Beyond Forgetting: Toward Mechanistically Grounded Continual Learning

Mohammad Rostami
Amazon Generative AI Innovation Center*
rostamii@amazon.com

Abstract

Continual learning (CL) is a prerequisite for autonomous intelligence. Despite decades of progress, the field has operated largely without a mechanistic account of where and why forgetting occurs, treating it as a monolithic phenomenon to be suppressed rather than a structured process to be understood. We argue that three recent shifts are collectively redefining the field: (1) a mechanistic turn that localizes forgetting to specific computational circuits; (2) the emergence of foundation models that change the problem entirely; and (3) the imperative to move beyond task-bounded protocols toward genuinely autonomous, boundary-free learning. Grounded in Complementary Learning Systems (CLS) theory, this work identifies fast and slow learning pathways inside large language models, characterizes the structural geometry of forgetting, and constructs deployable CL pipelines with biologically motivated signatures. These directions chart a path toward autonomous learning systems that regulate their own consolidation.

1 Introduction

Biological intelligence is defined by its continuity: organisms accumulate, reorganize, and selectively refine knowledge over a lifetime without relearning from scratch. This capacity, continual learning, is conspicuously absent in standard deep learning, which trains models once on a fixed dataset and freezes them thereafter. The gap is not a matter of scale; it reflects a fundamentally different relationship between the learner and time. The catastrophic forgetting problem [McCloskey and Cohen, 1989] captures the central failure mode: when a neural network is sequentially trained on new tasks, gradient-based updates overwrite prior knowledge. Decades of work have produced countermeasures such as experience replay [Shin *et al.*, 2017; Rebuffi *et al.*, 2017; Lopez-Paz and Ranzato, 2017; Rostami *et al.*, 2019; Rostami *et al.*, 2020b], parameter regularization [Kirkpatrick *et al.*, 2017; Zenke *et al.*, 2017; Aljundi *et al.*, 2018], architectural isolation [Rusu *et al.*, 2016; Mallya and Lazebnik, 2018; Yoon *et al.*, 2020], distillation [Li and Hoiem, 2018], and task-descriptor-based transfer [Rostami *et al.*, 2020a], yet none has solved the problem at scale and the emergence of large language models (LLMs) and vision language models (VLMs) has opened questions the CL literature treated as settled while making old methods inapplicable.

We argue the field is at an inflection point, driven by three intersecting developments. First, a *mechanistic turn*: the capacity to localize forgetting within a network's computational anatomy, attributing it to functional conflicts at identifiable circuits rather than to undifferentiated parameter drift. Second, *foundation models as CL substrates*: pretrained models with dual adaptation mechanisms, in-context learning and fine-tuning, that alter what forgetting means and what solutions are possible [Bommasani *et al.*, 2021; Wu *et al.*, 2024]. Third, the *autonomy imperative*: real-world deployment demands learning without explicit task boundaries, curated streams, or dense human annotation, a regime that classical benchmarks do not capture [Hadsell *et al.*, 2020]. This perspective traces these shifts through a body of research grounded in CLS theory [McClelland *et al.*, 1995; Kumaran *et al.*, 2016], synthesizes the current landscape, and charts concrete directions for autonomous CL.

2 The Classical Problem and Its Limits

2.1 Algorithms, Benchmarks, and Their Limits

Early connectionist work established that forgetting arises from distributed, overlapping representations [McCloskey and Cohen, 1989]: the parameters that enable flexible generalization also make networks vulnerable to overwriting. Elastic Weight Consolidation (EWC) [Kirkpatrick *et al.*, 2017] offered the first principled Bayesian treatment, using Fisher information to protect parameters critical to prior tasks. Synaptic Intelligence [Zenke *et al.*, 2017] and Memory Aware Synapses [Aljundi *et al.*, 2018] extended this to online settings. Gradient Episodic Memory [Lopez-Paz and Ranzato, 2017] and meta-learning approaches [Riemer *et al.*, 2019] framed CL as constrained optimization over episodic memory. Generative replay methods [Shin *et al.*, 2017; van de Ven *et al.*, 2020] drew on the CLS insight that the neocortex consolidates knowledge via replaying hippocampal traces, instantiating this as a dual-model architecture where a generator replays synthetic past experiences during new-

*Work is independent of position at Amazon.

task training. Architectural isolation methods, from Progressive Networks [Rusu *et al.*, 2016] to PackNet [Mallya and Lazebnik, 2018] and parameter decomposition [Yoon *et al.*, 2020], avoided interference by partitioning capacity across tasks. Knowledge distillation [Li and Hoiem, 2018] offered a complementary approach: preserve the prior model's output behavior through a teacher-student objective.

