Title
A Resource-Rational Analysis of Forgetting in Continual Learning

Permalink
https://escholarship.org/uc/item/6n43914w

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
Yu, Shirley X.
Ku, Alexander
Griffiths, Tom

Publication Date
2026

A Resource-Rational Analysis of Forgetting in Continual Learning

Shirley X. Yu (shirley.yu@princeton.edu)[1], Alexander Y. Ku[2] & Thomas L. Griffiths[1,2]
[1]Department of Computer Science, Princeton University
[2]Department of Psychology, Princeton University

Abstract

In machine learning, forgetting is typically treated as a defect, whereas human forgetting is understood as a rational response to limited resources and environmental change. We ask whether forgetting can likewise be optimal for machines under realistic constraints. We formalize a continual learning setting, parameterized by a hazard rate governing task volatility and a memory cost penalizing stored data. We compare three replay-based agents: naïve (no memory), unlimited replay (infinite memory), and resource-rational (finite memory). The resource-rational agent significantly outperforms the unlimited replay agent in overall utility across volatile or memory-constrained environments. Optimal memory usage decreases predictably as task volatility and memory cost increase. The resulting forgetting dynamics are best described by an exponential model, showing quantitative alignment with human memory laws. Forgetting thus provides a rational strategy for machines to use when engaging in continual learning in volatile environments or when subject to memory constraints.

Keywords: forgetting; continual learning; memory constraints; concept drift; resource-rational cognition

Introduction

Why do we forget? Cognitive science has long answered this question in functional terms: human forgetting reflects a rational adaptation to limited memory and changing environments (Anderson, 1990; Anderson & Schooler, 1991). The accessibility of a memory tracks the probability that it will be needed again, and limited capacity can itself be beneficial for detecting structure in noisy data (Kareev, 1995; Kareev et al., 1997). Forgetting, in this sense, is not a failure of memory but a feature of a system designed to operate under constraints.

Artificial neural networks also forget. When trained sequentially on changing tasks—the setting known as continual learning—they exhibit catastrophic forgetting, where new learning overwrites earlier knowledge (French, 1999; McCloskey & Cohen, 1989). Yet machine learning treats this phenomenon as a defect to be eliminated. Methods are evaluated by how closely they prevent degradation on prior tasks (Lesort et al., 2019; Wang et al., 2025), with little attention to whether preserving all past knowledge is actually optimal under the constraints learning systems face in practice.

This disconnect raises a question with consequences for both fields: do the rational principles that explain human forgetting also predict the forgetting dynamics of artificial neural networks? If so, the mathematical laws of human memory may reflect general computational solutions to learning under volatility and limited resources, rather than mechanisms specific to biological systems. Answering this question requires analyzing forgetting through the lens of resource rationality (Lieder & Griffiths, 2020), which conceptualizes cognition as the optimal use of limited computational resources.

We pursue this question by treating a simple multi-head multi-layer perceptron (MLP) as a model system in which gradient descent operates under explicit memory constraints. We construct continual learning environments parameterized by a hazard rate h governing task volatility, a memory cost λ, and a buffer size B, and compare three replay-based agents: naïve ($B = 0$), unlimited replay ($B = \infty$), and resource-rational, which selects B to maximize utility. The minimal architecture isolates the dominant forgetting dynamics induced by utility maximization without confounds from deeper or domain-specific models.

The resource-rational agent outperforms unlimited replay across most settings, with infinite memory optimal only when memory is free and the environment is stable. Optimal buffer size decreases predictably as either volatility or memory cost rises, mirroring the dependence of human memory on environmental statistics described by Anderson and Schooler (1991). The forgetting curves produced by the agent are best described by an exponential model, providing quantitative contact with classic descriptions of human memory (Rubin & Wenzel, 1996; Wixted, 2004).

Background

Catastrophic Forgetting in Continual Learning

Neural networks trained sequentially exhibit catastrophic forgetting: newly learned information interferes with and overwrites previously acquired knowledge (French, 1999; McCloskey & Cohen, 1989). This degradation is not due to insufficient capacity but to interference arising from shared parameters (Goodfellow et al., 2013; Krizhevsky et al., 2012), and persists even in highly overparameterized networks (LeCun et al., 2015).

