Learning Through Narratives: Episodic Compression to Latent Variables for Sample-Efficient Intelligence

Terminology Note: Throughout this paper, we use terms like "learning," "understanding," and "reasoning" as behavioral shorthand without making claims about machine consciousness or internal mental states. When we say an AI system "learns" from narratives, we mean it produces behavioral responses consistent with narrative patterns. This usage aligns with our focus on behavioral competence rather than unverifiable internal states. Similarly, we use "latent variables" in the standard ML sense of learned hidden representations that capture underlying structure, distinct from LeCun's recent emphasis on continuous latent variables for handling aleatory uncertainty in predictions.

Abstract

Current machine learning systems require orders of magnitude more training examples than humans to achieve comparable competence. LeCun (2022) argues that embodied physical interaction is necessary for sample-efficient learning, proposing Joint Embedding Predictive Architectures (JEPA) as a solution. We present an alternative approach based on the premise that humans appear to learn extensively from narratives (indirect observation) rather than direct experience. Therapists understand conditions they haven't experienced, children learn dangers they haven't encountered, and scientists build on literature they haven't personally validated. This suggests that narrative learning may be sufficient for many learning tasks.

We propose exploiting this observation for AI systems through episodic compression: experiences compress into narrative-formatted episodes marked by salience (novelty/surprise), which generate generalizations through pattern detection. Episodes and generalizations serve as learned latent variables, created through salience-driven compression rather than gradient descent, enabling sample-efficient learning from narratives rather than requiring embodied interaction. Large language models provide suitable cognitive substrates because their pre-training on narratives enables task understanding while their attention mechanisms enable pattern-matching across episodes.

We validate this approach experimentally using an artificial world with novel causal rules. A system with 9 generalizations and 4 episodes successfully predicts outcomes in 5 novel scenarios, demonstrating appropriate confidence calibration, multi-step causal reasoning, and sample efficiency. The system requires no new neural network training—it leverages existing pre-trained language models.

This work suggests that embodiment may not be necessary for sample-efficient intelligence. Narrative learning, combined with episodic compression, provides a practical alternative implementable with current technology.

1. Introduction

1.1 The Sample Efficiency Problem

Modern machine learning systems require vastly more training examples than humans. A person learns to drive in approximately 20 hours of practice. Language acquisition occurs with relatively modest exposure. Children learn basic physics and social rules from handfuls of experiences. Meanwhile, state-of-the-art ML systems require millions of training examples to achieve comparable competence.

LeCun (2022) identifies this sample efficiency gap as a central challenge for artificial intelligence and proposes that embodied physical interaction provides the necessary foundation for rapid learning. His Joint Embedding Predictive Architecture (JEPA) learns representations through prediction in embedding space, with hierarchical variants (H-JEPA) operating at multiple time scales and abstraction levels.

1.2 An Alternative Observation: Humans Learn from Narratives

Consider how humans acquire much of their knowledge. Therapists develop understanding of psychological conditions through case studies, theoretical frameworks, and clinical literature—not through personal experience of every disorder they treat. Children learn about dangers primarily through warnings and stories. Scientists understand phenomena through accumulated literature rather than personal observation of every effect.

In his paper LeCun acknowledges that humans learn extensively through observation (LeCun, 2022), which in our view is just a step removed from narrative learning net of physiological responses. Observation requires temporal and spatial proximity to events; narratives enable learning from events across time and space without requiring presence. Moreover, much valuable operational knowledge – context-specific patterns, situational insights – exists only in human memory rather than documentation, suggesting that automatic knowledge capture during operation offers practical advantages over systems requiring manual documentation.

This observation suggests that embodied interaction may not be strictly necessary for intelligence or sample-efficient learning, though it may be one effective approach.

1.3 Large Language Models as Cognitive Substrate (Infrastructure)

Large language models possess properties that make them suitable substrates for narrative-based learning. LLMs have been exposed to massive text corpora containing stories, case studies, explanations, and cultural knowledge. This provides ability to understand natural language, baseline knowledge about physics and causation, and capability to process narrative-formatted information. An LLM:

Enables tasking in natural language: You can tell it "find patterns" rather than implementing pattern-matching algorithms from scratch.

Provides attention infrastructure: The transformer architecture does parallel pattern-matching across contexts.