This accumulated toolkit, surveyed comprehensively in [Parisi *et al.*, 2019; Lange *et al.*, 2022; van de Ven *et al.*, 2022; Wang *et al.*, 2024], established three canonical evaluation scenarios (task-incremental, domain-incremental, class-incremental [van de Ven *et al.*, 2022]) and produced reliable results on controlled image-classification benchmarks. Several developments have since exposed the paradigm's limits. The TRACE benchmark [2023c] documented near-total collapse on mathematical reasoning under sequential fine-tuning of LLaMA2-Chat 13B, a severity unreachable on classical benchmarks. More fundamentally, Dohare et al. [2024] documented *loss of plasticity*: the progressive inability to learn new tasks, entirely independent of forgetting and worsened by standard tools like dropout and Adam. Neuro-inspired approaches have confirmed that incorporating active forgetting alongside stability-protection mechanisms, as in Drosophila-inspired multi-learner architectures [Wang *et al.*, 2023a], can partially alleviate this tradeoff. Together, these findings establish that CL systems face a dual obstacle, and that optimizing an update rule within a fixed network is only a partial answer.

2.2 The Foundation Model Disruption

A further challenge is that *architectural choice dominates algorithmic choice* in the foundation-model era: whether a backbone is frozen or fine-tuned matters more than which CL algorithm is applied to it [Zhang *et al.*, 2023; McDonnell *et al.*, 2023]. This insight, confirmed across class-incremental benchmarks, directly challenges the canonical method comparison paradigm. LLMs and VLMs also introduce parameter-efficient adaptation (LoRA [Hu *et al.*, 2022]) and new CL surfaces such as sequential language tasks [Scialom *et al.*, 2022; Jin *et al.*, 2021; Cai *et al.*, 2023; Ke and Liu, 2022], lifelong domain adaptation [Rostami, 2021; Rostami and Galstyan, 2023b], and multi-agent distributed settings [Rostami *et al.*, 2018], none of which the classical literature addressed. The relevant CL question has shifted: rather than training a randomly initialized network on task sequences, the challenge is adapting a rich pretrained prior to new experiences while preserving its generalist value [Wu *et al.*, 2024; Bommasani *et al.*, 2021]. A method that achieves zero forgetting by freezing the backbone satisfies the letter of the CL objective but fails the spirit if it forecloses genuine new learning, and distinguishing continual adaptation from frozen-feature retrieval.

3 The Mechanistic Turn

3.1 Complementary Learning Systems in Foundation Models

The CLS theory [McClelland *et al.*, 1995; Kumaran *et al.*, 2016] proposes that biological intelligence achieves continual learning by combining a fast hippocampal system (rapid, context-dependent, stateless episodic encoding) with a slow neocortical system (gradual, generalizing, persistent consolidation). Earlier work instantiated this framework computationally via generative replay spaces [Rostami *et al.*, 2019; Rostami *et al.*, 2020b; Rostami, 2021; Rostami *et al.*, 2021], showing that consolidating the internal data distribution, rather than storing raw exemplars, provides a practical approximation of neocortical replay. Neurally inspired extensions confirmed the value of brain-derived architectural priors across biological and artificial systems [van de Ven *et al.*, 2020; Shi *et al.*, 2025; Wang *et al.*, 2023a].

In foundation models, the CLS dual-memory structure is not an imposed architecture but an emergent property. In-context learning (ICL) functions as the fast path: rapid, context-dependent, and stateless (adaptation vanishes when the context window is cleared) [Garg *et al.*, 2022; Olsson *et al.*, 2022]. Fine-tuning functions as the slow path: gradual, generalizing, and persistent. A key mechanistic insight is that the fast path's capacity for *genuine in-context learning*, that is, acquiring capabilities that override pretraining priors rather than merely retrieving pretraining-consistent patterns, is a recipe-determined, editable circuit-level property, not an immutable function of model scale. Models trained with different alignment recipes on the same base checkpoint differ qualitatively in whether the fast path performs genuine acquisition or only prior-consistent retrieval; scale alone does not predict this distinction. Three mechanistic probes (causal lesioning, function-vector tracing, induction-head scoring) converge on early and mid decoder layers as the integration locus, distinct from the late-layer readout locus. This localization yields a principled consolidation trigger: when context-utilization approaches the fast path's capacity limit, adaptation should be consolidated into weights via generative replay. This logic has been instantiated as a deployable CL pipeline matching a joint-training oracle on 15-task benchmarks while operating strictly online, with biologically predicted signatures including superior offline batched consolidation consistent with the memory spacing effect.