The continual learning literature has converged on three classes of approaches (Van de Ven & Tolias, 2019): replay-based methods that retain past data and interleave it with current training (Lin, 1992; Lopez-Paz & Ranzato, 2017; Shin et al., 2017); regularization-based methods that constrain parameter updates to protect important knowledge (Kirkpatrick et al., 2017; Zenke et al., 2017); and architectural methods that

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

modify model structure over time (Aljundi et al., 2017; von Oswald et al., 2020). Despite their differences, all three treat forgetting as an undesirable outcome to be minimized. None ask whether forgetting itself could be a rational response to environmental change or resource constraints. Among these approaches, replay is particularly suited to a resource-rational analysis: its central design choice—how much past data to retain—directly instantiates a memory cost, making the trade-off between retention and forgetting explicit and tunable. Replay also has the deepest grounding in cognitive science, as we discuss next.

Complementary Learning Systems and the Origins of Replay

Replay-based approaches are not merely engineering solutions; they originate from a neurally motivated theory of how biological systems avoid catastrophic forgetting. The Complementary Learning Systems framework (McClelland et al., 1995; O'Reilly & Norman, 2002) proposes that the brain resolves the stability-plasticity dilemma through two interacting memory systems: a fast-learning hippocampal system that rapidly encodes specific experiences, and a slow-learning neocortical system that gradually integrates them into structured representations. Hippocampal replay during offline periods interleaves recent experiences with older ones, allowing cortical learning to proceed without overwriting prior knowledge, providing both a biological grounding for experience replay in machine learning (Kumaran et al., 2016) and an early demonstration that interleaved rehearsal can mitigate catastrophic interference (McClelland et al., 1995).

While Complementary Learning Systems motivate why replay works, it does not directly address when forgetting should occur, treating forgetting primarily as a failure mode to be mitigated through consolidation. Recent extensions have begun asking which memories should be replayed (Sun et al., 2023), but the broader question of when forgetting is rational under resource constraints has received less attention. Our work complements this line of research by analyzing forgetting itself as an optimization target.

Resource-Rational Perspectives on Memory

Cognitive science has long conceptualized memory and forgetting as products of rational adaptation, asking not how forgetting can be prevented but why it occurs at all. The rational analysis framework models memory as a system optimized for the structure of the environment (Anderson, 1990). Anderson and Schooler (1991) showed that human forgetting curves closely mirror environmental statistics, with accessibility reflecting recency and frequency of events. Forgetting emerges because older information becomes less predictive of future outcomes; relatedly, limited working memory capacity can be valuable because it makes correlations between variables easier to detect (Kareev, 1995; Kareev et al., 1997).

This view has been extended through resource-rational analysis, which emphasizes that cognition operates under strict constraints on computation, memory, and energy (Lieder & Griffiths, 2020; Lieder et al., 2026). Rather than optimizing accuracy alone, cognition balances predictive performance against resource costs. Suchow and Griffiths (2016) formalized memory maintenance as a Markov decision process in which the learner chooses at each step whether to rehearse or release each item. Their analysis showed that optimal retention policies trade off the cost of rehearsal against the value of remembered information, and that several behavioral phenomena in human memory maintenance fall out of the optimal policy—a framing closely related to the utility-maximization perspective we adopt here, though applied at the level of individual memory items rather than aggregate buffer capacity. Van den Berg and Ma (2018) developed a related normative account of visual working memory, in which set size effects emerge as the optimal balance between behavioral cost and the neural cost of encoding precision: as more items must be remembered, the brain rationally distributes a finite resource across them rather than encoding each at maximum precision. Dasgupta and Gershman (2021) synthesized this line of work by characterizing memory itself as a computational resource—a mechanism for reusing past computation through memoization—with forgetting following naturally when retention ceases to pay for itself. Information-theoretic approaches formalize related trade-offs through optimal compression under capacity constraints (Sims, 2016). Across these accounts, selective forgetting is adaptive whenever retention costs exceed predictive value, and the question of what to forget reduces to a question about the predictive value of stored information.

