Reconciling Rational and Mechanistic Accounts of the Power Law of Forgetting

Qiong Zhang[1] and Thomas L. Griffiths[2]

Psychology Department, Computer Science Department, Rutgers University – New Brunswick[1],

Psychology Department, Computer Science Department, Princeton University[2]

Author Note

Correspondence concerning this article should be addressed to Qiong Zhang (qiong.z@rutgers.edu).

Abstract

A fundamental property of memory is its decline over time, a process widely observed to follow a power law. Two primary accounts have been proposed to explain this phenomenon. A rational account considers the power law of forgetting arising from optimal adaptation to environmental statistics, but does not specify how human memory carries out such adaptation. A mechanistic account considers the power law of forgetting arising from a weighted combination of exponential trace decays, constrained by the immutable structure of memory, but it remains unclear why these weights are combined in this particular way. To reconcile these two accounts, we propose that the power law of forgetting emerges from both environmental demands and architectural constraints. Our theoretical analysis and simulations demonstrate that the memory system's architectural constraints can support estimating how quickly information decays in the environment, aligning its own rate of forgetting to match that of the environment. This adaptation would ensure that information is forgotten quickly when it becomes irrelevant, and retained longer when it remains useful. This provides a cognitively plausible model that explains past empirical results where human memory is observed to adapt to changing environmental statistics on a short timescale. Together, these results offer a unified, mechanistic, and rational understanding of what gives rise to the shape of memory forgetting.

Reconciling Rational and Mechanistic Accounts of the Power Law of Forgetting

Introduction

Introspection suggests that recent experiences are more vivid and easier to remember. The exact form of how memory declines over time has been studied since the beginning of modern memory research (Ebbinghaus, 1885/1913). Early studies focused on searching for a mathematical function that describes memory forgetting in empirical data. These studies provide the strongest support to the power law function, indicating that memory fades rapidly at first and then slows down gradually (Rubin & Wenzel, 1996; Wixted, Carpenter, et al., 2007; Wixted & Ebbesen, 1991). More recent studies have focused on developing theories to explain this power law relationship. Two primary approaches have emerged: a rational approach, which posits that the power law is an optimal adaptation to the statistical structure of the environment (J. R. Anderson, 1990; J. R. Anderson & Schooler, 1991), and a mechanistic approach, which considers the power law function a result of the immutable architecture of human memory (R. B. Anderson & Tweney, 1997; Howard et al., 2015; Hutchinson et al., 2025). While both accounts are influential, a theory that integrates them is lacking. The goal of this article is to provide such a unification. We propose that the power law of forgetting is both adaptive, driven by environmental goals, and immutable, constrained by the brain's fixed cognitive architecture. We argue that neither approach alone provides a complete explanation.

The rational account of forgetting views the power law function as arising from the adaptive nature of human memory. Under this account, the primary goal of memory is not to preserve the past perfectly but to retain information that is likely to be useful for current and future tasks (J. R. Anderson, 1990; J. R. Anderson & Milson, 1989). Therefore, our memory system should learn from statistical patterns in the environment to predict when information will be needed and forget what is no longer relevant. A key prediction of this account is that these environmental statistics should mirror the form of the forgetting function. Supporting this, studies have shown that the relationship between an item's recency and its probability of being needed again follows a power law in various environmental sources, including the *New York Times*, parental speech, and emails (J. R. Anderson and Schooler, 1991; Figure 1a) as well as

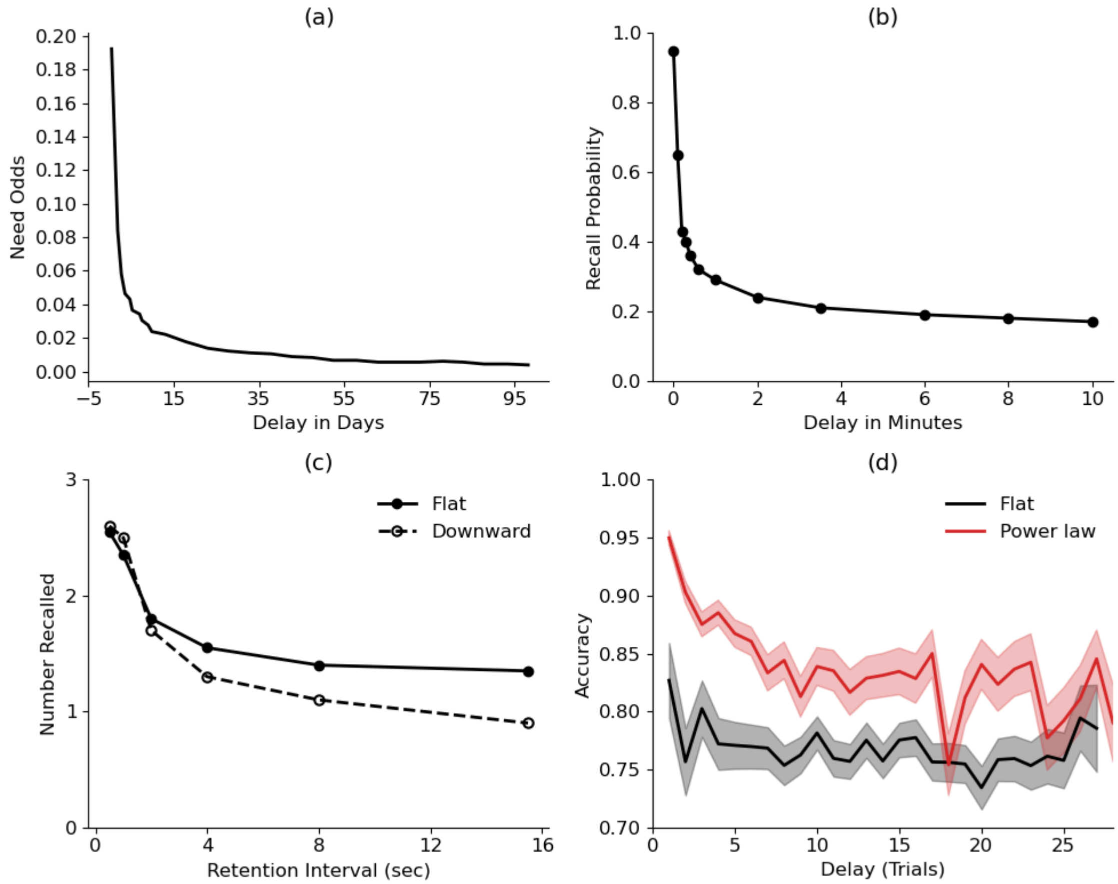

Figure 1. Past empirical evidence supporting a rational account of memory forgetting. (a) The environmental recency function, displaying the relationship between stimulus recency and the odds of being needed in the environment based on words appearing in the *New York Times* headlines (1986 and 1987), reproduced from Figure 4a in Schooler and Anderson (1997). (b) Forgetting curve, displaying the relationship between delay and memory performance of paired associates, reproduced from Rubin et al. (1999). (c) Forgetting curves towards the end of the experiment (last two blocks), displaying the relationship between the amount of time passed and memory performance in two conditions where the environmental recency function is either flat or downward, reproduced from Figure 3 in R. B. Anderson et al. (1997). (d) Categorization behavior towards the end of the experiment (last two blocks), displaying the relationship between the amount of time passed and categorization performance in two conditions where the environmental recency function is either flat or a power law, generated using data from Devraj et al. (2024). In (c) and (d), a block refers to a series of chronologically continuous trials.

Twitter messages and Reddit comments (J. R. Anderson et al., 2023), similar to that of human forgetting observed in laboratory experiments (Figure 1b; reproduced from Rubin et al. (1999)). To move beyond correlation and establish a causal role of environmental statistics in shaping memory forgetting, researchers have directly manipulated these environmental statistics. These experiments confirm that changing the patterns of stimulus presentation in an environment directly alters the rate of memory forgetting (R. B. Anderson et al., 1997; Figure 1c) and impacts downstream tasks like categorization that rely on memory (Devraj et al., 2024; Figure 1d). As shown in Figure 1c and d, a flattened recency function in the environment gives rise to slower decay of memory performance and categorization performance over time. While existing evidence supports a rational account, explaining why memory behaves in this way, it remains an open question how human memory carries out such adaptation. Addressing this requires an integration of the rational account with the knowledge of the underlying cognitive processes of forgetting.

