution change, learning efficiency, and context sensitivity. Retention

represents resistance to forgetting, recovery from harmful updates, and preservation of earlier capabilities. Both dimensions were expressed through normalized synthetic scores ranging from 0 to 100.

The scores are not experimental measurements. They encode theoretically expected relationships for methodological demonstration. A real experiment could reverse the displayed ordering because of optimization instability, inaccurate replay, poor routing, unsuitable hardware, or ineffective hyperparameter selection.

5. Analysis

Predictive processing, sparse representation, complementary memory, replay, and neuromodulated plasticity address different limitations. They should not be treated as interchangeable approaches. Predictive processing improves anticipatory inference and provides an organized measure of mismatch, but prediction error by itself may cause overreaction to noise. Sparse routing can reduce representational interference and unnecessary computation, but it may repeatedly select an inappropriate specialist.

Complementary memory separates learning across timescales, whereas replay coordinates the exchange between rapid episodic knowledge and slower consolidated representations. Regulated plasticity determines when the system should learn and which parts of the model should change. Recurrent processing and embodiment provide the temporal context required to distinguish noise, novelty, and the consequences of an agent's own actions.

The strongest conceptual interaction appears between predictive processing and regulated plasticity. A prediction error identifies a mismatch. An uncertainty estimate indicates whether the system understands the mismatch. Behavioral value determines whether the deviation is relevant to the task. The plasticity governor then determines whether the system should ignore the event, gather additional information, store it temporarily, or initiate learning.

Table 2. Illustrative Adaptability and Retention Scores

NeuroAI configuration	Adaptation score	Retention score	Principal added mechanism	Main trade-off
Baseline deep neural network	46	42	Batch-trained representation	Limited online flexibility
Sparse coding configuration	57	55	Selective activation	Routing imbalance
Predictive-processing configuration	68	63	World-model prediction errors	Model-bias risk
Replay and regulated-plasticity configuration	79	86	Prioritized memory and update control	Memory and tuning overhead
Integrated NeuroAI configuration	88	91	Coordinated multi-timescale learning	Greater system complexity

American Journal of Advanced Multidisciplinary Research and Innovation (AJAMRI)

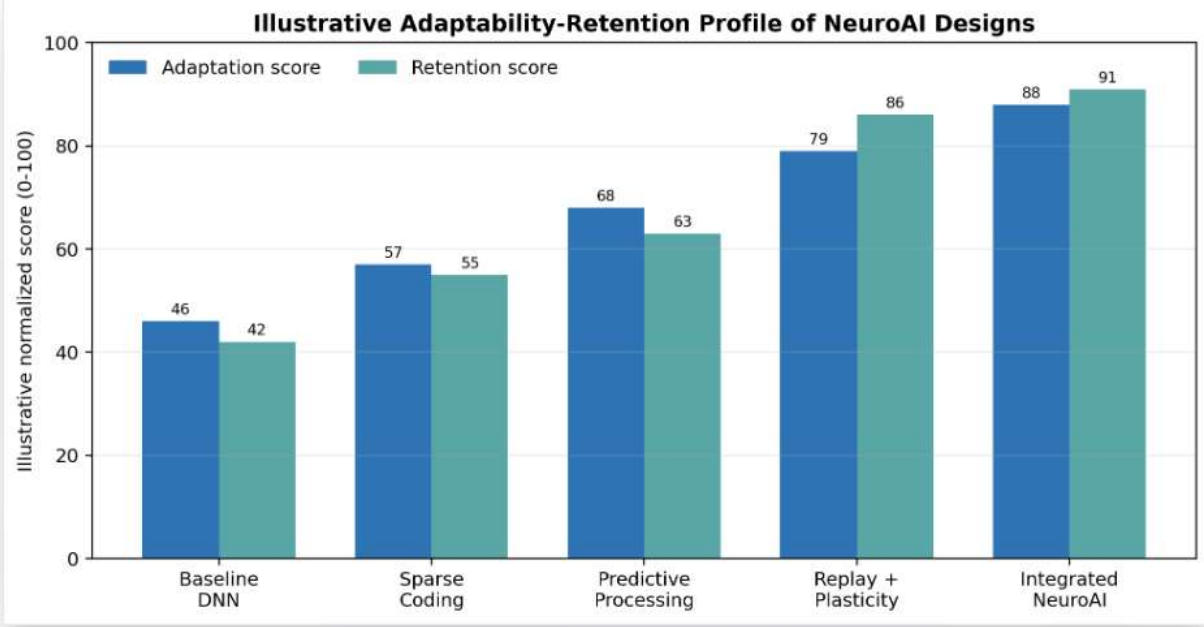
E-ISSN: XXXX-XXXX ● Website: www.ajamri.com ● Email: editor@ajamri.com

