

---

# When Does Continual Learning Improve Generalization?

---

Odysseus Kaloidas  
UC Berkeley

## Abstract

Does continuing to learn make an AI system better at solving unfamiliar problems, or merely better at remembering previous ones? Continual learning means learning from a succession of experiences; generalization means applying what was learned beyond the examples encountered. The two can improve together, but continued updating alone does not explain that improvement. A useful experience might reveal a reusable rule, improve adaptation from a few examples, or supply a previously unavailable fact. These possibilities shift attention from how long a model has been learning to how effectively it uses what it learned.

## 1 Remembering the Examples, Missing the Rule

Continual learning is often discussed as a missing ingredient in more capable AI: a model’s education could continue after deployment instead of ending with training. Recent proposals give persistent learning and memory a central role [3]. But what would that continuing education buy? Remembering yesterday’s answer and learning something that makes tomorrow’s unfamiliar problem easier are different achievements.

Consider a learner that records every sequence of zeros and ones it encounters, together with its assigned label, then assigns every unfamiliar string a default label. If the rule concerns whether the number of ones is even, perfect recall leaves many new cases wrong. Recognizing that rule, called parity, gives a way to answer for strings never seen before.

Continual-learning research already distinguishes retention, adaptation, and generalization [12]. Elastic weight consolidation discourages changes to parameters important for earlier tasks [7]; gradient episodic memory constrains updates using stored examples and also examines transfer [8]. The unresolved question is how preserving experience makes a different problem easier.

## 2 How Much Experience Does a New Task Require?

Generalization concerns performance beyond the examples used to form a predictor. Sample efficiency concerns how many examples are needed to reach a specified performance level. They are related, but a sample-efficient learner within a familiar task family may fail under a different kind of shift.

Chollet’s account emphasizes the efficiency of acquiring new skills, taking prior knowledge, assumptions, experience, and difficulty into account [2]. A long learning history need not make a learner efficient at resolving unfamiliar structure from limited evidence.

A learner must choose among possible rules. Suppose every bird in a small set of photographs is against a blue sky. Both wing shape and background color may predict the labels. More photographs from the same setting may reinforce both rules; photographs of birds indoors help separate them. Prior knowledge about objects and backgrounds can make those few examples especially informative.

Learning theory studies how the number of examples, the candidate rules, and the way examples are sampled affect performance beyond the training data [11]. Inductive bias means the assumptions that make some rules more plausible than others before the evidence settles the choice. Good assumptions can reduce the data needed; bad ones can make a learner confidently wrong. Results for independently sampled examples from one distribution do not automatically cover a changing lifetime of experience.

Baxter formalizes how related tasks can help identify a useful family of rules for later learning [1]. This gives continued experience a route to better generalization: earlier tasks can teach the learner what to pay attention to on the next one.

### 3 Adapting What Has Already Been Learned

Return to the bird photographs. A learner with useful prior knowledge may need only a few new examples to discover that wing shape matters here and background color does not. Those examples do more than enlarge its collection: they change its working hypothesis about the task. One way to make this adjustment is to keep the shared model fixed while changing a small internal representation.

The compact hypothesis can be represented by a vector  $z$ , often called a latent task code: a list of  $d$  adjustable numbers, written  $z \in \mathbb{R}^d$ . In the bird example, it could specify which visual features to use when distinguishing the new categories.

To describe how the code is adjusted, let  $f_\theta$  be the predictor with trained parameters  $\theta$ , and  $D = \{(x_i, y_i)\}_{i=1}^n$  the new examples: inputs  $x_i$  paired with answers  $y_i$ . An illustrative objective balances prediction error, measured by a loss function  $\ell$ , against departure from prior learning:

$$z_D^* \in \arg \min_z \left[ \frac{1}{n} \sum_{i=1}^n \ell(f_\theta(x_i, z), y_i) + \lambda \Omega_\theta(z) \right], \quad \hat{y} = f_\theta(x, z_D^*).$$

The first term measures mistakes on the new examples. The penalty  $\Omega_\theta$  discourages codes that conflict with prior learning;  $\lambda \geq 0$  sets its strength and  $n \geq 1$ . Assuming a minimum exists,  $z_D^*$  is the best compromise. Practical optimization may only approximate it. The shared knowledge stays fixed while the code changes, and the code can be discarded after the task. Useful adaptation therefore need not become permanent memory.