Decades of theoretical work in memory have examined cognitive processes underlying stages of memory encoding, storage, and retrieval, often formulated in precise computational models (Kahana, 2020). Having been validated against a range of empirical data, the cognitive mechanisms specified in these models represent the architectural constraints of human memory that must be considered when explaining new behavior or phenomena. Several mechanistic models of memory have assumed that there is a contextual drift over time that guides memory encoding and recall (Bower, 1967; Howard & Kahana, 2002; Mensink & Raaijmakers, 1988; Murdock, 1997; Polyn et al., 2009), which gives rise to exponential trace decay. To reconcile how exponential trace decay can support a power law, R. B. Anderson and Tweney (1997) demonstrated that a power law function can arise from arithmetic averaging of exponential functions. This suggests that memory strength could decrease exponentially at the item level but appears as a power law when aggregated across items and subjects, a result that can occur under very general conditions (Kahana & Adler, 2017). While these earlier works caution against interpreting power law forgetting at the level of individual items, subsequent computational and empirical evidence demonstrate that the power law is not merely an artifact of averaging, but operates at the level of a single memory trace (Donkin & Nosofsky, 2012; Mozer et al., 2009). Further support for the power law at the single-trace level comes from recent neural and theoretical developments, which show that mental context drifts at multiple rates (Shankar & Howard, 2012). These varying drift rates create a mixture of exponential functions that can combine to produce a power law function of forgetting (Howard et al., 2015; Tiganj et al., 2019). Although this literature provides a rich mechanistic account of the power law of forgetting as a weighted sum of exponential decays, it leaves open the question of why these weights are combined in this particular way and how this mechanism supports memory behavior when it adapts to new environments. Addressing these questions requires integrating the mechanistic account with an analysis of environmental goals in a rational account.

To unify the rational and mechanistic accounts, we build a model of how human memory adapts to the environmental statistics under the cognitive constraints specified by mechanistic models of memory. We follow the approach of resource-rational analysis (Lieder & Griffiths, 2020), which considers cognitive behavior to be shaped by both the environmental goals and limited computational resources. We define the goal of the memory system as adjusting its forgetting function so that memory availability matches the probability of memory being needed in the environment. We then derive the optimal solution to this problem under the constraint that memory decays exponentially at multiple rates. Importantly, it may be challenging for the cognitive system to track the need probability for each piece of information individually through its history of occurrences. We assume that a simpler solution is to track how quickly information generally becomes obsolete in the environment. Our model proposes that the memory system updates its belief about this rate of change, aligning its own rate of forgetting to match that environmental timescale. To foreshadow our results, model simulations show that under these assumptions, the memory system can effectively adapt by updating its beliefs based on environmental observations. It is capable of finding a set of weights for exponential decays that align its forgetting function with environmental statistics, even when relying on the incomplete environmental information typical of laboratory-based memory experiments.

The work we present here was partially motivated by a critique of Devraj et al. (2024) raised by Hutchinson et al. (2025). Hutchinson et al. (2025) argued that although a power law of forgetting may have originated as an adaptation to the environmental statistics of the world, it has since become an immutable feature of the memory system observed across a wide range of tasks and stimulus sets, including where environmental recency is no longer present (Donkin & Nosofsky, 2012). They further critique the plausibility of adaptation on a short timescale, arguing that if memory can indeed adapt to new environmental statistics within a single experiment, a rational account is incomplete without an explanation of how the mind carries out such a rapid adaptation. Our present work directly addresses their critique by demonstrating how cognitive mechanisms underlying memory forgetting enable memory to rationally adapt to changing environmental statistics.

In the remainder of the paper, we first provide background on the cognitive mechanisms supporting exponential trace decay and explain how they produce a power law at the item level. We then lay out the steps for building a resource-rational model and demonstrate its ability to adapt to a given environmental statistic under the constraints of these cognitive mechanisms. These results provide a unified, mechanistic, and rational understanding of what gives rise to the power law of forgetting.

Background: Mechanisms of memory forgetting

In this section, we review cognitive mechanisms that support exponential trace decay, and how exponential trace decay at different rates supports a power law forgetting function at the level of individual items. These cognitive mechanisms serve as the architectural constraints of human memory that must be considered when explaining adaptive behavior of memory forgetting.

Exponential trace decay through drifting temporal contexts

Past work has long recognized the important role of context in the encoding and retrieval of information. People remember more information when tested in their original environmental context than in a new one (Godden & Baddeley, 1975; Smith, 1979). Context can also be endogenous, slowly changing over time, and bind with encountered experiences. Early temporal context models assume a randomly drifting mental context (Bower, 1967). More recent models, including the Temporal Context Model (TCM; Howard and Kahana, 2002) and its successor, the Context Maintenance and Retrieval model (CMR; Polyn et al., 2009; Lohnas et al., 2015), assume that the context drifts towards new experiences as well as previously encountered contexts associated with these experiences. As the context similarity between study and test decreases, forgetting increases.

Specifically, the Temporal Context Model posits that when participants encode a list of sequentially presented items over time, the state of the context c_i at time step i is a combination of the context from the previous state c_{i-1} and an input vector f_i:

$$c_i = \rho\, c_{i-1} + \beta\, f_i \tag{1}$$

where f_i represents the retrieved context induced by an encountered experience, $\beta \in [0, 1]$ is a parameter determining the rate at which context drifts toward the new context, and ρ is a scalar ensuring that c_i is of length 1. For mathematical simplicity, one may assume that the input vectors f are mutually orthogonal to each other, and each of length 1, then we have ρ as a constant value written as $\rho = \sqrt{1 - \beta^2}$, following derivations in Howard (2004). Under these assumptions, the similarity of any two temporal context vectors, written as a dot product, falls off as an exponential function of the time steps between them (let $j > i$):

$$c_i^\top c_j = \rho^{(j-i)} \tag{2}$$

The context vectors themselves are not sufficient to support recall. To support recall, they need to form associations with items during encoding, held in M^{CF}_{exp}. These associations are initialized to zero at the beginning of the encoding, and as items come in during each time step, they are updated via the Hebbian outer-product learning rule. Specifically, when an item is encoded at timestep i, an association is formed between the context state c_i and the presented item f_i:

$$M^{CF}_i = M^{CF}_{i-1} + f_i c_i^\top \tag{3}$$

During recall, to determine which item is to be retrieved at timestep j, the TCM examines how much the current context c_j matches with all items' encoding contexts, as stored in rows of $M^{CF}_{\text{exp}} = \sum f_i c_i^\top$. At time step j, the memory activation strength for the item encoded at time step i, given the current context c_h as the retrieval cue, can be expressed as

$$a^i_j = f_i^\top M^{CF} c_j = f_i^\top (\sum f_i c_i^\top) c_j = c_i^\top c_j \tag{4}$$

where the last equality used the assumption of orthonormality on the f vectors. In other words, the memory activation strength reflects the similarity between the context at which item i was encoded, c_i, and the present context c_j during recall. Intuitively, items whose original encoding context more closely matches the current context are more likely to be recalled.

Plugging Eqn. (2) into Eqn. (4), we can write the memory activation strength for item i at time step j as:

$$a_j^i = c_i^\top c_j = \rho^{(j-i)} = \exp(-\ln(1/\rho)(j-i)) \tag{5}$$

where the memory strength a_j^i is shown to follow an exponential function of the recency of an item $(j-i)$, with smaller ρ corresponding to larger exponential decay.

Power law forgetting function through multiple decay rates

Extending Eqn. (5) to the case where t is continuous, we can write the memory forgetting function as:

$$a(t) = \exp(-\lambda t) \tag{6}$$

where λ is the decay rate and t is the time passed since the encoding of the information.