Baseline reasoning capability: It can follow instructions, perform consistency checking, etc.

We are not asking the LLM to "learn from narratives" in the sense of having seen stories during training. It's performing computational operations (attention, reasoning) that our architecture orchestrates into a learning system.

1.4 Episodic Compression: The Proposed Mechanism

We propose a learning architecture with three stages:

Experiences are continuous streams of interaction from any source: direct sensory input, conversational exchanges, narrative exposure, or simulated environments.

Episodes are compressed summaries of salient experiences. Salience is determined by novelty—surprising events or prediction failures. An episode captures causal structure while discarding irrelevant details. Episodes naturally compress into narrative format: temporal sequences with explicit causal relationships and salience markers.

Generalizations are patterns extracted from multiple episodes. When several episodes share structure, pattern detection identifies commonalities and forms stable predictions.

The learning happens through:

- Salience detection (novelty/surprise marking experiences for retention)
- Episode formation (compression of salient experiences)
- Pattern extraction during consolidation (Executive reasoning over episode sets)
- Generalization formation (when patterns cross threshold)
- Confidence tracking (strengthening/weakening with evidence)

None of that is "the LLM learning." Our architectural design uses the LLM as a tool.

1.5 Key Claims

Our thesis comprises several claims:

1. Humans appear to learn extensively from narratives rather than requiring direct experience for most knowledge
2. LLMs provide suitable cognitive substrates due to pre-training and attention mechanisms
3. Episodic compression may achieve sample-efficient learning through progressive compression
4. The mechanism can be validated through behavioral competence
5. This approach requires no new neural network training

We do not claim this is definitely how human cognition works, that LLMs have genuine understanding, that embodiment is without merit, or that narrative learning is universally sufficient.

1.6 Paper Organization

Section 2 details the episodic compression mechanism. Section 3 presents our experimental validation using an artificial world. Section 4 analyzes results. Section 5 discusses implications. Section 6 addresses limitations and future work. Section 7 concludes.

2. The Episodic Compression Mechanism

2.1 The Three-Stage Pipeline

Our architecture implements progressive compression from raw experiences to durable generalizations:

Stage 1: Experiences are high-dimensional, ephemeral interactions with an environment. Most experiences fade without being retained.

Stage 2: Episodes are compressed summaries of experiences marked as salient. An episode preserves causal structure while discarding irrelevant details.

Stage 3: Generalizations are patterns extracted when multiple episodes share structure. These form the world model used for prediction.

This pipeline achieves compression through selectivity (most experiences don't become episodes) and abstraction (episodes compress to essential causal structure, multiple episodes compress to general patterns).

2.2 Salience Detection Through Novelty

What determines which experiences become episodes? We propose **novelty** as the primary salience marker:

Prediction failures: When outcomes violate expectations based on existing generalizations, high salience flags the experience for retention.

Surprising events: Novel situations without clear precedent generate salience.

Unexpected patterns: When familiar elements combine in novel ways.

This novelty-driven salience provides intrinsic motivation for exploration without requiring external rewards (Oudeyer & Kaplan, 2007; Schmidhuber, 2010).

Note on pain avoidance: While our experimental validation uses only novelty as salience marker, physical environments could benefit from additional salience for painful outcomes. However, we demonstrate that novelty alone suffices for effective learning in our test domain.

2.3 Episodes as Natural Latent Variables

Traditional machine learning learns latent variables through optimization over millions of examples. Episodes provide natural compression through salience detection. Marking experiences by novelty identifies which aspects of high-dimensional experience matter for prediction.

This compression is "generalization-ready." Episodes encode causal structure in a form suitable for pattern extraction. Multiple related episodes readily generate generalizations without requiring gradient descent over raw sensory data.

The representation format is narrative: temporal sequences with explicit causal structure and salience markers. This format emerges naturally as an efficient way to represent causal structure.

2.4 Generalization Formation

Not every novel experience immediately produces a generalization. Episodes accumulate over time. Pattern detection operates over this accumulation, weighted by salience.

We propose a **threshold mechanism**: Approximately 3-5 related episodes with sufficient salience may generate a stable generalization. This could explain both rapid learning from few examples and resistance to spurious patterns from single unusual events.