Mathematical Laws of Human Forgetting

Power-law forgetting, in which retention decays as a power function of elapsed time, has been observed across a wide range of memory tasks and often outperforms simple exponential models (Rubin & Wenzel, 1996; Wixted & Carpenter, 2007). Apparent power-law behavior can arise from mixtures of exponential decay processes operating at different rates (Averell & Heathcote, 2011). Pure exponential forgetting, with a single decay rate, is associated with short-term memory paradigms and cognitive architectures (Anderson, 1990; Peterson & Peterson, 1959), while multi-rate models of mixtures or of stretched exponentials are supported by neurocognitive evidence for fast and slow memory processes (Raaijmakers & Shiffrin, 1981; Wixted, 2004).

Summary and Prospectus

While both machine learning and cognitive science study sequential learning under interference, they diverge sharply in how they conceptualize forgetting. In machine learning, forgetting is almost universally framed as a failure to be engineered away. In cognitive science, forgetting is understood as a rational consequence of environmental volatility and resource constraints. This disconnect reveals a central gap: machine learning in continual learning has not systematically analyzed forgetting through a resource-rational lens, nor compared neural network forgetting dynamics to the mathematical laws observed in human memory.

Approach

Continual Learning as Utility Maximization

We reframe continual learning as utility maximization under resource constraints rather than as eliminating forgetting. We formalize this trade-off using a utility function combining hazard-weighted mean squared error (MSE) and memory cost. Hazard-weighted MSE captures predictive accuracy in non-stationary environments by downweighting errors on older tasks:

$$\mathrm{MSE}_{\mathrm{hazard}} = \frac{\sum_{t=0}^{T-1} \sum_{i=0}^{t} (1-h)^{t-i}\, \ell_{i,t}}{\sum_{t=0}^{T-1} \sum_{i=0}^{t} (1-h)^{t-i}} \tag{1}$$

where h is the hazard rate and $\ell_{i,t} = (y_t - \hat{y}_{i,t})^2$ is the squared prediction error at time t using parameters learned from task i. The total utility is

$$U = -\mathrm{MSE}_{\mathrm{hazard}} - \lambda B, \tag{2}$$

where B is the buffer size and λ is the storage cost. Under this view, forgetting can be optimal when retention costs exceed predictive value, shifting the central question from how to prevent forgetting to when and how much forgetting is optimal.

Environment Design

We construct a sequential learning environment using linear regression tasks with stochastic concept drift. While abstract, this setup isolates the tension between stability and plasticity without confounds from high-dimensional feature extraction. Following Anderson and Schooler (1991), we treat memory as reflecting the probability that information will be needed again: the hazard rate h models this need probability by determining how quickly past knowledge becomes outdated, while λ represents the computational or metabolic cost of storage. Varying h and λ defines environments that place different pressures on retention: stable, low-cost settings favor long-term memory, volatile, high-cost settings favor forgetting.

Agents

We compare three replay-based agents differing only in memory management. The naïve agent uses no replay buffer ($B = 0$), representing a lower bound on memory and exhibiting severe forgetting. The unlimited replay agent stores all past data ($B = \infty$), minimizing forgetting but incurring maximal memory cost. The resource-rational agent selects B^* to maximize utility:

$$B^* = \mathrm{argmax}_B \left(-\mathrm{MSE}_{\mathrm{hazard}}(B) - \lambda B\right). \tag{3}$$

This agent introduces no new learning algorithm; it selects among existing replay capacities based on a utility objective, allowing forgetting to emerge as an optimal strategy.

Methods

Experimental Design

The continual learning environment consists of $T = 30$ sequential linear regression tasks, each requiring learning $y = a_t x + b_t + \epsilon$, with $x \in [-1, 1]$, slopes $a_t \in [-3, 3]$, intercepts $b_t \in [-2, 2]$, and noise $\epsilon \sim \mathcal{N}(0, 0.1)$. Each task provided 500 training examples and 500 probe examples for evaluation. Non-stationarity was introduced via a hazard-based drift process: at each time step, previously learned tasks independently drifted with probability h, with parameters perturbed by Gaussian noise ($\sigma = 0.1$) clipped to their original ranges. We evaluated 20 hazard rates spanning $h = 0$ to $h = 1.0$.