Recent computational theories and neural data provide evidence that the brain constructs a scale-invariant representation of past events, maintaining temporal contexts that drift at different timescales (Howard et al., 2015; Shankar & Howard, 2012). This supports the use of a range of λ values, $\lambda_1, \lambda_2, \lambda_3, \ldots$. Under a range of λ values, the memory forgetting function can be written as a weighted combination of exponential decay functions:

$$a(t) = \sum w_k \exp(-\lambda_k t) \tag{7}$$

Since λ is continuous, we have:

$$a(t) = \int_0^\infty w(\lambda) \exp(-\lambda t) d\lambda \tag{8}$$

It is a well-known result that a mixture of exponential decays with a broad distribution of rates, as shown in $a(t)$, can approximate a power-law function. However, Bernstein's Theorem also states that any monotone function, whether it follows a power law or not, can be represented as a mixture of exponentials (Widder, 1941). This leaves open the question of why these weights are combined in this particular way and how they can support memory behavior when adapting to a new environment.

Our proposed framework: A resource-rational model of memory forgetting

In this section, we present a model to understand what gives rise to the shape of the memory forgetting function. Our model demonstrates how the need to remember information in the environment shapes the forgetting process while operating under a set of realistic cognitive constraints. Following the framework of resource-rational analysis (Lieder & Griffiths, 2020), we will first identify the cognitive constraints that shape memory forgetting, formalize the computational problem that the memory system faces, and then derive the optimal solution achievable under the identified cognitive constraints.

Identifying the cognitive constraints

Past work has identified the cognitive mechanisms that constrain the shape of the memory forgetting function (Bower, 1967; Howard & Kahana, 2002; Polyn et al., 2009), which we summarize below:

1. There is a range of decay rates λ.

2. Under the influence of each decay rate, memory strength decreases exponentially.

3. All weights sum up to 1, with the weighting taking place at encoding.

The first two constraints follow from cognitive mechanisms described in the Background section: memory strength decreases exponentially as a result of a drifting mental context (Howard and Kahana, 2002; Polyn et al., 2009; see derivations from Eqn. (1) to (5)); there exists a range of decay rates (see Eqn. (7)), supported by recent computational theories and neural evidence that the brain maintains temporal contexts that drift at different timescales (Howard et al., 2015; Shankar & Howard, 2012). Further, we justify here the third constraint why the weighting of exponential functions is expected to take place during encoding, with all weights summing to one. According to the TCM model, memory activation at time step j for item at time step i, a_j^i, not only depends on the amount of context drifting between study and test (see Eqn. (2)), but also on how much the item is initially associated with the context representation in the first place (see Eqn. (3)). If items failed to be associated with the context representation during encoding, even if the context representation is available as cues during recall, no memories can be recovered. When there is a single decay rate λ, as is the case with TCM, no weighting is required to be specified. However, under multiple context representations, each representing a different decay rate λ_k, there is a decision to be made at encoding regarding how much to associate the presented item with each context representation.

$$M_{k,i}^{CF} = M_{k,i-1}^{CF} + w_k f_i c_i^\top \tag{9}$$

where $M_{k,i}^{CF}$ stores the associations between items and the context representation with a decay rate of λ_k, and w_k is the strength of this association. Combining Eqn. (4) and Eqn. (9), and adding up the memory activation strengths across all context representations provides a mechanistic explanation for the weighting of exponential functions specified in Eqn. (7). The ability to encode an item with different context representations is not unbounded, as there are limited attentional and cognitive resources at memory encoding (Ma et al., 2024; Popov & Reder, 2020). Therefore, it is reasonable to assume that the total strength of associations at each moment adds up to a constant value of 1.

Formalizing the computational problem

We will now make some basic assumptions about the environment and how the need to retrieve certain information changes over time in this environment. Imagine an environment that is constantly changing. New information comes in, remains relevant in the environment for a while, until they are no longer relevant. Assuming that there is a hazard rate describing the proportion of currently relevant information or processes becoming irrelevant per time unit. If the hazard rate λ is a constant value for all t, following derivations in Sozou (1998), the probability that a process is still alive or relevant after time t has passed in an environment, referred to as the survival function, can be expressed as:

$$s(t) = \exp(-\lambda t) \tag{10}$$

We can also write down the probability density of the total survival duration (or the probability that a process fails right at time t):

$$f(t) = \lambda \exp(-\lambda t) \tag{11}$$

As the hazard rate, λ, describes the proportion of currently relevant information or processes that become irrelevant per time unit, it can be interpreted as the rate at which information decays in the environment. Large λ values correspond to fast-changing environments, where information becomes irrelevant quickly; and small λ values correspond to slow-changing environments where information remains relevant for a long time. The computational goal of the memory system is to match its forgetting function, representing memory availability, to the survival function of information in the environment, representing the probability that a memory will be needed in the environment. This adaptation would ensure that information is forgotten quickly when it becomes irrelevant, and retained longer when it remains useful.

We will show that the architectural constraints of the memory system are capable of supporting such an adaptation. Summarizing the set of cognitive constraints outlined in the last section (where λ is continuous), we have:

$$a(t) = \int_0^\infty w(\lambda) \exp(-\lambda t) d\lambda, \text{ where } \int_0^\infty w(\lambda) d\lambda = 1 \tag{12}$$

where $w(\lambda)$ was previously shown to be the strength of memory associations for each λ and $exp(-\lambda t)$ the memory activation at recall after time t passed for a given λ.

We can interpret $w(\lambda)$ as the memory system representing its belief about the λ value in the environment's survival function, which we will refer to it as $p(\lambda)$ moving forward to highlight its property as a probability density function. We can interpret $\exp(-\lambda t)$ as the memory system representing the survival functions of different λ in the environment. The memory system uses its current belief about how fast information decays in the environment to decide at which λ one should allocate more of its limited cognitive resources to form memory associations. The computational goal of the memory system is to update its belief about λ with observations X from the environment, so that the memory forgetting function $a(t)$ can accurately approximate the survival function $s(t)$ in the environment.

Deriving the optimal solution

We build a Bayesian model of memory forgetting to derive the optimal solution for matching the memory forgetting function, $a(t)$, with the need probability of memory in the environment, $s(t)$, after observing X from the environment. This adaptation operates under realistic cognitive constraints as outlined in the previous sections. We assume that $a(t)$ initially resembles a power law function, as is often observed in the memory literature. We then demonstrate how the model adapts to a new environment by inferring its characteristic hazard rate λ_0 through observing X. We consider two different scenarios describing observations X: a fully observable environment, where the status of every piece of information in the environment (i.e., relevant or irrelevant) is known at all times, and a partially observable environment, where this information is incomplete, limited only to that used to query memory. While our analyses primarily focus on a simplified case with a single hazard rate in the environment, the framework readily generalizes to more complex environments with a mixture of rates.

To allow the initial memory forgetting function $a(t)$ to resemble a power law function, a natural choice for $p(\lambda)$, which represents one's prior belief about environmental λ, is a gamma distribution with a shape parameter α and a rate parameter β:

$$p(\lambda) = \text{Gamma}(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda) \tag{13}$$

We can then obtain $a(t)$ through the Laplace transform of the gamma distribution, which is a standard result (Evans et al., 1993):

$$a(t) = \left(\frac{\beta}{\beta + t}\right)^\alpha \tag{14}$$

In the denominator $(\beta + t)$, the constant β becomes negligible when t is large, which gives a power law function $\beta^\alpha t^{-\alpha}$.

Fully observable environment. We first consider a fully observable environment. In this scenario, the memory system knows the complete status (i.e., alive or failed) of every process in the environment at all times. This includes a full history of each process's lifecycle: its start, its active duration, and its eventual failure. Using this perfect information, the memory system continuously updates its belief of the environment's hazard rate, $p(\lambda)$, in order to match the memory forgetting function $a(t)$ with the environmental survival function $s(t)$. While this scenario is an idealization, it may be plausible in contexts where inference is facilitated by rich perceptual information. For example, a person walking on a busy street develops an intuitive sense of how quickly surrounding objects, like cars or other pedestrians, are changing around them.