Rusu and colleagues' Latent Embedding Optimization turns this idea into a particular adaptation procedure: the code generates classifier parameters, which are then also fine-tuned [10]. Tests on miniImageNet and tieredImageNet demonstrate learning new image categories from few examples. This is evidence for that complete procedure on those tasks; the adjustable code alone does not explain its success.

Model-agnostic meta-learning instead prepares model parameters for rapid adjustment on a new task [4]. Efficient transfer therefore need not use latent optimization. Both approaches seek to make prior learning reusable; permanently retaining each adjustment is a different decision.

A poorly chosen objective can fit the examples without finding the intended rule. A small latent code gives no guarantee by itself. What the code controls, which assumptions guide its adjustment, and how the new task relates to earlier ones still matter.

Reizinger and colleagues show that models with identical optimal test loss can behave differently beyond the conditions represented in training [9]. Finitely many binary-string labels also fit rules other than parity. Preferring parity requires an assumption; more computation cannot make those observations logically determine every unseen label.

### 4 When Continual Learning Strengthens Generalization

Continual learning becomes consequential when new experience improves representations, adaptation procedures, or the evidence available for a task. Dorovatovs and colleagues discuss combining changes stored in model weights with information supplied through context [3]. Persistent memory can support later adaptation even when the model's weights stay fixed.

Harrington and colleagues find more efficient context compression, condensing information supplied to a model, without substantially better new-task learning [5]. Hu and colleagues instead report joint

retention, adaptation, and generalization gains from continual reward-based learning in models that turn images and language into actions [6]. The capacities can improve together, but improvement in one does not establish improvement in the others.

An arbitrary new rule can also make fresh evidence indispensable: no inference procedure can recover a private decision never communicated to it. Continued learning supplies that evidence. Its larger promise is that experience could improve the way later experience is interpreted. A model that has seen more birds might need fewer examples of an unfamiliar species because it has learned which features travel. Understanding when continued learning produces that kind of change would tell us much more than knowing that a system can keep updating.

## References

- [1] Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000. doi: 10.1613/jair.731.
- [2] François Chollet. On the measure of intelligence, 2019. URL <https://arxiv.org/abs/1911.01547>. arXiv:1911.01547.
- [3] Vaggelis Dorovatas, Malte Schwerin, Andrew D. Bagdanov, Lucas Caccia, Antonio Carta, Laurent Charlin, Barbara Hammer, Tyler L. Hayes, Timm Hess, Christopher Kanan, Dhireesha Kudithipudi, Xialei Liu, Vincenzo Lomonaco, Jorge Mendez-Mendez, Darshan Patil, Ameya Prabhu, Elisa Ricci, Tinne Tuytelaars, Gido M. van de Ven, Liyuan Wang, Joost van de Weijer, Jonghyun Choi, Martin Mundt, and Rahaf Aljundi. Position: Modular memory is the key to continual learning agents, 2026. URL <https://arxiv.org/abs/2603.01761>. arXiv:2603.01761. ICML 2026 position paper; June 16 version.
- [4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017. URL <https://proceedings.mlr.press/v70/finn17a.html>.
- [5] Anne Harrington, Nayan Saxena, Michael Murphy, et al. When does continual learning require learning, 2026. URL <https://arxiv.org/abs/2607.07847>. arXiv:2607.07847, preprint.
- [6] Jiaheng Hu, Jay Shim, Chen Tang, Yoonchang Sung, Bo Liu, Peter Stone, and Roberto Martin-Martin. Simple Recipe Works: Vision-Language-Action Models are Natural Continual Learners with Reinforcement Learning. arXiv:2603.11653, 2026. URL <https://arxiv.org/abs/2603.11653>.
- [7] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL <https://arxiv.org/abs/1612.00796>.
- [8] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient Episodic Memory for Continual Learning. arXiv:1706.08840, 2017. URL <https://arxiv.org/abs/1706.08840>.
- [9] Patrik Reizinger, Szilvia Ujváry, Anna Mészáros, Anna Kerekes, Wieland Brendel, and Ferenc Huszár. Position: Understanding LLMs requires more than statistical generalization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42365–42390. PMLR, 2024. URL <https://proceedings.mlr.press/v235/reizinger24a.html>.
- [10] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019. URL <https://arxiv.org/abs/1807.05960>. Conference version, revised March 2019.
- [11] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019.
- [12] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A Comprehensive Survey of Continual Learning: Theory, Method and Application. arXiv:2302.00487, 2023. URL <https://arxiv.org/abs/2302.00487>. Version 3, revised February 2024.