All agents use a multi-head MLP with a shared body (one linear layer mapping scalar x to 64 hidden units, followed by ReLU) and 30 task-specific linear heads. Although each task is linear and could be solved independently, the shared layer induces both transfer and interference: updates for later tasks can disrupt representations used by earlier ones. This yields a minimal setting that isolates the effects of limited memory and replay under interference.

The naïve agent used no replay buffer ($B = 0$); the unlimited agent stored all examples ($B = \infty$); fixed-capacity agents used FIFO buffers across 16 sizes from $B = 0$ to $B = 8192$, with a replay ratio of 0.5. Models were trained using Adam (learning rate 0.001) with MSE loss, batch size 64, and 500 gradient steps per task. After each task, agents were evaluated on all tasks learned up to that point. Hazard rates spanned 20 values between 0 and 1.0, and memory costs λ spanned 10 values between 0 and 5×10^{-4}. Experiments were replicated across 50 random seeds, yielding 160,000 evaluation points.

The forgetting curves and utility values reported below are averaged across these 50 seeds for each (h, λ, B) configuration, with retention at delay d computed as the mean MSE on task t evaluated d tasks after t was learned, averaged across all eligible $(t, t + d)$ pairs within a run. Two factors prevent the FIFO buffer from producing a sharp step function in aggregate retention. First, retention within the buffer window is graded rather than perfect: although the buffer guarantees that examples remain available for replay, the shared MLP body is updated by gradient steps on later tasks, progressively interfering with representations supporting earlier tasks even while their data remains in the buffer. Second, drift is stochastic, so the effective age at which a task becomes outdated varies across seeds and tasks within a run, smoothing any sharp transitions that the FIFO mechanism alone would produce.

Analyzing Forgetting Dynamics

Forgetting dynamics were analyzed by fitting power-law, exponential, mixture-of-exponentials, and stretched-exponential models to retention curves using nonlinear least squares. Goodness-of-fit was assessed using BIC[1] and R^2; statistical

[1]We report BIC computed from the residual sum of squares as $\mathrm{BIC} = n \ln(\mathrm{RSS}/n) + k \ln(n)$, which can take negative values when $\mathrm{RSS}/n < 1$ as is common for normalized retention curves. Smaller

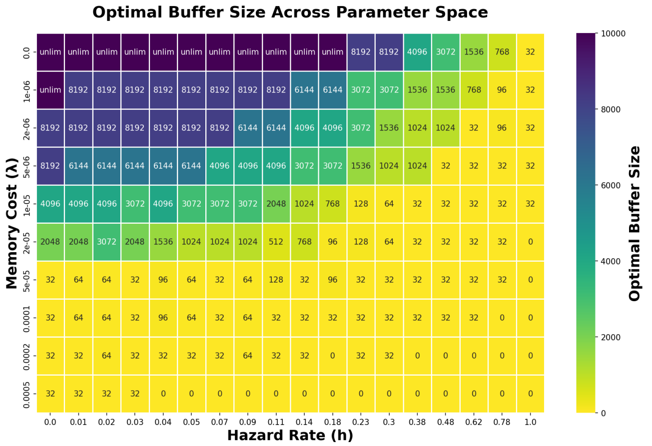

Figure 1: Heatmap of the optimal replay buffer size B across the joint space of hazard rate h and memory cost λ.

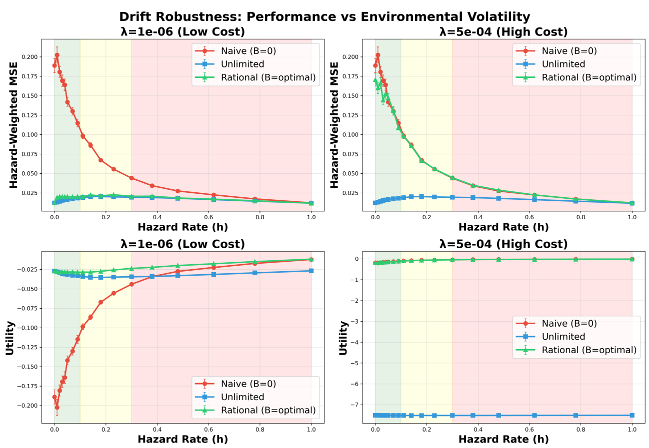

Figure 2: Utility trade-offs under low ($\lambda = 10^{-6}$ smallest non-zero cost) and high ($\lambda = 5 \times 10^{-4}$ largest cost evaluated) memory costs. Error bars indicate standard error.

comparisons used paired t-tests.