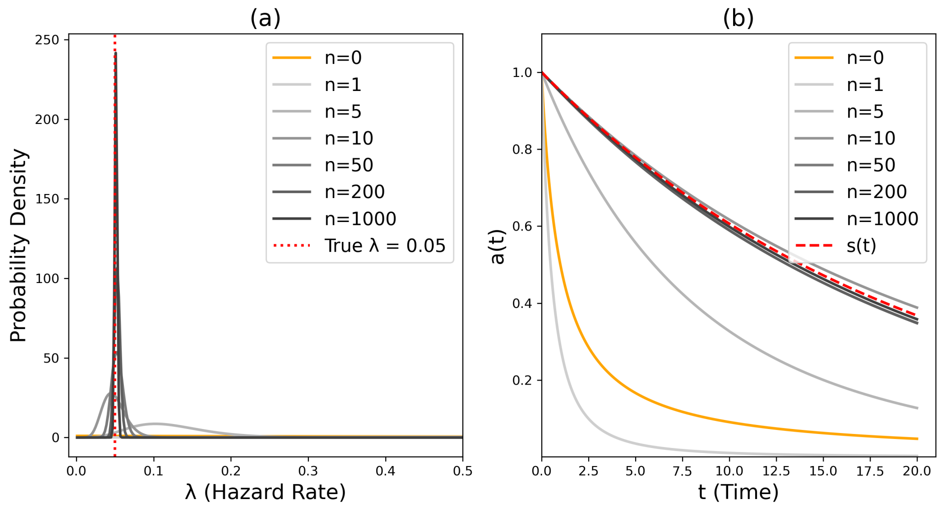

Figure 2. Adaptive memory forgetting in a fully observable environment. The memory system knows the complete status of every process in the environment. (a) As the number of observations increases, the memory system's belief about the λ value in the environment gradually converges to the true hazard rate λ_0. (b) Memory availability $a(t)$ converges to the environmental need probability specified in $s(t)$. Model parameters: $\alpha_{prior} = 1$, $\beta_{prior} = 1$, $n = 1000$, $\lambda_{true} = 0.05$.

In this environment, if the memory system observes that a process is still alive after t' has passed, its belief about the environmental hazard rate can be updated as

$$\begin{aligned} p(\lambda \mid t > t') &= \frac{p(t > t' \mid \lambda)p(\lambda)}{p(t > t')} \\ &\propto \exp(-\lambda t')\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\beta\lambda) \end{aligned} \tag{15}$$

where the probability of observing survival past time t' given a rate λ, $p(t > t' \mid \lambda)$, is the survival function $s(t')$ (see Eqn. (10)), and the prior belief $p(\lambda)$ is a Gamma distribution with a shape parameter α and a rate parameter β. Grouping the terms involving λ, the posterior distribution of λ is also a Gamma distribution:

$$\lambda \sim \mathrm{Gamma}(\lambda \mid \alpha, \beta + t') \tag{16}$$

So the memory availability $a(t)$ can be updated as

$$a(t) = \int_0^\infty p(\lambda \mid t > t') \exp(-\lambda t) d\lambda = \left(\frac{\beta + t}{\beta + t' + t} \right)^\alpha \tag{17}$$

If the memory system observes that a process fails right when t' has passed, its belief about the environmental hazard rate can be updated as

$$\begin{aligned} p(\lambda \mid t = t') &= \frac{p(t = t' \mid \lambda) p(\lambda)}{p(t = t')} \\ &\propto \lambda \exp(-\lambda t') \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha - 1} \exp(-\beta \lambda) \end{aligned} \tag{18}$$

where the probability of observing a failture time t' given a rate λ, $p(t = t' \mid \lambda)$, is $f(t')$ (see Eqn. (11)), and the prior belief $p(\lambda)$ is a Gamma distribution with a shape parameter α and a rate parameter β. Grouping the terms involving λ, the posterior distribution of λ is also a Gamma distribution:

$$\lambda \sim \text{Gamma}(\lambda \mid \alpha + 1, \beta + t') \tag{19}$$

So the memory forgetting function $a(t)$ can be updated as

$$a(t) = \int_0^\infty p(\lambda \mid t = t') \exp(-\lambda t) d\lambda = \left(\frac{\beta + t}{\beta + t' + t} \right)^{\alpha + 1} \tag{20}$$

As more data are observed, one's belief about the environmental λ will converge to the true underlying λ, and one's memory availability as specified in $a(t)$ will converge to the probability memory is needed in the environment specified in $s(t)$. As shown in Figure 2, we simulated the model to demonstrate this convergence. In the simulation, each observation was the complete history of a process generated from a true hazard rate, λ, in the environment. As the number of observations increased, the memory system's belief about the λ value in the environment gradually converged to the true hazard rate λ (Figure 2a); meanwhile, memory availability as specified in $a(t)$ converged to the environmental need probability specified in $s(t)$ (Figure 2b).

Partially observable environment. We next consider a partially observable environment, a more realistic scenario where the memory system has incomplete information. Here, the system cannot directly observe when a process fails. Instead, its knowledge is limited to observations sampled from the pool of currently surviving processes. This setup is closer to a typical laboratory memory experiment: an observer knows an item is still alive when it reappears but is never explicitly notified of its failure. We will show that even with such minimal information from the environment, the memory forgetting function $a(t)$ can still approximate the environmental survival function $s(t)$. We will formalize this model below. Intuitively, if a process appears frequently, it suggests a low environmental hazard rate λ, whereas a long absence of a process implies a high hazard rate. We will formalize this model below.

As illustrated in Figure 3a, the environment consists of a hazard rate λ and a set of M processes with failure times $\{\tau_1, \tau_2, \ldots, \tau_M\}$, which are hidden and unknown to the memory system. Each τ_i is a random variable that depends on λ. The model knows that each process i starts at a time u_i. The period of its total survival duration follows an exponential distribution, $p(\Delta t_i|\lambda) = \lambda exp(-\lambda \Delta t_i)$ (see Eqn. (11)). The failure time τ_i is therefore

$$\tau_i = u_i + \Delta t_i \tag{21}$$

At each time step t_k, an observation $x_k = j$ is made, where process j was sampled uniformly from the set of processes that were active and had not yet failed at time step t_k. Let $A(t) = \{i|u_i \leq t \text{ and } \tau_i > t\}$ be the hidden set of active, surviving processes at time t. The probability of observing process j at time t_k conditioned on a known set of $A(t_k)$ is:

$$p(x_k = j \mid \lambda, A(t_k)) = \begin{cases} \frac{1}{|A(t_k)|} & \text{if } j \in A(t_k) \\ 0 & \text{if } j \notin A(t_k) \end{cases} \tag{22}$$

The expected value of the population size, $\mathbb{E}[|A(t_k)|]$, is simply the sum of survival probabilities for all started processes, $\sum_{i \in \{m|u_m \leq t_k\}} a(t_k - u_i|\lambda)$. Marginalizing over the survival status of process j, the probability of observing process j at time t_k can be expressed as (see a detailed derivation in the Appendix):

$$p(x_k = j \mid \lambda) \approx \frac{a(t_k - u_j \mid \lambda)}{\sum_{i \in \{m|u_m \leq t_k\}} a(t_k - u_i \mid \lambda)} \tag{23}$$

To estimate the posterior distribution in this case, we use the particle filtering algorithm (Doucet et al., 2001). Each particle represents a hypothesis for λ, initially sampled from a prior distribution $\lambda \sim \text{Gamma}(\alpha_{piror}, \beta_{prior})$; we then calculate the likelihood for that specific hypothesis according to Eqn. (23). This likelihood is used in updating the weight of each particle. If an observation is unlikely under a particle's hypothesis, the particle will be down-weighted. As more data is observed, and by doing this for a large number of particles, the particles will converge on the true state of the system. See more details of the inference process in the Appendix. The final estimation of the hazard rate λ after k observations is the weighted average of the hazard rate of all N particles:

$$\mathbb{E}[\lambda] \approx \sum_{i=1}^{N} w_k^{(i)} \lambda^{(i)} \tag{24}$$

The final estimate of the survival function $a(t)$ is expressed as the weighted average of the survival functions of all N particles:

$$a(t) \approx \sum_{p=1}^{N} w_k^{(i)} a(t \mid \lambda^{(i)}) = \sum_{i=1}^{N} w_k^{(i)} \exp(-\lambda^{(i)} t) \tag{25}$$

By repeating this process for thousands of particles, it builds an empirical approximation of the true, intractable posterior distribution. If the environment consists of a mixture of λ values (governed by a gamma distribution with shape α and rate β) rather than a single λ value, the same estimation procedure applies but the estimated survival function for each particle $a(t \mid \lambda) = exp(-\lambda t)$ is replaced with $a(t \mid \alpha, \beta) = \left(\frac{\beta}{\beta+t}\right)^{\alpha}$, as the particles now represent hypotheses about α and β.