3.2 The Structural Geometry of Forgetting

A fundamental reorientation concerns *what causes forgetting*. The dominant view, that representational overlap between tasks causes interference [Kirkpatrick *et al.*, 2017; Wang *et al.*, 2023b], has been directly challenged by controlled experiments. When multi-hop reasoning skills are constructed to span the full range of semantic conflict while holding within-pair circuit overlap constant (dependency cosine > 0.91), forgetting magnitude sweeps the full $[0, 1]$ range based solely on functional conflict. Compatible skills sharing the same circuit are absorbed without forgetting; skills requiring contradictory readouts at the same locus produce near-total forgetting. Structural parameter overlap is the scaffold; functional conflict is the trigger. This dissociation means that parameter-isolation methods (EWC [Kirkpatrick *et al.*, 2017], O-LoRA [Wang *et al.*, 2023b]) protect against the wrong variable: they prevent parameter overlap but do not prevent functional conflict at shared circuits. The implication is that the network location of adaptation, relative to the layers the prior skill depends on, is as important as the

adaptation method. Early and mid layers of LLM decoders serve as the integration circuit for both ICL and sequential fine-tuning; routing a conflicting update to non-critical layers reduces forgetting as a data-free, regularization-free intervention competitive with the standard toolbox.

Forgetting also has structure in a second dimension: fine-tuning intensity. Rather than varying continuously with update magnitude, forgetting exhibits a *phase transition* as a function of total weight drift. Below a drift threshold, forgetting is selective and task-specific relational signals exist, making methods like EWC [Kirkpatrick *et al.*, 2017] and replay [Rostami *et al.*, 2019; Rostami, 2021] appropriate interventions. Above the threshold, however, every capability collapses non-selectively toward chance, and no relational predictor, whether gradient overlap, representational similarity, or in-context conflict scores, can forecast which capabilities are lost. This two-regime structure explains why forgetting predictors work in some evaluations and fail entirely in others: the two settings straddle the phase boundary. Critically, post-collapse performance loss is genuine representational overwriting, not format disruption: linear probes at all network depths fail to recover the prior task's correct answer from hidden states, and re-training on prior data recaptures only a fraction of lost capability. The actionable lever is therefore *total weight drift*: staying below the phase boundary through learning-rate reduction or explicit norm capping retains most of prior capability while fully acquiring the new task, a simpler and more universally applicable recipe than existing geometric methods that manipulate gradient geometry without controlling drift magnitude.

3.3 Forgetting in Vision-Language Models

The mechanistic picture extends to multimodal systems with a sharper localization question enabled by the tripartite VLM architecture (vision encoder, projector, language backbone). In controlled experiments where adversarial visual skills conflict at the same scene content, forgetting localizes to the *language backbone* regardless of whether the conflict was delivered visually or linguistically. The vision encoder and projector, though involved in computing the answer, contribute negligible forgetting when targeted by parameter-matched adapters. This "cross-modal funneling" establishes that VLM forgetting concentrates at the shared computation substrate, not at the modality pathway through which new information entered. Connector- and encoder-side CL mitigations address a secondary locus; backbone-side intervention is where forgetting is controlled. This result also enables a practical *cross-modal repair* strategy: capabilities degraded through one modality can be restored via the other, provided the repair targets the shared language backbone, an immediately actionable technique for VLM-CL systems [Cai *et al.*, 2023].

4 Foundation Models as Continual Learning Substrates

4.1 Parameter Isolation at Scale

LoRA [Hu *et al.*, 2022] has enabled parameter-efficient task partitioning within shared foundation models. O-LoRA [Wang *et al.*, 2023b] constrains task-specific updates to mutually orthogonal low-rank subspaces, preventing new-task learning from overwriting prior-task parameters. Mixture-of-Experts extensions [Liu *et al.*, 2026] partition experts by task while sharing the backbone, addressing the isolation-transfer tradeoff. Prefix-based methods [Scialom *et al.*, 2022] freeze the backbone entirely, trading expressivity for zero-forgetting guarantees. A key caveat is that orthogonality of update subspaces does not guarantee avoidance of functional conflict. Two orthogonal updates can produce contradictory computations at the same circuit locus, because which parameters a circuit depends on is a property of the computational graph, not of parameter-space geometry. Structural parameter isolation and functional conflict avoidance are distinct problems; the field has addressed the former while not the latter. Bridging these two analyses is a priority for future [Wang *et al.*, 2023b; Yoon *et al.*, 2020].