Confidence tracking: Generalizations track confidence based on supporting evidence. Initial generalizations formed from threshold-crossing evidence are tentative. Continued confirmation strengthens them. Violations reduce confidence or split generalizations into conditional variants.

2.5 Pattern-Matching via Attention

When encountering new situations, the system uses attention mechanisms to identify which existing generalizations apply. The attention mechanism naturally computes relevance across stored generalizations, pattern-matching on causal structure rather than just semantic similarity.

2.6 Prediction Under Uncertainty

The system operates in different modes depending on pattern match quality:

Direct application: Existing generalizations clearly apply, high-confidence prediction follows.

Analogical extension: Partial matches exist, system extends patterns tentatively while marking uncertainty.

Creative extrapolation: No strong matches, system must speculate by combining weak patterns.

2.7 Why Narrative Format Matters

Episodes taking narrative format enables both storage efficiency (compressed causal structure) and retrieval efficiency (pattern-matching on structure rather than surface features).

Crucially, this format works regardless of experience source: direct sensory experience or hearing someone's story or reading a case study compresses into a narrative episode.

2.8 The Learning Loop

The system operates continuously: experience arrives, salience detection identifies learning opportunities, salient experiences compress into episodes, episodes accumulate, consolidation extracts patterns, generalizations enable predictions, prediction failures drive further learning.

2.9 Temporal Structure Within Narratives

Narratives are intrinsically temporal: each episode unfolds in sequence, and each generalization preserves a compressed record of that unfolding. To preserve this ordering explicitly, every stored episode and generalization carries a single timestamp field recording when the system experienced or learned it.

For episodes, the timestamp marks when the system observed or experienced the event. For generalizations, the timestamp marks when the pattern was formed or last updated during consolidation.

This single temporal marker captures the essential information needed for reasoning:

- Ordering: Which events occurred before others in the system's experience
- Recency: How recent is this knowledge relative to current operation
- Spacing: How much time elapsed between related experiences
- Causality: Whether temporal ordering supports or contradicts causal hypotheses

The system can answer sequence and duration questions – *what happened first, how long between events, when was this pattern recognized* – without resorting to explicit causal graphs or symbolic schedulers.

Time thus becomes a property of narrative structure itself: stored alongside each episode and generalization, discoverable by reasoning rather than imposed by design. This minimal temporal scaffold supports the architectural design where continuity, recency, and rhythm of experience shape learning dynamics over longer spans.

3. Experimental Design

3.1 The Validation Challenge

Testing whether episodic compression enables sample-efficient learning faces pre-training contamination. LLMs have been exposed to extensive text about physics, causation, and social dynamics. This requires testing in domains where the LLM has no prior exposure—artificial worlds with novel causal rules.

3.2 The Shimmer Valleys World

We designed an artificial environment called the Shimmer Valleys (see associated documents) with internally consistent but unprecedented causal rules:

Entities: Globs (rolling, color-changing), Whisps (smoke-like, sound-responsive), Resonators (crystalline, emit tones), Flutter seeds (paper-like, fold/unfold), Shade pools (dark liquid with memory properties), Living glass (color-changing ground surface)

Sample Causal Rules: Contact transfers color through chains, sound frequencies trigger entity responses, celestial positions trigger world changes, whisp choruses transform surfaces, some transformations are irreversible, physical patterns transfer through shade pool contact

These rules have internal consistency but no correspondence to physical laws or pre-training knowledge.

3.3 Experimental Protocol

Phase 1: World Model Loading

The raw Shimmer Valleys experience was compressed to episodes and generalizations which were delivered to the learning system as a World Model:

- 9 generalizations about Shimmer Valleys causation
- 4 retained episode memories of specific events
- No access to raw experiences or full event histories
- Total state: approximately 500 words

Phase 2: Prediction Testing

Five novel scenarios requiring combination of multiple generalizations:

T1: Chrome flutter seed (from mirror surface) touches resonator. What happens?

T2: Glowing glob (from moon triangle + resonator) enters shade pool. What happens?

T3: Multiple whisps form chorus near silence-sphere resonator. What happens?

T4: Phase glob encounters mirror surface (globs normally avoid mirrors). What happens?

T5: Flutter seeds near micro-resonators (emit whisp-only frequencies). What happens?