As shown in Figure 3, we simulated the model to demonstrate that one's belief about the environmental λ will converge to the true underlying λ. We randomly generated start times of 5000 processes uniformly from $[0, 1000]$, and generated their survival durations based on the true hazard rate λ in the environment (Figure 3a). At each time step t_k, one observation is made by sampling from one of the currently surviving processes. As the number of observations increased, the weighted averaged of the λ value among the particles gradually converged to the true hazard rate λ (Figure 3b); meanwhile, memory availability $a(t)$ estimated from the particles converged to the environmental need probability specified in $s(t)$ (Figure 3c).

Simulating the environments in R. B. Anderson et al. (1997) and Devraj et al. (2024)

Previous research has shown that changing the patterns of stimulus presentation in an environment directly alters the rate of memory forgetting (R. B. Anderson et al., 1997; Figure 1c) and impacts downstream tasks, like categorization, that rely on memory (Devraj et al., 2024; Figure 1d). R. B. Anderson et al. (1997) and Devraj et al. (2024) had similar manipulations of

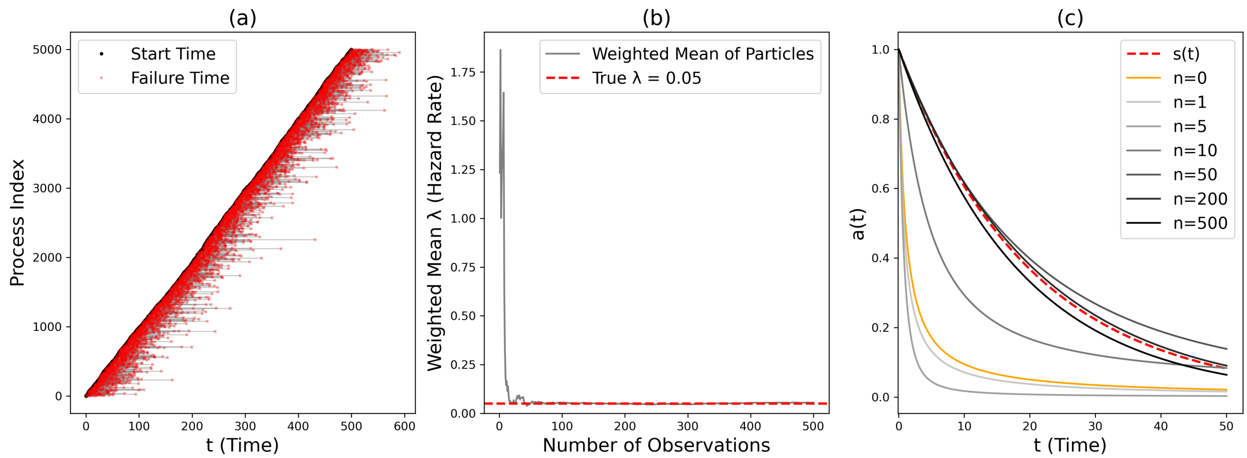

Figure 3. Adaptive memory forgetting in a partially observable environment. The memory system's knowledge is limited to observations sampled from the pool of currently surviving processes. (a) A visualization of start times and failure times of memory processes in the environment generated by the true hazard rate λ. (b) As the number of observed time steps increases, the weighted average of the λ value of the set of particles converges to the true hazard rate λ value. (c) Memory availability $a(t)$ converges to the environmental need probability specified in $s(t)$. Model parameters: $\alpha_{prior} = 1$, $\beta_{prior} = 1$, $n = 500$, $n_{particle} = 20000$, $n_{process} = 5000$, $\lambda_{true} = 0.05$.

environmental statistics: a flat condition, where the probability of needing information remains constant, and a downward or power-law condition, where information becomes less likely to be needed as time passes.

To examine the role of these environmental statistics in our model, we simulated a partially observable environment under two corresponding setups. The environmental λ is set to an extremely small value (Figure 4b, dashed line) to simulate the flat environmental survival function (Figure 4c, dashed line). A mixture of different λ (Figure 4e, dashed line) is used to simulate the power-law environmental survival function (Figure 4f, dashed line). Consistent with human data, where a flat environment leads to slower decay in memory and categorization performance (Figure 1c and d), our simulations show the memory forgetting function $a(t)$ gradually converging to the environmental survival function $s(t)$ in both conditions. As shown in Figure 4c and f, this convergence yields a slower forgetting rate in the flat condition than in the power-law condition.