4.2 Self-Supervised and Retrieval-Augmented CL

Self-supervised CL methods [Fini *et al.*, 2022] extend forgetting-avoidance to the label-free regime by converting SSL losses into distillation mechanisms that preserve representation structure across sequential data presentations. Online, task-free paradigms [Pham *et al.*, 2021] address the more demanding setting where task boundaries are absent and inference must be possible at every training moment. Lifelong domain adaptation, the task of adapting to sequentially arriving unlabeled target domains without retaining labeled source data, presents a structurally related challenge where consolidated internal distributions serve as a compact, privacy-preserving stand-in for raw exemplar replay [Rostami, 2021; Rostami and Galstyan, 2023b; Rostami and Galstyan, 2023a]. This approach, proposed for visual domain shift [Rostami, 2021] and extended to concept drift [Rostami and Galstyan, 2023a] and security-critical settings such as novel face presentation attack detection [Rostami *et al.*, 2021], establishes that consolidating a model's internal distribution in a shared latent space, rather than storing samples, is a general and deployable CL primitive. Retrieval-augmented generation offers a non-parametric alternative: store knowledge externally and never modify the parameters that might forget it. For many deployments this is compelling, but it does not eliminate CL; it localizes it to the retrieval interface. When the gap between the query distribution and the retriever's training assumptions becomes large enough that retrieval fails, parametric fine-tuning of query reformulation resurfaces the forgetting problem. CL for RAG middleware, adapting query understanding and reranking while keeping the LLM frozen, is an active applied track that inherits the same structural challenges [Riemer *et al.*, 2019; Rostami *et al.*, 2020a].

5 Toward Autonomous Continual Learning

5.1 The Task-Boundary Problem

Virtually all CL benchmarks assume explicit task boundaries: the system is told when one task ends and the next begins. In real deployments, task identity is unobservable: data streams are continuous, non-stationary, and shift without segmentation cues [Hadsell *et al.*, 2020]. Without boundaries, the consolidation trigger must be inferred from the stream itself,

through a distributional shift detector, a confidence monitor, or an information-theoretic signal that the fast path is approaching capacity. Promising evidence suggests that capacity pressure, rather than a precisely tuned threshold, is the operative consolidation cue, consistent with the biological prediction that hippocampal-to-neocortical replay is triggered by saturation rather than by explicit episode termination [Kumaran *et al.*, 2016; McClelland *et al.*, 1995]. The online, task-free setting [Pham *et al.*, 2021; Fini *et al.*, 2022] is therefore not merely an engineering difficulty but a conceptual test of whether the CLS-inspired trigger mechanism is robust enough to operate without human-defined boundaries, and a key question for autonomous deployment.

5.2 Future Directions

The mechanistic and foundation-model perspectives suggest several concrete directions for autonomous CL research.

Drift-aware, locus-targeted consolidation. Current CL methods apply uniform protection strategies. The phase-transition finding implies that the forgetting regime, selective or non-selective, should determine the intervention: below the drift threshold, locus-specific regularization suffices; above it, the intervention must prevent crossing the boundary rather than managing forgetting within it. A practical autonomous CL system needs a drift monitor that switches strategies accordingly, concentrating protection at the layers where functional conflict has been shown to localize.

Cross-modal autonomous adaptation. The cross-modal funneling result suggests that when one modality provides a rich supervision signal (e.g., language feedback on a failure), a repair at the shared language backbone can transfer to visual performance without requiring paired visual data. This "cross-modal repair transfer" could enable autonomous adaptation to visual distribution shift using only textual supervision, applicable wherever language provides a reliable signal about visual failures such as downstream task degradation.

Plasticity-forgetting joint management. The dual obstacle [Dohare *et al.*, 2024], networks losing both old knowledge and the capacity for new learning, has no unified treatment. Methods that prevent forgetting reduce plasticity; methods that maintain plasticity risk forgetting. Architectures with structurally distinct fast and slow pathways are immune to this tradeoff at the pathway level: the slow path can be protected while the fast path retains full plasticity [Pham *et al.*, 2021; Shi *et al.*, 2025]. Operationalizing this principle for foundation models at scale, and designing consolidation handoffs between pathways to maximize positive transfer rather than merely minimize forgetting [Rostami *et al.*, 2020a; Rostami *et al.*, 2019], remains a key open challenge.