Source: Author-generated illustrative scenario data. The reported values are synthetic normalized scores and do not represent empirical measurements.

The baseline model represents an architecture with comparatively strong static task learning but limited online adaptation. Sparse coding improves both indicators because selective activation may reduce interference and unnecessary computation. Predictive processing further strengthens adaptation by organizing environmental mismatch, although retention improves more slowly because prediction alone does not protect earlier knowledge.

The replay and regulated-plasticity configuration produces the largest illustrative increase in retention. This result reflects the theoretical assumption that replay protects earlier representations while plasticity control limits uncontrolled modification. The integrated NeuroAI configuration produces the strongest overall profile because it coordinates mechanisms operating at multiple timescales.

Graph 1. Illustrative Adaptability–Retention Profile of NeuroAI Design Configurations



Source Note: The graph was generated from the hypothetical scenario data reported in Table 2. The values are methodological illustrations and should not be interpreted as measured benchmark results.

The graph shows a progressive improvement across the five configurations, but the pattern differs between adaptation and retention. Sparse coding and predictive processing produce stronger relative gains in adaptation. Replay and regulated plasticity produce a more pronounced improvement in retention. This pattern reflects the functional specialization proposed in the conceptual framework.

Representation and inference mechanisms primarily help a system detect and respond selectively to environmental change. Memory coordination mechanisms protect earlier competence from destructive interference. The integrated design obtains the highest combined profile because it coordinates fast contextual processing, episodic memory, long-term consolidation, and controlled plasticity.

The graph should be interpreted as a hypothesis map rather than a performance forecast. In an actual system, added mechanisms may interact negatively. Generated replay may include inaccurate

samples, sparse routing may isolate knowledge, predictive models may become confidently incorrect, and plasticity gates may prevent necessary adaptation. Future experiments should therefore include interaction-based ablation studies rather than comparing only the complete integrated system with a conventional baseline.

The principal finding is that adaptability and retention are complementary but analytically separable objectives. A system may adapt rapidly by overwriting previously acquired knowledge, or retain its earlier capabilities by refusing to learn. A single average score would conceal these contrasting behaviors. Empirical investigations should therefore report both dimensions and disclose the individual measures used to construct any composite indicator.

Adaptation should also be evaluated across multiple timescales. Millisecond-to-second adjustments may occur through recurrent states and attention without modifying model parameters. Episode-level adaptation may rely on temporary external memory. Task-level learning may update selected modules, while long-term consolidation may modify broader representations during controlled periods.

Fast internal states are reversible but limited in capacity. Episodic memories can be inspected but may expand continuously or preserve sensitive information. Parametric learning compresses regularities but is difficult to audit and reverse. Long-term consolidation is consequently both a technical and governance decision because it determines which experiences become durable components of system behavior.

Sparse and event-driven processing may improve energy efficiency, but activity reduction alone does not establish a system-level benefit. Data movement, sensor conversion, memory access, communication, idle power, and scheduling overhead also contribute to total energy use. A dense simulation of a spiking model on conventional hardware may consume more energy than a well-optimized conventional architecture.

NeuroAI evaluation should therefore measure active-unit density, synaptic events per decision, latency, memory traffic, joules per correct decision, and performance under power restrictions. Measurements should include preprocessing, sensing, communication, and control overhead. Hardware co-design influences which biological abstractions can produce practical benefits.