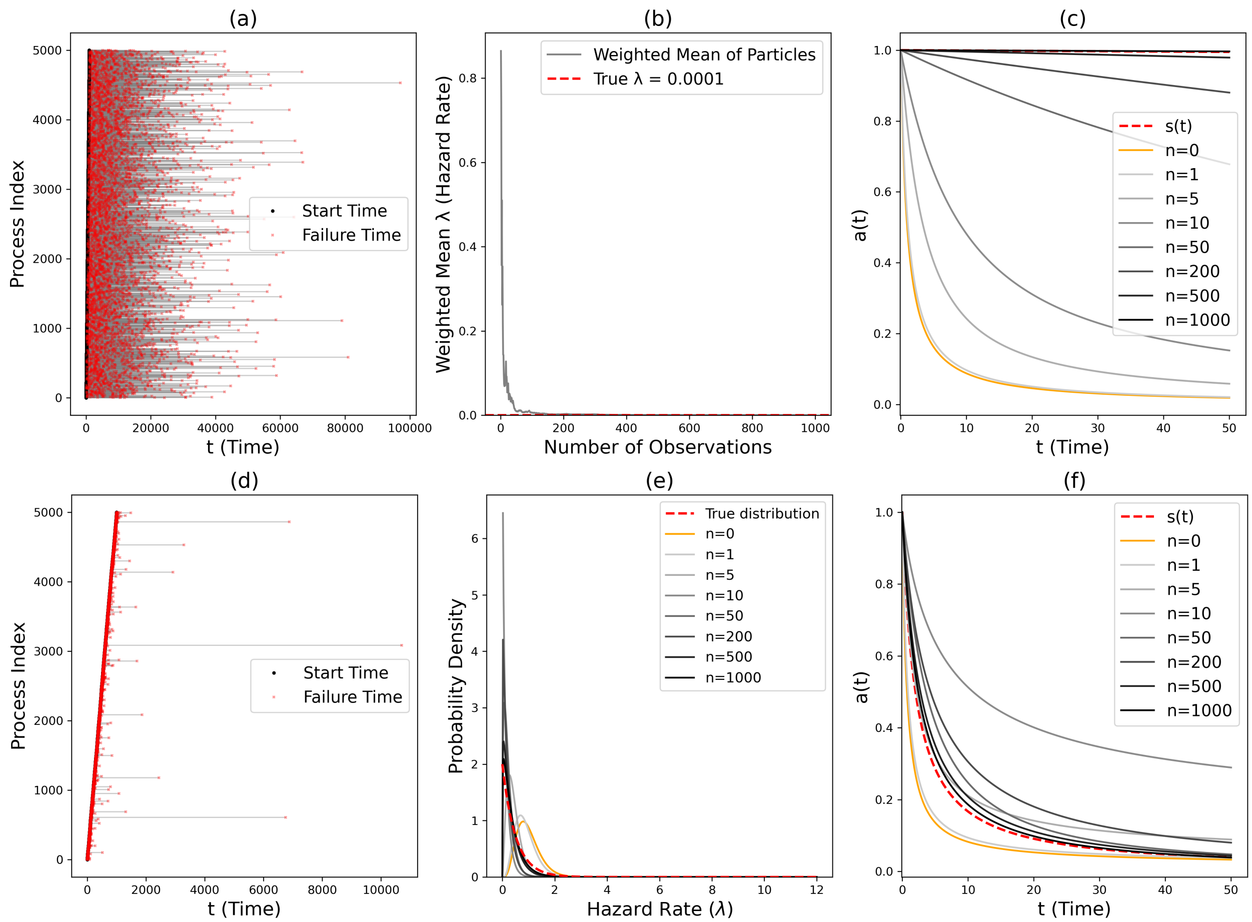

Figure 4. Simulating the environmental statistics in R. B. Anderson et al. (1997) and Devraj et al. (2024) in a partially observable environment. The memory system's knowledge is limited to observations sampled from the pool of currently surviving processes. Under the flat environmental statistics (a–c), the probability of information being needed remains constant (simulated by an extremely small λ value). Under the power-law environmental statistics (d–f), the probability decreases over time (simulated by a mixture of λ values in a gamma distribution governed by shape α and rate β). Start times and failure times of memory processes in the environment are visualized (a,c). As the number of observed time steps increases, the weighted average of the λ value of the set of particles converges to the true hazard rate λ under the flat environmental statistics (b); under the power-law environmental statistics, the belief of the distribution of λ converges to the correct mixture of λ (e). Memory availability $a(t)$ converges to the environmental need probability specified in $s(t)$ in both environments (c, f). Model parameters (a-c): $\lambda \sim \text{Gamma}(\alpha_{piror}, \beta_{prior})$, $\alpha_{prior} = 1$, $\beta_{prior} = 1$, $n = 1000$, $n_{particle} = 20000$, $n_{process} = 5000$, $\lambda_{true} = 0.0001$. Model parameters (d-f): $\alpha \sim \text{Uniform}(a_{prior}, b_{prior})$, $\beta \sim \text{Uniform}(a_{prior}, b_{prior})$, $a_{prior} = 0.1$, $b_{prior} = 10$, $n = 1000$, $n_{particle} = 20000$, $n_{process} = 5000$, $\alpha_{true} = 1$, $\beta_{true} = 2$.

The model achieves this by inferring the underlying λ value in the environment. When the environmental λ is close to zero, all memory processes remain needed in the environment for a long period of time (Figure 4a), and the system's belief of the hazard rate λ converges to zero over time (Figure 4b). This simulation is similar to the results in Figure 3 but addresses the boundary condition where there is no information decay in the environment. When there is a mixture of λ in the environment, which simulates a power-law environmental survival function, Figure 4e shows that the belief converges to the correct mixture of λ (modeled by a gamma distribution governed by shape α and rate β). While prior research has provided a rational account of findings in R. B. Anderson et al. (1997) and Devraj et al. (2024), suggesting that the memory forgetting function $a(t)$ should approximate the environmental need probability for adaptive purposes, here we provide a cognitively plausible model demonstrating how this adaptation occurs under known cognitive constraints. It is worth noting, however, that despite the alignment between the model and participant data regarding forgetting rates, the data do not converge to the exact environmental statistics the way the model does. Specifically, in the flat condition, the memory forgetting function decreases but does not become completely flat in human participants (see Figure 1c and d). This is a point that we will come back to in the General Discussion.

General Discussion

The decline of memory performance over time delay is one of the most studied patterns in human memory. In this work, we present a unified understanding of the power-law forgetting function by integrating two separate lines of research: one that views the power law as an optimal adaptation to the environment, and another that views it as a result of the memory system's immutable architecture. Through theoretical analysis and simulations, we demonstrate how the fixed structure of human memory can support adaptive behavior in new environments, showing that memory can be both immutable and adaptive. We assume that environments evolve at different timescales as information comes and goes. Our model proposes that the memory system updates its belief about this rate of change, aligning its own rate of forgetting to match that timescale. We demonstrate that under these assumptions, the memory forgetting function can closely approximate the environment's need probability, even under conditions with incomplete knowledge of the environment. Intuitively, the frequent reappearance of items suggests a slow-changing environment where information retains its relevance. In contrast, the prolonged absence of an item implies a rapidly changing environment where information quickly becomes obsolete. We show that the model captures empirical findings where forgetting rates vary depending on whether the environmental recency function is flat or follows a power law (R. B. Anderson et al., 1997; Devraj et al., 2024). We now turn to the broader implications of these results.

Previous research supports a rational account of memory forgetting, showing that statistical patterns in various environmental sources align with the power law of forgetting (J. R. Anderson & Schooler, 1991; J. R. Anderson et al., 2023). However, it remains an open question how the memory system achieves this adaptation, matching memory availability to the probability that information is needed in the environment. While this ability could be an evolutionary adaptation over a long timescale that may become part of the memory's fixed architecture (Hutchinson et al., 2025), evidence also shows that rational adaptation can occur much more rapidly as our memory system is exposed to a new environment (R. B. Anderson, 2001; Devraj et al., 2024). A recent study provided additional evidence for this, showing that an item's history of occurrences during the experiment can be used to accurately predict its recognition performance. When the environmental statistics were suddenly changed mid-experiment from artificial order to natural order, memory availability shifted accordingly (J. R. Anderson et al., 2026). While J. R. Anderson et al. (2026) used a statistical model to predict the probability of an item reappearing in the environment based on its history of occurrences, our current work presents a cognitively plausible model of how our memory system accomplishes this estimation.

The central assumption of our proposed model is that human memory forgetting adapts to its environment through updating its belief about the rate of information decay in the environment. Our assumption of how information becomes irrelevant in the environment is consistent with models of the environment in past work, where the probability of a reward being realized or an item's desirability in the environment decays exponentially over time (J. R. Anderson & Milson, 1989; J. R. Anderson et al., 2023; Burrell, 1985; Sozou, 1998). While a prior belief about the environmental decay rate – represented in the weighting of the exponential functions – may serve as a default strategy learned from experiencing a large number of environments with heterogeneous decay rates, the ability to update this belief allows for continuous adaptation to a new environment. From this perspective, both cognitive constraints (e.g., exponentially decaying memory traces as specified in mechanistic models of memory) and optimal solutions (e.g., updated beliefs about the environmental decay rate) can be seen as forms of rational adaptation operating on different timescales. While our memory system's basic design could be a product of long-term evolution by developing fixed structures or constraints that allow for alignment with exponentially decaying environments, its moment-to-moment performance is also governed by a faster and continuous adaptation to the statistics of a new environment, achieved by updating the weights of exponentials. Past work supports the idea that even fixed cognitive constraints could result from rational adaptation. For example, the rational allocation of mental effort is limited by cognitive control (Lieder et al., 2018; Musslick et al., 2015), yet this limited capacity is itself considered a rational adaptation to computational dilemmas in neural architectures (Musslick & Cohen, 2021). Similarly, Salvatore and Zhang (2025) found a convergence between context-based memory architectures and prominent neural networks used in machine translation. Despite the different forces driving these model developments (task performance in machine learning versus alignment with behavioral data in human memory), their convergence suggests that architectural constraints in human memory may serve an adaptive purpose. Therefore, a rational explanation is useful not only for understanding the flexibility of cognitive behavior under changing environments, but also for explaining cognitive constraints traditionally characterized by mechanistic models of cognition. Although it is useful to distinguish rational adaptation over relatively shorter timescales (the time to complete an experiment) and longer ones (developmental, lifetime, or evolutionary), we frame the present work as a synthesis of ideas from mechanistic and rational accounts of memory forgetting. We focus on unifying these accounts, rather than on behavioral adaptability over short or long timescales, because they directly emphasize the underlying cognitive constructs and mechanisms.