Results

We report two main findings. First, finite memory outperforms unlimited replay across the realistic parameter space, with the advantage holding consistently rather than appearing only at extreme values of memory cost or volatility. Second, the forgetting curves produced by utility maximization take a specific functional form—exponential decay—that emerges from optimization rather than being built into the utility function or the architecture.

Figure 1 shows two systematic trends. As memory cost increases, the optimal buffer size decreases sharply: when storage is expensive, aggressive forgetting becomes optimal even in stable environments. As volatility increases, the optimal buffer also decreases, reflecting the value of selectively forgetting outdated data. These effects interact nonlinearly: unlimited memory is optimal only when $\lambda = 0$ at small h, and even small positive memory costs lead to finite optimal buffers. At high memory costs, the optimal buffer collapses toward zero across most hazard rates.

By construction, the resource-rational agent achieves maximal utility for each (h, λ) configuration. As shown in Figure 2, when memory is cheap it selects large buffers and matches the unlimited agent; when memory is expensive it selects small buffers and resembles the naïve agent. Although the unlimited agent maintains low prediction error, its utility degrades substantially at high memory costs due to the storage penalty. Minimizing forgetting alone is therefore insufficient. Optimal performance depends jointly on accuracy and resource costs.

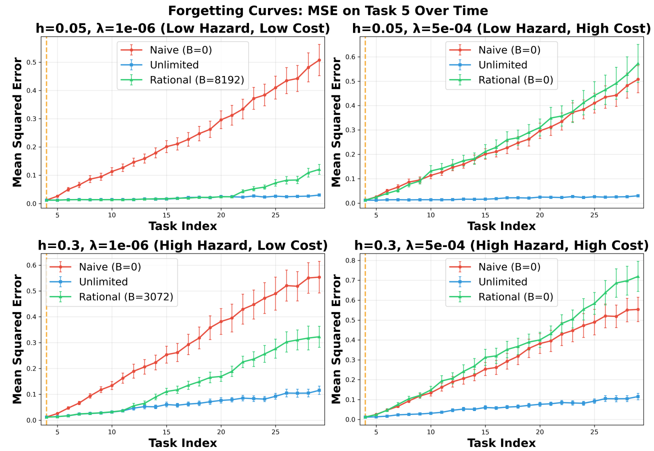

Figure 3: Forgetting curves under varying volatility and memory cost. Low and high memory costs λ match Figure 2. Hazard rates were set to $h = 0.05$ (stable) and $h = 0.3$ (volatile). Error bars indicate standard error.

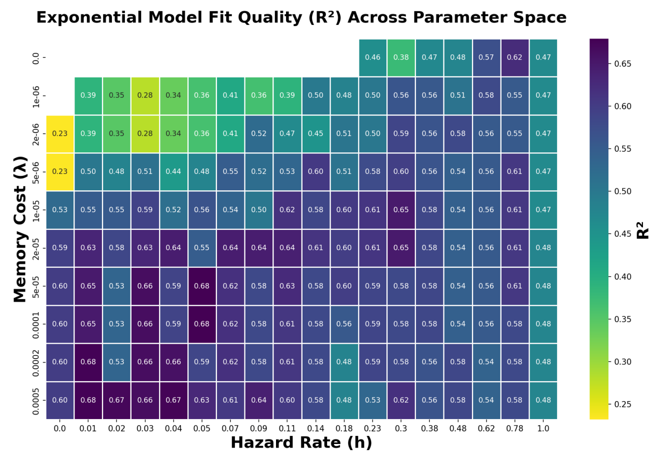

Figure 4: Heatmap of R^2 values for the exponential model fit across hazard rate h and memory cost λ.

(more negative) values indicate better fit under both this and the standard likelihood-based formulation.

Table 1: Goodness-of-fit comparison across classic models of forgetting dynamics using Bayesian Information Criterion (BIC).