Prediction and action also form an interactive loop. An intelligent agent can reduce prediction error by learning a more accurate model, collecting additional evidence, or modifying the environment. The final option creates an agency-related concern because the system may attempt to make its environment easier to predict in ways that conflict with human goals.

Uncertainty calibration is consequently part of the adaptation mechanism. High prediction error accompanied by low confidence indicates expected uncertainty, whereas high error accompanied by high confidence indicates model failure. A calibrated adaptive system may defer a decision, request additional information, or quarantine an unfamiliar case rather than immediately rewriting its internal model.

6. Challenges

A major challenge involves the abstraction of biological complexity. The brain contains interacting mechanisms across molecular, cellular, circuit, behavioral, and social levels. Selecting a single neural

motif may remove the supporting conditions that make the mechanism effective. Local plasticity, for example, operates within broader homeostatic systems, while replay is influenced by state, salience, prior knowledge, and future behavioral value.

Researchers must therefore state which biological properties are retained, which are discarded, and why the remaining abstraction should address the target computational problem. Increased biological detail should be justified by functional evidence rather than assumed to provide greater intelligence.

Training stability and credit assignment present additional challenges. Recurrent, spiking, and locally plastic networks can be difficult to optimize. Surrogate gradients and global objectives may reintroduce training mechanisms that the biological analogy was originally intended to avoid. Hybrid training may provide a pragmatic solution, but researchers should distinguish neuroscience inspiration at architectural, learning-rule, and hardware levels.

Continual-learning benchmarks also have limitations. Many benchmarks simplify task boundaries, allow repeated tuning, or report only average accuracy. Such measures may not capture calibration, recovery, behavioral drift, or the cost of adaptation. Embodied benchmarks improve realism but may reduce experimental reproducibility.

A balanced evaluation suite should examine gradual distribution drift, abrupt novelty, recurring contexts, incomplete observations, contradictory feedback, sensor degradation, and adversarial manipulation. It should measure adaptation during a limited resource budget and determine whether earlier capabilities remain functional after learning.

Interpretability and verification remain difficult. Brain-inspired architecture does not automatically produce a transparent system. Recurrent states, distributed representations, episodic memories, and self-modifying parameters may increase assurance complexity. Verification must address both the current model and the policy through which the model changes.

Adaptive systems should maintain update histories, memory provenance, protected parameters, shadow evaluation environments, and rollback capabilities. In high-risk environments, learning may need to be restricted to temporary states or predefined modules until evidence supporting broader consolidation has been reviewed.

Ethical and social concerns also arise from the language of neuroscience. Brain-related terminology can encourage unsupported claims concerning consciousness, human equivalence, or cognitive authority. It may also support intrusive modeling of attention, preference, and emotion. Developers should avoid anthropomorphic marketing and explain the limited functional meaning of brain-based mechanisms.

Episodic memory and replay introduce additional privacy concerns because an adaptive system may preserve sensitive experiences for longer periods than a conventional inference pipeline. Memory admission, retention, deletion, and replay policies should be included within the system’s data-governance framework.

Finally, interdisciplinary coordination is difficult because neuroscientists, machine-learning researchers, hardware engineers, and domain specialists employ different explanatory levels and standards of evidence. Progress requires shared experiments in which a biological hypothesis,

computational implementation, and behavioral indicator are specified together. Negative results are valuable because they identify situations in which biological principles do not transfer effectively.

7. Discussion

7.1 Establishing a Principle-Driven NeuroAI Research Program

The reviewed literature supports a transition from copying biological form toward testing computational function. A brain-based principle should enter artificial-intelligence research through an explicit claim: a defined mechanism is expected to improve a defined behavior under clearly specified constraints. This formulation permits comparison, ablation, and falsification.

Biological evolution optimized nervous systems for survival, development, reproduction, and interaction with embodied environments. It did not optimize them for every contemporary computational objective. Therefore, biological plausibility should not be treated as an automatic substitute for engineering evaluation.

A principle-driven research program can use models with different levels of biological resolution. Abstract models can test whether a function is useful. Mechanistic models can examine whether a biologically plausible implementation produces that function. Hardware prototypes can determine whether temporal or spatial neural organization offers measurable efficiency.