Agentic and open-world CL. Tool-using LLM agents operating over extended deployments face a CL problem that no current benchmark captures: task identity is unobservable, distribution shift is gradual, and the model must simultaneously use and update its knowledge. What "forgetting" means for a reasoning agent whose world model is updated by external feedback, and whose own actions generate the non-stationarity, requires rethinking core assumptions of the field. Multi-agent distributed CL [Rostami *et al.*, 2018], where agents collectively acquire knowledge without centralizing data, represents one promising structural template.

Forward transfer as a design objective. Most CL research minimizes backward transfer while treating forward transfer as a secondary benefit. A mechanistic account of when new learning updates shared representations in ways that improve prior-task generalization, because the new task provides additional examples of the same circuit structure at a compatible locus, suggests that designing adaptation strategies to exploit circuit geometry for positive transfer is a largely unexplored and potentially high-yield direction [Rostami *et al.*, 2020a; Jin *et al.*, 2021; Rostami and Galstyan, 2023a].

6 Discussion and Conclusion

Continual learning is entering its most productive phase in decades, driven by mechanistic understanding of forgetting, foundation models that instantiate CLS architecture at scale, and practical urgency for systems that adapt to the world without periodic full retraining [Hadsell *et al.*, 2020]. The field has accumulated a rich toolkit of algorithmic solutions [Parisi *et al.*, 2019; Lange *et al.*, 2022; Wang *et al.*, 2024] and a growing library of biologically grounded insights [van de Ven *et al.*, 2020; Shi *et al.*, 2025; Wang *et al.*, 2023a], yet it lacks a unified theoretical account of where, why, and under what conditions forgetting occurs.

We have argued for a shift from algorithmic to mechanistic thinking. The question "which update rule prevents forgetting?" has generated hundreds of contested methods. The question "at which circuit, under what conflict conditions, and above what drift magnitude does forgetting occur?" admits controlled experiments and causal interventions that generalize across architectures. CLS theory [McClelland *et al.*, 1995; Kumaran *et al.*, 2016] provides the unifying framework connecting these results to biological memory, grounding deployable pipelines and generating falsifiable predictions. The productive frontier lies in making this framework precise: quantifying functional conflict geometry, characterizing phase boundaries across model families, and designing autonomous systems that regulate their own consolidation.

The path toward autonomous continual learning runs through this mechanistic understanding. Systems that detect their own capacity limits, trigger consolidation appropriately, route conflicting adaptations to non-critical circuits, exploit cross-modal repair transfer, and maintain plasticity through structural pathway separation represent a qualitatively different class of learner. Building them, drawing on the accumulated CL toolkit and the emerging mechanistic science, is the central challenge of the next decade of CL research.

References

[Aljundi *et al.*, 2018] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 144–161, 2018.

[Bommasani *et al.*, 2021] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney

von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models, 2021.

[Cai *et al.*, 2023] Yuliang Cai, Jesse Thomason, and Mohammad Rostami. Task-attentive transformer architecture for continual learning of vision-and-language tasks using knowledge distillation. In *Findings of the Association for Computational Linguistics*, pages 6986–7000, 2023.

[Dohare *et al.*, 2024] Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774, 2024.

[Fini *et al.*, 2022] Enrico Fini, Victor G. Turrisi da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal. Self-supervised models are continual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[Garg *et al.*, 2022] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[Hadsell *et al.*, 2020] Raia Hadsell, Dushyant Rao, Andrei A. Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, 2020.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[Jin *et al.*, 2021] Xisen Jin, Bill Yuchen Lin, Mohammad Rostami, and Xiang Ren. Learn continually, generalize rapidly: Lifelong knowledge accumulation for few-shot learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 714–729, 2021.

[Ke and Liu, 2022] Zixuan Ke and Bing Liu. Continual learning of natural language processing tasks: A survey, 2022.

[Kirkpatrick *et al.*, 2017] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[Kumaran *et al.*, 2016] Dharshan Kumaran, Demis Hassabis, and James L. McClelland. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7):512–534, 2016.

[Lange *et al.*, 2022] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022.

[Li and Hoiem, 2018] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

[Liu *et al.*, 2026] Yang Liu, Toan Nguyen, and Flora D. Salim. CP-MoE: Consistency-preserving mixture-of-experts for continual learning. *arXiv preprint arXiv:2605.20247*, 2026.