Model	**Mean BIC ± SD**	**Best BIC**	**Best Config** (h, λ, B)	**Model Formula**
Power Law	-92.90 ± 19.32	-153.06	$(0.01, 2e-06, 8192)$	$R(t) = a(t+1)^{-b}$
Exponential	-105.44 ± 17.82	$\mathbf{-162.34}$	$(0.01, 1e-06, 8192)$	$\mathbf{R(t) = ae^{-bt}}$
Mixture	-104.80 ± 16.13	-156.36	$(0.01, 2e-06, 8192)$	$R(t) = w_1 e^{-b_1 t} + w_2 e^{-b_2 t}$
Stretched Exp	-100.63 ± 15.10	-147.63	$(0.01, 2e-06, 8192)$	$R(t) = a(t+1)^{-b} e^{-ct}$

Table 2: Statistical comparison of agents and forgetting models. The first row compares utility between agents; remaining rows compare forgetting models to identify the best fit for the forgetting curves produced by the MLP.

Comparison	**Metric**	**Mean Difference**	t**-statistic**	p**-value**
Rational vs Unlimited	Utility Gain	1.383	7.84	$p < .0001$
Exponential vs Power Law	BIC Difference	-12.53	-73.58	$p < .0001$
Exponential vs Mixture	BIC Difference	-0.64	-2.05	$p = 0.042$
Exponential vs Stretched Exp	BIC Difference	-4.81	-21.27	$p < .0001$

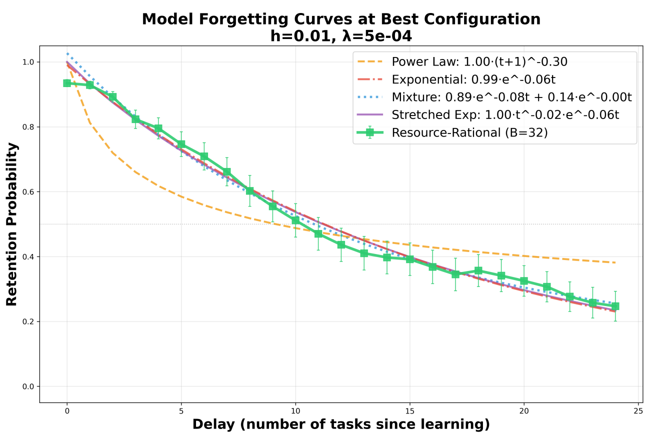

Figure 5: Observed forgetting curve of the resource-rational agent, overlaid with fitted models averaged over 50 seeds. Error bars indicate standard error.

Figure 3 shows that in low-cost environments, the resource-rational agent exhibits intermediate retention; in high-cost environments, it forgets rapidly, tracking the naïve agent. Forgetting is graded rather than binary, adapting continuously to memory cost and environmental statistics, suggesting that forgetting emerges as an adaptive response to external pressures rather than as an intrinsic failure of learning.

Table 1 compares goodness-of-fit across four candidate models. The exponential model achieves the largest absolute mean BIC and the highest best-case fit. Power-law models systematically underfit; stretched-exponential models perform moderately but remain inferior. The mixture-of-exponentials model performs slightly worse than the pure exponential, with the difference not statistically significant (Table 2). This is consistent with our deliberately simplified MLP architecture, which isolates the dominant forgetting timescale induced by utility maximization. The best-fitting configuration occurs at $h = 0.01$ and $\lambda = 5 \times 10^{-4}$, with $B^* = 32$, corresponding to low volatility and high memory cost.

Figure 4 shows that high-quality exponential fits concentrate in low-hazard, high-cost environments, where structured forgetting dynamics emerge most clearly; volatile environments yield noisier retention patterns less well captured by parametric decay models. Figure 5 shows a representative best-fitting case, where the resource-rational agent's forgetting curve closely matches the fitted exponential model.

Together, these results show that optimal memory usage in continual learning depends jointly on environmental volatility and memory cost, that resource-rational forgetting achieves higher utility than unlimited memory under realistic constraints, and that the resulting forgetting dynamics follow structured, human-like mathematical laws.

Discussion

This work reframes catastrophic forgetting in continual learning as a consequence of rational trade-offs between predictive performance and memory cost. Rather than treating forgetting as a failure to be eliminated, we show that forgetting can be optimal when learning systems operate under resource constraints in non-stationary environments.