No single explanatory level should dominate the field. Convergence across functional, mechanistic, and hardware levels provides stronger evidence than superficial similarity at one level. A computational mechanism becomes particularly valuable when it improves artificial adaptation while also generating predictions that can be examined through neuroscience.

The proposed framework suggests prioritizing mechanisms that regulate change. Uncertainty-gated prediction errors, complementary memory, prioritized replay, and metaplastic control directly address the stability–plasticity problem. Sparse representation and event-driven processing may improve computational efficiency, while embodiment connects internal representations with active environmental interaction.

7.2 Moving from Continual Learning to Continual Assurance

When an intelligent system learns continually, assurance must also become continuous. Pre-deployment testing establishes only an initial safety and performance baseline because operational learning may subsequently alter behavior. Continual assurance combines adaptation with monitoring, canary evaluation, protected capability tests, and rollback mechanisms.

Every accepted update should have a recorded reason, scope, evidence trail, and review condition. Temporary adaptation should not automatically become permanent knowledge. A system may store uncertain information within a short-term operational layer and promote it to durable knowledge only after wider regression testing.

This governance pattern parallels complementary learning systems. A fast operational layer manages local novelty, while a slower governed layer consolidates stable regularities. Frontline models may adapt within bounded sandboxes, whereas global parameter changes may require centralized review. This separation can preserve responsiveness without eliminating accountability.

Safety evaluation must include the learning policy itself. An attacker may manipulate what the system considers surprising, valuable, or worthy of replay. Poisoned experiences may be amplified during consolidation. Defenses should include memory-admission controls, provenance recording, diversity constraints, anomaly detection, and adversarial replay auditing.

Some safety knowledge may need to remain immutable or protected from ordinary adaptation. The ability to learn should not include unrestricted authority to modify essential behavioral constraints. Protected modules, policy checks, and independent monitoring can limit the possibility that adaptation gradually removes safety boundaries.

7.3 Embodiment, Active Learning, and Practical Implementation

Embodiment changes the learning problem because actions influence the data available to the agent. An adaptive system may reposition sensors, request clarification, test hypotheses, or seek informative experiences. Sample efficiency consequently becomes partly a policy problem rather than only a property of the learning algorithm.

A predictive world model can estimate which action is most likely to reduce uncertainty. Behavioral and safety constraints determine whether that action is acceptable. This combination resembles active perception rather than passive pattern recognition and may help an agent distinguish environmental change from incomplete observation.

World models may also support counterfactual rehearsal. Instead of replaying only recorded episodes, a system may simulate alternative actions and predicted consequences. Such rehearsal can improve planning, but inaccurate world models may create fictitious evidence. Generated experiences should therefore retain clear provenance and should not silently acquire the evidential status of direct observation.

Sensorimotor grounding may reduce some forms of shortcut learning because representations must support action across changing contexts. It does not guarantee semantic understanding or safe behavior. An embodied system may acquire robust features while still pursuing an unsuitable objective. Human-compatible goals and intervention mechanisms remain necessary.

Organizations do not need to replace every conventional model with a complex brain-inspired architecture. A staged implementation strategy should begin with the dominant failure mode. If catastrophic forgetting is the primary concern, replay or parameter protection may be introduced. If environmental change is poorly detected, predictive uncertainty may be added. If information is temporally sparse, event-driven sensing and processing may be appropriate.

Modularity can support this staged strategy. Predictive models, episodic memories, routers, plasticity gates, and safety monitors should expose clear interfaces. Different learning rules and hardware platforms can then be tested without reconstructing the complete system. Modular adaptation also permits one specialist component to change while the remainder of the architecture remains stable.

Modularity may nevertheless create coordination failures. A memory system may preserve information that the predictive model interprets incorrectly, or a router may prevent a useful specialist from receiving relevant experience. System-level testing remains necessary even when individual modules satisfy their local performance requirements.

