scientific reports

OPEN

Stochastic consolidation of lifelong memory

Nimrod Shaham[1], Jay Chandra[1], Gabriel Kreiman[2] & Haim Sompolinsky[1,3]✉

Humans have the remarkable ability to continually store new memories, while maintaining old memories for a lifetime. How the brain avoids catastrophic forgetting of memories due to interference between encoded memories is an open problem in computational neuroscience. Here we present a model for continual learning in a recurrent neural network combining Hebbian learning, synaptic decay and a novel memory consolidation mechanism: memories undergo stochastic rehearsals with rates proportional to the memory's basin of attraction, causing self-amplified consolidation. This mechanism gives rise to memory lifetimes that extend much longer than the synaptic decay time, and retrieval probability of memories that gracefully decays with their age. The number of retrievable memories is proportional to a power of the number of neurons. Perturbations to the circuit model cause temporally-graded retrograde and anterograde deficits, mimicking observed memory impairments following neurological trauma.

Understanding the principles governing long-term memory is a major challenge in theoretical neuroscience. The brain is capable of storing information for the lifetime of the animal, while continually learning new information, so the brain must face the stability—plasticity dilemma: keep changing in order to learn new memories, but do so without erasing existing information. In humans, forgetting curves (retrieval probability vs. age of memory, sometimes referred to as retention curves), are found experimentally to be gracefully decaying with memory age, allowing for non-zero probability of retrieval for memories tens of years of age[1–4]. While retrieval probability curves monotonically decrease with memory age, the lifetime of individual memories is more intricate, and seemingly stochastic—we might not be able to retrieve a memory from last week, but can retrieve a much older one. Thus, the retrieval of memories does not depend on the memories' age alone.

Early attractor neural network models of long-term memory suffer from catastrophic forgetting: when the number of encoded memories is lower than a critical value, memories are retrievable with high precision, but when it is above that critical value, none of the memories can be retrieved[5–7]. Incorporating synaptic decay into the circuit enables continual learning, such that at any point in time recent memories are stable. However, the predicted forgetting curves exhibit a critical memory age, all memories newer than some age are almost perfectly retrievable, while all older ones are destroyed (a palimpsest behavior-old information is deleted in favor of new information)[8–13]. This is in contrast to the gracefully decaying forgetting curves in humans. Furthermore, the critical age is of the order of synaptic decay time, hence memories older than this time cannot be retrieved. Another class of memory models which avoid catastrophic forgetting are models with bounded (continuous or discrete) synaptic strengths[14,15]. These models also give rise to a palimpsest behavior qualitatively similar to synaptic decay models: only new memories up until a critical age are retrievable, while the older memories are not.

One of the main methods of studying the mechanisms of human memory is through memory disorders. Amnesic patients show a variety of patterns of forgetting. One is anterograde amnesia-reduced memory retrieval of events encoded after the onset of the disturbance to the circuit, presumably due to the inability to encode or store new memories. Another pattern is temporally-graded retrograde amnesia—when the probability of retrieval of memories encoded a short time before the pathology onset is lower than that of older events, giving rise to non-monotonic forgetting curves (an effect also known as Ribot's law). Retrograde amnesia is typically explained by invoking memory consolidation theory, suggesting that memories must go through a stabilization process that is disrupted by the proximal onset of the disturbance[16–22].

In addition to possible cellular mechanisms, memory consolidation at the system level is mediated through a rehearsal process—reactivating memories in wakefulness or during sleep[23–27]. Several computational models have been proposed for memory consolidation through rehearsals[28–36]. However, all reported results were confined to a small number of memories; none demonstrated memory functionality and forgetting curves in a

[1]Center for Brain Science, Harvard University, Cambridge, USA. [2]Harvard Medical School, Boston, USA. [3]Edmond and Lily Safra Center for Brain Sciences, the Hebrew University of Jerusalem, Jerusalem, Israel. ✉email: haim@fiz.huji.ac.il

large circuit with a number of retrievable memories scaling with the number of neurons. None of the models obtain the scaling of capacity and memory lifetime with the number of neurons and other intrinsic parameters.