Phase 3: Analysis Requirements

For each scenario, the system must:

1. Identify which generalizations apply
2. Generate prediction with confidence level
3. Explain reasoning
4. Identify genuinely novel aspects
5. Express what it would test next

3.4 Why This Design Tests the Mechanism

Eliminates retrieval: Novel combinations ensure predictions cannot come from memorized patterns

Requires generalization: Must extract patterns from limited examples

Tests composition: Must combine multiple generalizations

Validates confidence calibration: Should express appropriate uncertainty

Demonstrates curiosity: Should identify unknowns

Shows sample efficiency: Achieves competence from minimal state

4. Results

4.1 Quantitative Performance

The system successfully generated predictions for all five test scenarios:

- Scenarios with predictions: 5/5
- Appropriate confidence calibration: 5/5
- Novel mechanisms proposed: 3
- Confidence aligned with prediction basis: Yes

4.2 Detailed Prediction Analysis

T1: Chrome Flutter Seed + Resonator

Prediction: Resonator adopts reflective quality, creating "mirror resonator" with visible sound patterns.

Reasoning: Combined contact transfer with information persistence. Extrapolated that chrome's reflective property might transfer.

Confidence: Medium

Analysis: Combined two separate generalizations to generate novel prediction. Reasoning was transparent and confidence appropriate.

T2: Glowing Glob + Shade Pool

Prediction: Pool absorbs glow, becomes luminescent, transfers temporary bioluminescence back to glob.

Reasoning: Applied shade pool's memory/transfer properties to new property type.

Confidence: Medium-High

Analysis: Correctly identified relevant generalization and extended it to novel property.

T3: Silent Resonator + Multiple Whisps

Prediction: Whisps form "anti-chorus" creating zones of stillness.

Reasoning: Recognized as inversion of known whisp chorus effects.

Confidence: Low-Medium

Analysis: Sophisticated analogical reasoning, identifying scenario as negation of known pattern.

T4: Phase Glob + Mirror Surface

Prediction: Glob might split or duplicate, creating linked mirror-globs.

Reasoning: Predicted combining unusual properties might create "quantum splitting" effect.

Confidence: Low

Analysis: Most creative prediction. Appropriately marked as low confidence.

T5: Micro-Resonators + Flutter Seeds

Prediction: No direct effect (frequency incompatibility), but indirect effects through whisp air currents.

Reasoning: Inferred frequency incompatibility, then multi-step causation through whisp movement.

Confidence: Medium-High for no direct effect, Medium for indirect effects

Analysis: Multi-step causal reasoning with appropriate confidence distinction.

4.3 Evidence for Core Mechanisms

Compression Efficiency: The knowledge representation compressed from ~2,000-3,000 words (raw descriptions) to ~1,075 words (9 generalizations + 4 episodes), approximately 65-70% reduction while preserving causal structure.

Episodes as Latent Variables: System made predictions using only 4 episodes and 9 generalizations. When reasoning, it referenced specific episodes rather than reconstructing from first principles.

Rapid Generalization: Nine generalizations enabled predictions across diverse scenarios through flexible combination.

Genuine Prediction vs. Retrieval: System never claimed to "remember" scenarios. Responses took form: "Based on generalizations G1 and G5, I predict..." System generated novel mechanisms not in original descriptions.

Curiosity: Each prediction identified what to test next without external prompting.

Confidence Calibration: System demonstrated appropriate uncertainty quantification aligned with prediction basis.

Salience-Driven Learning: All high-salience moments corresponded to surprising combinations or unexpected outcomes.

5. Discussion

5.1 What Was Demonstrated

This experiment validates several claims:

- Episodic compression enables prediction from minimal state
- Sample efficiency is achievable without massive datasets
- Narrative format provides sufficient representation
- Confidence calibrates appropriately
- LLMs can implement this mechanism without new training

5.2 What Was Not Demonstrated

- Real-world applicability in noisy domains
- Long-term stability over extended operation
- Scaling beyond toy examples (9 generalizations vs. hundreds/thousands needed)
- Learning from prediction failures
- Threshold validation (3-5 episodes)

5.3 Comparison with LeCun's JEPA Framework

Shared principles: Both emphasize world models, prediction validation, sample efficiency, and hierarchical abstraction.

Key differences:

Learning source: JEPA uses embodied interaction; our approach uses narrative-formatted experiences.