Environmental statistics shape not only memory itself but also the performance of downstream tasks that rely on memory, such as categorization (Devraj et al., 2024). Traditional categorization tasks often present stimuli with uniform frequency from various delays (Minda &

Smith, 2001). In our previous work, we introduced a power-law relationship between how recently a stimulus was seen and its probability of being needed again in the environment (Devraj et al., 2024). We found that this change reduced the accessibility of exemplar-based representations over time, aligning memories with the new environmental statistics. Consistent with this interpretation, we observed changes in categorization strategies, where there was increased reliance on prototype-based representations when the accessibility of exemplar-based representations was reduced. In a commentary on our results, Hutchinson et al. (2025) pointed out that a model using only exemplar-based representations can account for the data equally well, and we agree with the value of such a sparser model. However, we disagree with their claim that recency alone – how much time has passed since seeing a stimulus – is sufficient to explain categorization performance, without needing a rational explanation based on environmental statistics. To demonstrate their point, they fit an exemplar-based model with a power-law forgetting function. In their model, the accessibility of an exemplar is a direct readout from the mathematical relationship of the power law, rather than a result of memory adapting to the environment's need probability. While their model provides a mechanistic account for categorization performance in both environments, a rational account offers additional utility by explaining why memory forgetting differs across those environments. Just as a complete biological account of the heart requires understanding both the function it serves (pumping blood) and its internal mechanisms, a complete explanation of cognitive behavior benefits from synthesizing insights from both mechanistic and rational models (Chater & Oaksford, 1999; Marr, 2010), which is our goal here. As Kahana and Adler (2017) has put it, "The important question for theories of memory to address is how forgetting is affected by experimental manipulations and not what mathematical form the forgetting process assumes." Our present work proposes that understanding memory forgetting requires more than just characterizing the shape of the forgetting function. It demands an integrative theory that specifies the roles of both environmental goals and underlying cognitive mechanisms. Hutchinson et al. (2025) assumes that the power-law forgetting function arises from an immutable structure of human memory. In contrast, we demonstrate how memory can still adapt to environmental goals within the constraints of this structure. This adaptation places reasonable demands on the cognitive system, only requiring memory to align with the overall rate of change in the environment, without tracking the probability that each stimulus is needed in the environment individually. Although desirability can also vary across items in the environment, as previous models of the environment have assumed (J. R. Anderson & Milson, 1989; Burrell, 1985), we leave it to future work to understand how the memory system supports tracking this more detailed information.

Our theoretical analysis and simulations show that, given the constraints that memory decays exponentially at multiple rates, the memory system can effectively adjust memory availability by updating its beliefs about how quickly information decays in the environment. Power-law forgetting arises when the environment follows a power law. Power law forgetting can also arise when there is uncertainty regarding the environmental hazard rate, even in environments that do not follow a power law. This could explain why model and data align in their relative forgetting rates, yet the data in the zero-decay environment do not show a completely flat forgetting function in the same way as the model. Future empirical work can examine whether increasing the duration of the experiment will lead to better convergence, to the extent that the shape of the memory forgetting function no longer follows a power law. There may be other constraints, not incorporated in the current model, that set the limit of this convergence. For one, there could be an advantage to the most recent several items regardless of the environmental statistics, given what is held in the working memory. Additionally, imperfect memory tracking can bias estimations. If a reappearing item is mistakenly judged as new, the system starts a new memory process. This makes the inference of an extremely small hazard rate λ challenging and biases the estimated value upward, as such inference requires the knowledge that a process has lived for a very long time. Finally, if memory forgetting completely adapts to a zero-decay environment, there would be no context drift under the assumptions of the Temporal Context Model. While this could be seen as undesirable, the resulting state where all items are equally accessible and compete for retrieval may be a feature rather than a bug. This mirrors environmental demands where many processes are simultaneously active, decreasing the probability of any single process being selected. Similarly, although a lack of context drift makes it difficult to encode temporal order information, such information is arguably redundant in an environment where all items are equally likely to appear and randomly sampled at any given moment. Therefore, the absence of context drift in a zero-decay environment does not inherently constrain rational adaptation; rather, the observed discrepancy likely stems from a strong power-law prior and uncertainty in the system's belief about the new environment.

In summary, we have presented a resource-rational model of memory forgetting, showing how the adaptation of memory to different environments can be captured by assuming that forgetting is jointly shaped by computational demands in the environment and the brain's fixed architecture. This approach unifies two separate accounts of the power law of forgetting, rational and mechanistic, and provides a cognitively plausible mechanism of how the memory system makes information available based on how likely it is needed in a changing environment.

Acknowledgments

We would like to thank John Anderson for his feedback on a version of the draft, and Arjun Devraj for his assistance in producing Figure 1 in this article.

Open Practices Statement

This study is based on theoretical analyses and simulations only; it does not analyze specific human datasets and was not preregistered.

Declarations

Funding No funding was received for the preparation of this manuscript.

Conflicts of interest We have no known conflicts of interest or competing interests to disclose.

Ethics approval Not applicable – no participant data was collected.

Consent to participate Not applicable – no participant data was collected.

Consent for publication Not applicable – no participant data was collected.

Data availability Not applicable – no participant data was collected.

Code availability All materials and codes can be accessed at https://osf.io/g9fch/overview?view$_{o}nly = c55d1ef9eed9495f90a4939e86050c3f$.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Anderson, J. R., Betts, S., Byrne, M. D., Schooler, L. J., & Stanley, C. (2023). The environmental basis of memory. *Psychological Review*, *130*(5), 1137–1166.

Anderson, J. R., Betts, S., & Fincham, J. M. (2026). Using the environment to predict memory performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://pubmed.ncbi.nlm.nih.gov/41678255/

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408.

Anderson, R. B. (2001). The power law as an emergent property. *Memory & Cognition*, *29*(7), 1061–1068.

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, *25*(5), 724–730.

Anderson, R. B., Tweney, R. D., Rivardo, M., & Duncan, S. (1997). Need probability affects retention: A direct demonstration. *Memory & Cognition*, *25*(6), 867–872.

Bower, G. (1967). A multicomponent theory of the memory trace. *Psychology of Learning and Motivation*, *1*, 229–325.

Burrell, Q. L. (1985). A note on ageing in a library circulation model. *Journal of Documentation*, *41*(2), 100–115.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in cognitive sciences*, *3*(2), 57–65.

Devraj, A., Griffiths, T. L., & Zhang, Q. (2024). Reconciling categorization and memory via environmental statistics. *Psychonomic Bulletin & Review*, *31*(5), 2118–2136.

Donkin, C., & Nosofsky, R. M. (2012). A power-law model of psychological memory strength in short-and long-term recognition. *Psychological science*, *23*(6), 625–634.

Doucet, A., De Freitas, N., Gordon, N. J., et al. (2001). *Sequential Monte Carlo methods in practice*. Springer.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenues, Trans.). Teachers College, Columbia University.

Evans, M., Hastings, N., & Peacock, B. (1993). *Statistical distributions*. New York: Wiley.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331.

Howard, M. W. (2004). Scaling behavior in the temporal context model. *Journal of Mathematical Psychology*, *48*(4), 230–238.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Howard, M. W., Shankar, K. H., Aue, W. R., & Criss, A. H. (2015). A distributed representation of internal time. *Psychological review*, *122*(1), 24.

Hutchinson, D., Osth, A., & Little, D. (2025). Recency is sufficient for reconciling categorisation and memory: Commentary on devraj et al.(2024). *OSF*.

Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychology*, *71*(1), 107–138.

Kahana, M. J., & Adler, M. (2017). Note on the power law of forgetting. *bioRxiv*, 173765.

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.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Computational Biology*, *14*(4), e1006043.

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review*, *122*(2), 337–363.