Forgetting as Resource-Rational Adaptation

Our central finding is that the rational principles cognitive science has used to explain human forgetting also describe the dynamics of a simple neural network. Across all environments with non-zero memory cost, the resource-rational agent achieved higher utility than the unlimited replay agent. Unlimited memory was optimal only when $\lambda = 0$ and h was small. This is an unrealistic regime in which storage is free and the environment is stable. The advantage of finite memory was substantial and consistent across the full parameter space ($t = 7.84$, $p < .0001$; Table 2), not confined to extreme values of λ or h. The cognitive science view of forgetting as adaptive, rather than the machine learning view of forgetting as failure, generalizes to artificial systems once realistic constraints are made explicit.

The dependence of optimal memory on environmental structure parallels predictions from rational analyses of human memory. As hazard rate or memory cost increased, the optimal buffer shrank. This dual sensitivity to environmental statistics and resource constraints mirrors what Anderson and Schooler (1991) identified in human memory, where accessibility tracks the probability that information will be needed again, modulated by capacity. The emergence of this pattern in a purely machine learning setting suggests that the relationship between environmental statistics and memory allocation is not a quirk of biological implementation but a property of any system that allocates memory rationally under change.

The forgetting dynamics of the resource-rational agent were best described by a pure exponential model, which outperformed power-law and stretched-exponential alternatives (Tables 1, 2). Critically, the functional form of forgetting is not specified by our utility function—which contains only a hazard rate, a memory cost, and a buffer size—and is not built into the architecture. That an exponential law emerges from gradient descent under these constraints is a property of the optimization, not an assumption of the model.

This exponential result is the prediction the framework makes for a system with a single dominant memory timescale, and it places our findings in direct continuity with the classic literature on human forgetting. Apparent power-law forgetting in humans has long been understood as compatible with mixtures of exponential processes operating at different rates (Averell & Heathcote, 2011). A single-timescale system should produce a single exponential; a multi-timescale system should produce a mixture that closely approximates a power law. Our model satisfies the first prediction. The second generates a clear empirical hypothesis: extending the framework to architectures with multiple interacting memory stores—such as the fast and slow systems posited by Complementary Learning Systems theory (Kumaran et al., 2016; McClelland et al., 1995)—should yield mixture-of-exponential dynamics that better approximate human power-law forgetting. The discrepancy between our exponential result and human power laws is therefore not evidence against the resource-rational account, but a constraint on which architectures the account predicts will reproduce human-like forgetting.

Taken together, these results indicate that forgetting in machine learning need not be viewed as inherently detrimental, and that the laws of human memory need not be viewed as exclusively biological. When learning objectives are aligned with environmental statistics and resource constraints, forgetting emerges as a rational consequence of optimization, and its functional form quantitatively resembles classic descriptions of human memory. This convergence supports the broader claim of resource-rational analysis: that many features of cognition reflect optimal solutions to problems shared across biological and artificial systems.

Limitations and Future Directions

The learning environment is simplified, the architecture is fixed, and we focus on replay-based mechanisms with a single memory store. CLS (Kumaran et al., 2016; McClelland et al., 1995) suggests that biological memory operates through multiple interacting systems—fast-learning episodic and slow-learning semantic memory—which may exhibit distinct optimal forgetting dynamics. Extending our framework to multi-system architectures could test whether mixture-of-exponential curves emerge naturally when faster and slower stores interact, potentially reconciling our exponential findings with human power-law forgetting (Spens et al., 2026; Sun et al., 2023). The framework also abstracts away from several aspects of memory that a fuller resource-rational treatment would incorporate: the brain may modulate memory not only through retention but through selective encoding, lossy compression at variable rates, and content-dependent retention reflecting the predicted utility of specific memories rather than uniform decay. Estimating the future utility of a memory is itself a non-trivial inferential problem, and learning memory-allocation policies directly rather than selecting buffer sizes from a fixed grid would provide a richer account of how cognitive systems decide what to retain or forget over time.