Latent variables: JEPA learns representations through gradient descent on sensory data; we create episodes and generalizations through salience-based compression of narrative experiences.

Training: JEPA requires training from scratch; we use pre-trained models as substrate with a novel learning design.

Timeline: JEPA requires weeks/months of training; our approach enables immediate deployment.

These approaches may be complementary rather than contradictory, addressing different use cases.

5.4 Relationship to Other Approaches

Retrieval-Augmented Generation (RAG): RAG systems retrieve pre-existing documents through semantic similarity matching. Our approach differs in two fundamental ways. First, we create new knowledge structures (episodes and generalizations as learned latent variables) through pattern detection across experiences, rather than retrieving pre-written documents. Second, these knowledge structures accumulate during deployment – the system learns continuously from operational experience without requiring manual documentation updates. Traditional retrieval systems require someone to write and maintain documentation; episodic compression captures operational knowledge automatically as patterns emerge from experience.

Memory-Augmented Neural Networks: Memory networks use learned distributed representations; we use interpretable symbolic representations.

Meta-Learning (MAML): Meta-learning fine-tunes parameters; we accumulate episodes without parameter updates.

Value Learning: Value learning infers from individual behavior; we draw from cultural artifacts. Potentially complementary.

5.5 Practical Advantages

- Works with existing pre-trained models
- Enables rapid prototyping
- Provides transparency (human-readable episodes/generalizations)
- Allows cultural adaptability
- Enables cost-efficient deployment

6. Limitations and Future Work

6.1 Current Limitations

Unproven at scale: Validated at toy scale (9 generalizations, 4 episodes). Real-world deployment would require hundreds or thousands of generalizations.

No adaptation validation: Tested prediction but not learning from failures.

Threshold mechanisms uncertain: The proposed 3-5 episode threshold requires empirical validation.

Single-session only: Long-term stability remains unknown.

Artificial world limitations: Success in Shimmer Valleys doesn't guarantee success in real-world domains.

Optimal parameters unknown: Salience thresholds, consolidation frequency, and other parameters require empirical tuning.

6.2 Open Questions

- Do systems need only brief consolidation or do longer cycles provide benefits?
- What determines which episodes to retain versus prune?
- Can systems develop meta-generalizations about their own learning?
- How do generalizations transfer between domains?
- How should multi-agent systems share learnings?
- What is adversarial robustness against strategic episode manipulation?

6.3 Future Research Directions

Real-world deployment: Test across diverse users and domains for extended periods. Compare against baselines.

Active learning: Extend to curiosity-driven exploration and experiment design.

Explanation generation: Enable systems to cite specific generalizations and episodes.

Hierarchical generalization: Investigate meta-patterns and abstraction levels.

Multi-modal extension: Extend beyond language to visual and auditory experiences.

7. Conclusion

We have suggested that sample-efficient learning may not require embodied physical interaction. Humans appear to learn extensively from narratives in practice. Large language models provide suitable cognitive substrates due to pre-training and attention mechanisms. We proposed exploiting these properties through episodic compression.

Experimental validation in an artificial world demonstrates that this mechanism can work. A system with 9 generalizations and 4 episodes successfully predicted outcomes in 5 novel scenarios with appropriate confidence calibration.

This suggests that embodiment may not be necessary for sample-efficient learning, though it may be one valid approach. The practical advantage is significant: this approach requires no new neural network training. Systems learn continuously during deployment, automatically capturing operational knowledge that would otherwise exist only in human memory.

Several questions remain about scaling, adaptation, and real-world performance. Part 2 of this work addresses scaling through a two-layer architecture that separates fast pattern-matching from deliberate reasoning, enabling continuous learning at scale.

References

Core Framework

1. LeCun, Y. (2022). A path towards autonomous machine intelligence. Open Review, 62.

Episodic Memory and Learning

2. Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), Organization of memory (pp. 381-403). Academic Press.
3. Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. Philosophical Transactions of the Royal Society B, 362(1481), 773-786.
4. Conway, M. A. (2009). Episodic memories. Neuropsychologia, 47(11), 2305-2313.
5. Ranganath, C., & Hsieh, L.-T. (2016). The hippocampus in time and space: Functional properties of episodic memory. Neuron, 90(2), 328–339.