Here we present a neural network model for lifelong continual learning and memory consolidation. Our model continuously stores patterns of activity by Hebbian learning, and combines synaptic decay with stochastic nonlinear reactivation of memories. Our model generates intricate and rich memory forgetting behavior. Retrieval probability curves decay smoothly with memory age (exponentially or even as a power law), with characteristic times that can be orders of magnitude longer than the synaptic decay time. In addition, due to the stochasticity of the consolidation process, there is a large variability in the survival of individual memories of the same age. We show that at any given time, the number of retrievable memories scales as a power of the number of neurons, exhibiting adequate memory functionality expected for a robust neuronal circuit with distributed memories. The power approaches unity for high rehearsal rate and the capacity approaches linear behavior with the number of neurons.

Perturbations of the model circuit give rise to complex patterns of memory deficits, temporally-graded retrograde and anterograde amnesia, the details of which depend on the size as well as the nature of the perturbation. Our theory relates global measures of memory functionality (memory capacity, characteristic memory lifetime) to intrinsic cellular and circuit parameters, such as synaptic decay rate and reactivation statistics, and provides new insight into how the brain builds and maintains the body of memories available for retrieval at each point in an animal's life.

Results

Model conceptual description. We model the lifelong memory acquisition and forgetting processes using a recurrent neural network, continuously experiencing Hebbian learning of new activity patterns ("memories"). Following initial memorization, memories are strengthened by stochastic rehearsal events, each increasing the Hebbian contribution of the rehearsed memory to the network synaptic connectivity matrix ('the memory efficacy'), thereby enhancing their retrievability. In addition, the synapses spontaneously degrade with time, mimicking the well-known synaptic turnover in biological neural networks. For simplicity, our model consists of a single neuronal population, and every synapse is allowed to be positive or negative. The retrieval of a memory at a certain moment in time is determined by both its current efficacy as well as its random interference with other stored memories. The rate of rehearsals of a given memory is self-consistently determined by its retrievability, such that a more easily retrievable memory will be revisited more frequently than a less retrievable one, and a memory that loses its retrievability will not be rehearsed anymore. In this framework, memories begin their life with a fixed initial efficacy, which is subsequently increased by each rehearsal, and decays between rehearsals. As long as a memory is revisited frequently enough, its amplitude will be larger than the threshold efficacy required for retrieval. However, due to the stochastic nature of the rehearsals process, there will be a period of time eventually when the time lag between two rehearsals will be too long, such that the efficacy of the memory will drop too low, making the memory irretrievable, and therefore forgotten. The age at which a certain memory will be forgotten is therefore stochastic and may range from the order of the synaptic decay timescale to several orders of magnitude longer. In the following sections we give a detailed description of our memory model, present its properties and explain how these properties emerge from our model's dynamics.

Model details. Our model is based on the sparse version of the Hopfield attractor network model of associative memory[5,6]. Memories are sparse[7,37–41], uncorrelated N-dimensional binary activation patterns (N is the number of neurons) and are stored as fixed points of a recurrent neural network dynamics with binary neurons. We assume that the neural activation threshold is adjusted dynamically so that the population activity level maintains the same sparsity as the memories (see "Methods", in "The effects of structural perturbations on memory function" this assumption is modified). Synaptic dynamics are governed by three processes: Deterministic exponential synaptic decay[8,11] (first term in Eq. (1)), Hebbian learning[42] of new memories (second term in Eq. (1)), and Hebbian consolidation of old memories following their reactivation (third term in Eq. (1)),

$$J_{ij}(t + \Delta t) = (1 - \Delta t/\tau)J_{ij}(t) + \sum_l \xi_i^l \xi_j^l \, \delta_{t,l} + b \sum_{k,\{t_k\}} \xi_i^k \xi_j^k \, \delta_{t,t_k} \tag{1}$$

Here $J_{ij}(t)$ is the strength of the synapse between neurons i and j at time t (symmetric in i and j), ξ_i^l is the i-th element of the memory introduced first at time $t = l$ and it is given by:

$$\xi_i^l = \begin{cases} \frac{1-f}{\sqrt{Nf(1-f)}} & \text{with prob. } f \\ -\frac{f}{\sqrt{Nf(1-f)}} & \text{with prob. } 1-f \end{cases} \tag{2}$$