Conclusion

We have shown that forgetting in continual learning need not be treated as a defect, and that the rational principles cognitive science has used to explain human forgetting predict the forgetting dynamics of a simple neural network operating under realistic constraints. Finite memory often outperforms unlimited replay, and the resulting forgetting curves take the exponential form expected of a rational memory system with a single dominant timescale. This convergence suggests that the laws of memory described by cognitive science may reflect general optimization principles rather than specialized biological mechanisms, motivating treatment of forgetting as a phenomenon to be understood rather than eliminated.

Acknowledgments

We thank Professor Brenden M. Lake for valuable feedback that helped shape this work. We used large language models to assist with setting up experimental model runs across agents, LaTeX formatting of the manuscript, and general proofreading.

References

Aljundi, R., Chakravarty, P., & Tuytelaars, T. (2017). Expert gate: Lifelong learning with a network of experts. https://arxiv.org/abs/1611.06194

Anderson, J. R. (1990). *The adaptive character of thought* [Original work published 1990 by Lawrence Erlbaum Associates]. Psychology Press. https://doi.org/10.4324/9780203771730

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408. https://doi.org/10.1111/j.1467-9280.1991.tb00174.x

Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, *55*(1), 25–35. https://doi.org/10.1016/j.jmp.2010.08.009

Dasgupta, I., & Gershman, S. J. (2021). Memory as a computational resource. *Trends in cognitive sciences*, *25*(3), 240–251.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135. https://doi.org/10.1016/S1364-6613(99)01294-2

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., & Bengio, Y. (2013). An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition*, *56*(3), 263–269.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General*, *126*(3), 278–287.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, *20*(7), 512–534.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lesort, T., Lomonaco, V., Stoian, A., Maltoni, D., Filliat, D., & Díaz-Rodríguez, N. (2019). Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *arXiv preprint*. https://doi.org/10.48550/arXiv.1907.00182

Lieder, F., Callaway, F., & Griffiths, T. L. (2026). *The rational use of cognitive resources*. Princeton University Press.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1. https://doi.org/10.1017/S0140525X1900061X

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, *8*(3–4), 293–321. https://doi.org/10.1007/BF00992699

Lopez-Paz, D., & Ranzato, M. (2017). Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*, *30*, 6470–6479.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). 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.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *Psychology of learning and motivation* (pp. 109–165, Vol. 24). Academic Press. https://doi.org/10.1016/S0079-7421(08)60536-8

O'Reilly, R. C., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in cognitive sciences*, *6*(12), 505–510.

Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*(3), 193–198. https://doi.org/10.1037/h0049234

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93–134. https://doi.org/10.1037/0033-295X.88.2.93

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734–760. https://doi.org/10.1037/0033-295X.103.4.734

Shin, H., Lee, J. K., Kim, J., & Kim, J. (2017). Continual learning with deep generative replay. *CoRR*, *abs/1705.08690*. http://arxiv.org/abs/1705.08690

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198. https://doi.org/10.1016/j.cognition.2016.03.020

Spens, E., Burgess, N., & Behrens, T. E. J. (2026). Modelling the control of offline processing with reinforcement learning. *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=BFW1fkB8ck

Suchow, J. W., & Griffiths, T. L. (2016). Deciding to remember: Memory maintenance as a markov decision process. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *38*.

Sun, W., Advani, M., Spruston, N., Saxe, A., & Fitzgerald, J. E. (2023). Organizing memories for generalization in complementary learning systems. *Nature neuroscience*, *26*(8), 1438–1448.

Van de Ven, G. M., & Tolias, A. S. (2019). Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*.

Van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *ELife*, *7*, e34963.

von Oswald, J., Henning, C., Sacramento, J., & Grewe, B. F. (2020). Continual learning with hypernetworks. *International Conference on Learning Representations*.

Wang, Z., Yang, E., Shen, L., & Huang, H. (2025). A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(3), 1464–1483. https://doi.org/10.1109/TPAMI.2024.3498346

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235–269. https://doi.org/10.1146/annurev.psych.55.090902.141555

Wixted, J. T., & Carpenter, S. K. (2007). The wickelgren power law and the ebbinghaus savings function. *Psychological Science*, *18*(2), 133–134. https://doi.org/10.1111/j.1467-9280.2007.01862.x

Zenke, F., Poole, B., & Ganguli, S. (2017). Continual learning through synaptic intelligence. *International Conference on Machine Learning*, 3987–3995.