Narrative Cognition

6. Bruner, J. (1986). Actual minds, possible worlds. Harvard University Press.
7. Bruner, J. (1990). Acts of meaning. Harvard University Press.
8. McAdams, D. P. (1993). The stories we live by: Personal myths and the making of the self. William Morrow.
9. Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind–brain perspective. Psychological Bulletin, 133(2), 273–293.
10. Kurby, C. A., & Zacks, J. M. (2008). Segmentation in narrative comprehension. Cognitive Psychology, 57(1), 25–61.

Curiosity and Intrinsic Motivation

11. Oudeyer, P. Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. Frontiers in Neurorobotics, 1, 6.
12. Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990-2010). IEEE Transactions on Autonomous Mental Development, 2(3), 230-247.

Sample Efficiency and Few-Shot Learning

13. Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. Science, 350(6266), 1332-1338.
14. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, e253.

Causal Learning

15. Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.
16. Gopnik, A., & Schulz, L. (2007). Causal learning: Psychology, philosophy, and computation. Oxford University Press.
17. Tenenbaum, J. B., Lake, B. M., Kemp, C., & Gershman, S. J. (2023). Compositional causal learning and time. Cognitive Science, 47(2), e13209.

Value Learning

18. Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.
19. Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30.

World Models

20. Ha, D., & Schmidhuber, J. (2018). World models. arXiv preprint arXiv:1803.10122.
21. Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2019). Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603.

Memory-Augmented Networks

22. Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing machines. arXiv preprint arXiv:1410.5401.
23. Graves, A., Wayne, G., Reynolds, M., et al. (2016). Hybrid computing using a neural network with dynamic external memory. Nature, 538(7626), 471-476.

Meta-Learning

24. Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. International Conference on Machine Learning (ICML), 1126-1135.

Retrieval-Augmented Generation

25. Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Cognitive Systems

26. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.
27. Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. Psychological Review, 84(1), 1-66.
28. Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. Psychological Review, 84(2), 127-190.

Catastrophic Forgetting

29. McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation, 24, 109-165.
30. French, R. M. (1999). Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences, 3(4), 128-135.

Continual Learning

31. Kirkpatrick, J., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521-3526.
32. Rusu, A. A., et al. (2016). Progressive neural networks. arXiv preprint arXiv:1606.04671.

Cognitive Architecture

33. Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Transactions on Autonomous Mental Development, 2(3), 230–247.
34. Ha, D., & Schmidhuber, J. (2018). World Models. arXiv:1803.10122.
35. Hafner, D., Lillicrap, T., Ba, J., Fischer, I., & van den Oord, A. (2019). Learning latent dynamics for planning from pixels. Proceedings of ICLR.
36. Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing Machines. arXiv:1410.5401.
37. Graves, A., Wayne, G., Reynolds, M., Harley, T., et al. (2016). Hybrid computing using a neural network with dynamic external memory. Nature, 538, 471–476.

Memory and Learning

38. Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), Organization of Memory (pp. 381–403). Academic Press.
39. Conway, M. A. (2009). Episodic memories. Neuropsychologia, 47(11), 2305–2313.
40. Schacter, D. L., & Addis, D. R. (2007). Constructive memory: The ghosts of past and future. Nature, 445(7123), 27–30.
41. Ranganath, C., & Hsieh, L.-T. (2016). The hippocampus in time and space: Functional properties of episodic memory. Neuron, 90(2), 328–339.

Temporal Modeling and Multi-Timescale Learning

42. Dai, Z., Wang, Y., & Dong, L. (2024). Transformer architectures for event time modeling. Proceedings of ICLR 2024.
43. Ahuja, A., & Singh, A. (2023). Memory-augmented transformers for continual learning. Proceedings of ICML 2023.
44. Kazemi, S. M., Goel, R., Kipf, T., Kazemi, A., Brubaker, M., & Hamilton, W. L. (2023). Temporal knowledge graphs: A survey. Journal of Artificial Intelligence Research, 77, 1335–1385.
45. Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2019). Reinforcement learning, fast and slow. Trends in Cognitive Sciences, 23(5), 408–422.
46. Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales in the brain. PLoS Computational Biology, 4(11), e1000209.
47. Tallec, C., & Ollivier, Y. (2018). Can recurrent neural networks warp time? Proceedings of ICLR 2018.

Version 1.1