[Lopez-Paz and Ranzato, 2017] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems 30*, pages 6467–6476, 2017.

[Mallya and Lazebnik, 2018] Arun Mallya and Svetlana Lazebnik. PackNet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7765–7773, 2018.

[McClelland *et al.*, 1995] James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995.

[McCloskey and Cohen, 1989] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

[McDonnell *et al.*, 2023] Mark D. McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[Olsson *et al.*, 2022] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

[Parisi *et al.*, 2019] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[Pham *et al.*, 2021] Quang Pham, Chenghao Liu, and Steven C. H. Hoi. DualNet: Continual learning, fast and slow. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 16131–16144, 2021.

[Rebuffi *et al.*, 2017] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Riemer *et al.*, 2019] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*, 2019.

[Rostami and Galstyan, 2023a] Mohammad Rostami and Aram Galstyan. Cognitively inspired learning of incremental drifting concepts. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, pages 3058–3066, 2023.

[Rostami and Galstyan, 2023b] Mohammad Rostami and Aram Galstyan. Overcoming concept shift in domain-aware settings through consolidated internal distributions. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, pages 9623–9631, 2023.

[Rostami *et al.*, 2018] Mohammad Rostami, Soheil Kolouri, Kyungnam Kim, and Eric Eaton. Multi-agent distributed lifelong learning for collective knowledge acquisition. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 712–720, 2018.

[Rostami *et al.*, 2019] Mohammad Rostami, Soheil Kolouri, and Praveen K. Pilly. Complementary learning for overcoming catastrophic forgetting using experience replay. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 3339–3345, 2019.

[Rostami *et al.*, 2020a] Mohammad Rostami, David Isele, and Eric Eaton. Using task descriptions in lifelong machine learning for improved performance and zero-shot transfer. *Journal of Artificial Intelligence Research*, 67:673–704, 2020.

[Rostami *et al.*, 2020b] Mohammad Rostami, Soheil Kolouri, Praveen Pilly, and James McClelland. Generative continual concept learning. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 5545–5552, 2020.

[Rostami *et al.*, 2021] Mohammad Rostami, Leonidas Spinoulas, Mohamed E. Hussein, Joe Mathai, and Wael Abd-Almageed. Detection and continual learning of novel face presentation attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14831–14840, 2021.

[Rostami, 2021] Mohammad Rostami. Lifelong domain adaptation via consolidated internal distribution. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 11172–11183, 2021.

[Rusu *et al.*, 2016] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks, 2016.

[Scialom *et al.*, 2022] Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, 2022.

[Shi *et al.*, 2025] Qianqian Shi, Faqiang Liu, Hongyi Li, Guangyu Li, Luping Shi, and Rong Zhao. Hybrid neural networks for continual learning inspired by corticohippocampal circuits. *Nature Communications*, 16:1272, 2025.

[Shin *et al.*, 2017] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems (NIPS 2017)*, 2017.

[van de Ven *et al.*, 2020] Gido M. van de Ven, Hava T. Siegelmann, and Andreas S. Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11:4069, 2020.

[van de Ven *et al.*, 2022] Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4:1185–1197, 2022.

[Wang *et al.*, 2023a] Liyuan Wang, Xingxing Zhang, Qian Li, Mingtian Zhang, Hang Su, Jun Zhu, and Yi Zhong. Incorporating neuro-inspired adaptability for continual learning in artificial intelligence. *Nature Machine Intelligence*, 5(12):1356–1368, 2023.

[Wang *et al.*, 2023b] Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671, 2023.

[Wang *et al.*, 2023c] Xiao Wang, Yuansen Zhang, Tianze Chen, Junjie Ye, Tao Gui, Qi Zhang, Xuanjing Huang, Shengding Hu, Zhiyuan Liu, and Maosong Sun. TRACE: A comprehensive benchmark for continual learning in large language models, 2023.

[Wang *et al.*, 2024] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024.

[Wu *et al.*, 2024] Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey, 2024.

[Yoon *et al.*, 2020] Jaehong Yoon, Saehoon Kim, Eunho Yang, and Sung Ju Hwang. Scalable and order-robust continual learning with additive parameter decomposition. In *International Conference on Learning Representations*, 2020.

[Zenke *et al.*, 2017] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995, 2017.

[Zhang *et al.*, 2023] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. SLCA: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.