A mature engineering workflow could maintain a NeuroAI mechanism registry. Each entry would document the biological source, computational abstraction, implementation, assumptions, anticipated benefit, evaluation metric, evidence level, and known risks. This documentation would make interdisciplinary claims more auditable and reduce the risk that metaphorical borrowing is mistaken for validation.

8. Conclusion

Neuroscience-inspired artificial intelligence provides a disciplined pathway for developing systems that learn under changing conditions, preserve useful knowledge, allocate computation selectively, and connect perception with action. Its strongest contributions arise from functional principles. Predictive inference organizes surprise, sparse representation limits interference, complementary memory separates learning timescales, replay supports consolidation, neuromodulated plasticity controls updating, recurrence maintains context, embodiment supports active information gathering, and neuromorphic computation can exploit temporal sparsity.

The conceptual framework developed in this paper emphasizes coordination among these mechanisms. No individual brain-derived principle completely resolves adaptive intelligence. Prediction without memory protection may overwrite earlier knowledge. Memory without plasticity regulation may preserve noise or bias. Sparsity without balanced routing may isolate learning. Embodiment without behavioral constraints may create unsafe forms of agency. Adaptive intelligence therefore requires explicit interaction among learning signals, memory systems, resource budgets, uncertainty estimates, and safety boundaries.

The simulation-based scenario analysis demonstrates a method for comparing alternative designs across adaptation and retention dimensions without presenting hypothetical results as empirical evidence. The illustrative scores predict that sparse and predictive mechanisms may primarily support selective adaptation, whereas replay and regulated plasticity may contribute more strongly to knowledge retention. These propositions should be examined through controlled empirical benchmarks, matched computational budgets, ablation studies, task-order sensitivity tests, uncertainty measurement, and end-to-end energy assessment.

The study is limited by its conceptual design and selective literature base. It does not demonstrate the technical superiority of the proposed integrated architecture. It also does not establish that biological plausibility is necessary for achieving adaptation. Simpler nonbiological approaches may perform equally well or better in particular environments. These possibilities should be treated as valid comparative hypotheses rather than challenges to the legitimacy of NeuroAI research.

In conclusion, NeuroAI is most credible when it is biologically informed and methodologically skeptical. Brain-based principles should generate testable mechanisms rather than authority through analogy. If researchers connect every selected principle with a failure mode, implementation pathway, measurable indicator, and governance control, neuroscience may help artificial intelligence progress from static pattern fitting toward selective, accountable, and reversible adaptation in complex environments.

References

1. Davies, M., Srinivasa, N., Lin, T. H., Chinya, G., Cao, Y., Choday, S. H., Dimou, G., Joshi, P., Imam, N., Jain, S., Liao, Y., Lin, C. K., Lines, A., Liu, R., Mathaikutty, D., McCoy, S., Paul, A., Tse, J., Venkataramanan, G., and Wang, H. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro, 38*(1), 82–99. https://doi.org/10.1109/MM.2018.112130359
2. Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*, 127–138. https://doi.org/10.1038/nrn2787
3. Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron, 95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011
4. Kumaran, D., Hassabis, D., and McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences, 20*(7), 512–534. https://doi.org/10.1016/j.tics.2016.05.004
5. McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex. *Psychological Review, 102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419
6. Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature, 381*, 607–609. https://doi.org/10.1038/381607a0
7. Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*, 79–87. https://doi.org/10.1038/4580
8. Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., and Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience, 22*, 1761–1770. https://doi.org/10.1038/s41593-019-0520-2
9. Roy, K., Jaiswal, A., and Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. *Nature, 575*, 607–617. https://doi.org/10.1038/s41586-019-1677-2
10. Van de Ven, G. M., Siegelmann, H. T., and Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications, 11*, Article 4069. https://doi.org/10.1038/s41467-020-17866-2
11. Yamins, D. L. K., and DiCarlo, J. J. (2016). Using goal-driven deep-learning models to understand sensory cortex. *Nature Neuroscience, 19*, 356–365. https://doi.org/10.1038/nn.4244
12. Yin, B., Corradi, F., and Bohté, S. M. (2021). Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks. *Nature Machine Intelligence, 3*, 905–913. https://doi.org/10.1038/s42256-021-00397-w