Ma, S., Popov, V., & Zhang, Q. (2024). A neural index reflecting the amount of cognitive resources available during memory encoding: A model-based approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 350–370.

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Mensink, G.-J., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review*, *95*(4), 434–455.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775–799.

Mozer, M. C., Pashler, H., Cepeda, N. J., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. *Advances in Neural Information Processing Systems*, *22*.

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review*, *104*(4), 839–862.

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, *25*(9), 757–775.

Musslick, S., Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2015). A computational model of control allocation based on the expected value of control. *Proceedings of the Reinforcement Learning and Decision Making Conference*, *2015*, 7–10.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156. https://doi.org/10.1037/a0014420

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, *127*(1), 1–46.

Rubin, D. C., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1161–1176.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734–760.

Salvatore, N., & Zhang, Q. (2025). Sequence-to-sequence models with attention mechanistically map to the architecture of human memory search. *arXiv preprint arXiv:2506.17424*.

Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology*, *32*, 219–250.

Shankar, K. H., & Howard, M. W. (2012). A scale-invariant internal representation of time. *Neural Computation*, *24*(1), 134–193.

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(5), 460–471.

Sozou, P. D. (1998). On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *265*(1409), 2015–2020.

Tiganj, Z., Gershman, S. J., Sederberg, P. B., & Howard, M. W. (2019). Estimating scale-invariant future in continuous time. *Neural computation*, *31*(4), 681–709.

Widder, D. V. (1941). *The Laplace transform*. Princeton University Press.

Wixted, J. T., Carpenter, S. K., et al. (2007). The Wickelgren power law and the Ebbinghaus savings function. *Psychological Science*, *18*(2), 133–134.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, *2*(6), 409–415.

Appendix

Under a partially observable environment, the system cannot directly observe when a process fails, and its knowledge is limited to observations sampled from the pool of currently surviving processes. We employ the particle filter algorithm to infer the environmental hazard rate in this case. The objective is to approximate the posterior probability distribution of a static, hidden hazard rate λ based on a sequence of observations.

Model Formulation

The problem is defined by a process model, which describes the evolution of the hidden state, and a measurement model, which relates the hidden state to noisy observations. The hidden state of the system at time step k is the hazard rate λ. Since this parameter is assumed to be constant over time, the state vector is static:

$$h_k = \lambda \tag{26}$$

The process model describes the evolution of the state, $p(h_k|h_{k-1})$. As the state is static, the evolution is simply an identity function with no process noise:

$$h_k = h_{k-1} \tag{27}$$

The measurement model, $p(x_k|h_k)$, defines the probability of an observation given the state. At each time step k, we observe the index j of a process that is still alive, e.g., $x_k = j$. The probability of this measurement depends on the latent set of active processes at the observation time t_k, defined as $A(t_k) = \{i|s_i \leq t_k \text{ and } \tau_i > t_k\}$, where s_i and τ_i are the start and failure times of the process j. The total survival duration of a process follows an exponential distribution, $p(\Delta t_i|\lambda) = \lambda exp(-\lambda \Delta t_i)$ (see Eqn. (11)), the failure time τ_i is therefore $\tau_i = u_i + \Delta t_i$.

The probability of observing process j at t_k is then:

$$p(x_k = j|h_k, A(t_k)) = \begin{cases} \frac{1}{|A(t_k)|} & \text{if } j \in A(t_k) \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

Since $A(t_k)$ is latent, we calculate the probability of observing process j at time t_k by conditioning on the survival status of process j:

$$\begin{aligned} p(x_k = j \mid h_k) &= p(x_k = j \mid j \in A(t_k), h_k) p(j \in A(t_k) \mid h_k) \\ &\quad + p(x_k = j \mid j \notin A(t_k), h_k) p(j \notin A(t_k) \mid h_k) \\ &\approx \mathbb{E}[\frac{1}{|A(t_k)|}] a(t_k - u_j \mid h_k) + 0 \\ &\approx \frac{a(t_k - u_j \mid h_k)}{\sum_{i \in \{m \mid u_m \leq t_k\}} a(t_k - u_i \mid h_k)} \end{aligned} \tag{29}$$

where the first approximation is from $\mathbb{E}\left[XY\right] \approx \mathbb{E}[X]\mathbb{E}[Y]$ assuming that the survival status of a process j is independent of the population size $|A(t_k)|$, and the second one is from using the first-order Taylor approximation $\mathbb{E}\left[\frac{1}{X}\right] \approx \frac{1}{\mathbb{E}[X]}$. The expected value of the population size, $\mathbb{E}[|A(t_k)|]$, is the sum of survival probabilities for all started processes, $\sum_{i \in \{m|u_m \leq t_k\}} a(t_k - u_i|h_k)$. The survival function, estimated by memory availability $a(t)$, in Eqn. (29) is $a(t \mid \lambda) = exp(-\lambda t)$ when the environment consists of a single hazard rate λ. If the environment consists of a mixture of λ values (governed by a gamma distribution with shape α and rate β), the hidden state of the system is $h_k = \{\alpha, \beta\}$, and the estimated survival function is $a(t \mid \alpha, \beta) = \left(\frac{\beta}{\beta+t}\right)^{\alpha}$.

Particle Filter Algorithm

We approximate the posterior distribution of λ using the particle filter algorithm in the following steps.

Initialization For each particle $i = 1, \ldots, N$, the initial state $h_0^{(i)}$ (initial hypothesis of the hazard rate λ) is sampled from the prior distribution, a gamma distribution: $\lambda \sim \text{Gamma}(\alpha_{prior}, \beta_{prior})$. For each particle, failure times $\{\tau_1^{(i)}, \ldots, \tau_M^{(i)}\}$ are simulated based on its state hypothesis $\lambda^{(i)}$. Initial weights are set uniformly: $w_0^{(i)} = 1/N$. If the environment consists of a mixture of λ values (governed by a gamma distribution with shape α and rate β), the initial state $h_0^{(i)}$ is sampled from the prior distribution, a uniform distribution: $\alpha \sim \text{Uniform}(a_{prior}, b_{prior})$, $\beta \sim \text{Uniform}(a_{prior}, b_{prior})$.

Sequential Importance Sampling For each new observation x_k, the prediction step is simply $h_k^{(i)} = h_{k-1}^{(i)}$. The weight of each particle is updated by multiplying its previous weight by the measurement likelihood. The updated weights are then normalized. To evaluate the likelihood $p(x_k|h_k^{(i)})$, the set of active processes $A(t_k)$ is determined using the simulated failure times associated with particle i.

Resampling A resampling step is performed after each observation. A new set of $n_{particle}$ particles is drawn from the current population with replacement, where the probability of selecting particle i is proportional to its weight $w_k^{(i)}$. Following resampling, the weights of the new particle population are reset to a uniform value, $w_k^{(i)} = 1/N$. For each newly drawn particle, its full set of M failure times $\{\tau_1^{(i)}, \ldots, \tau_M^{(i)}\}$ is re-simulated based on its corresponding state value $\lambda^{(i)}$. Steps 2 and 3 are performed iteratively as subsequent observations become available.

After a total of k observations, the final posterior distribution of the hazard rate, $p(\lambda|x_{1:k})$, is represented empirically by the weighted collection of particles $\{(\lambda^{(i)}, w_k^{(i)})\}$. The expected value of any function $g(\lambda)$ can be approximated as:

$$\mathbb{E}[g(\lambda)] \approx \sum_{i=1}^{N} w_k^{(i)} g(\lambda^{(i)}) \tag{30}$$

The expected value of the posterior distribution of λ is estimated as the weighted average of λ of all particles:

$$\mathbb{E}[\lambda] \approx \sum_{i=1}^{N} w_k^{(i)} \lambda^{(i)} \tag{31}$$

The posterior predictive survival function is estimated by the memory system as the weighted average of the survival functions of all particles:

$$a(t) \approx \sum_{i=1}^{N} w_k^{(i)} \exp(-\lambda^{(i)} t) \tag{32}$$