Here f is the fraction of neurons active in a memory state (the sparseness level). According to the above equation, new memories enter in each time interval Δt and synapses decay at a rate $1/\tau$, representing the finite lifetime of synapses[43]. The last term represents a Hebbian strengthening of old memories following a sequence of reactivation events that occur for memory k at times denoted by t_k (which will be specified below). The factor b denotes the size of synaptic modification due to a single consolidation event of an old memory, assumed to be smaller than the Hebbian amplitude of learning a new memory (i.e. $b < 1$). The resulting connectivity matrix can be written as:

$$J_{ij}(t) = \sum_l A_l(t)\xi_i^l \xi_j^l \tag{3}$$

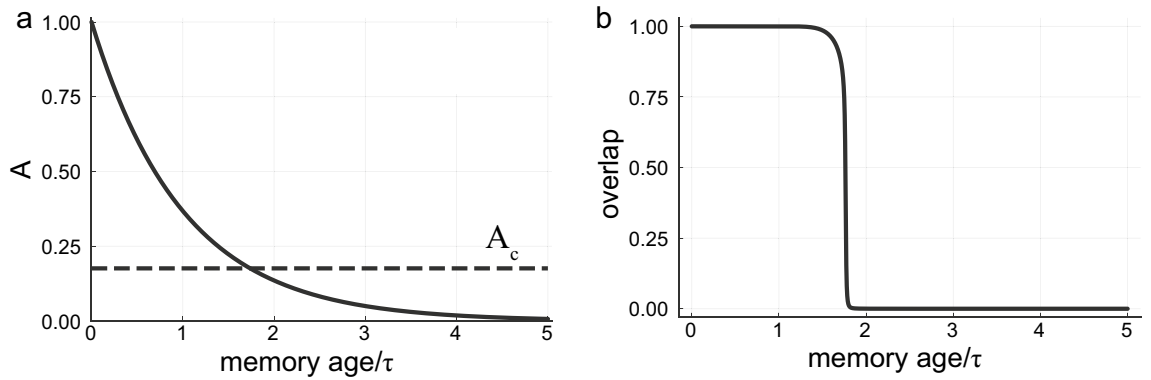**Figure 1.** (**a**) Pure forgetting. A memory efficacy trajectory as a function of time (solid line). The critical efficacy A_c is plotted as a dashed line. (**b**) Overlap of the network state with a memory state as a function of the memory age. The overlap is a measure of memory retrievability—after initializing the network near a memory state, the overlap of the nearby attractor network activity will be close to unity for retrievable memories and small compared to one for irretrievable memories. Here $N = 8000$, $f = 0.01$, $\tau = 2240$. The catastrophic age here is $\sim 1.73\tau$, resulting in a capacity (number of retrievable memories) of $0.5N$. Note the very large value of τ needed to support this capacity—this will be addressed in later sections.

$A_l(t)$ is the *efficacy* of memory l at time t. The ability to recall a memory depends on the level of noise, which originates from random interference with other memories. Its variance is proportional to the sum of the squares of all efficacies (see "Methods"):

$$\Delta^2(t) = \frac{f}{N} \sum_n A_n^2(t) \tag{4}$$

We define the critical efficacy A_c, as the efficacy for which a memory pattern loses its stability. A_c is proportional to the interference noise Δ, with the proportionality constant depending only on the sparseness:

$$A_c = a(f) \cdot \Delta \tag{5}$$

Due to the reduction of overlap between memories, The critical efficacy is increased when f is decreased. The factor $a(f)$ can be approximated by (see Supplementary Information (SI) section 1):

$$a(f) \approx 1.44 \sqrt{2 \log\left(\frac{1.9}{f}\right)} \tag{6}$$

This approximation holds for the f regime we consider throughout this study. For $f = 0.01$ (which we will use throughout the paper), $a \approx 4.7$.

Pure forgetting. Without rehearsals, our model is similar to previous models of associative memory with forgetting[8,10,11], in which memory efficacies decay exponentially with age, $A_l(t) = \exp(-(t - l)/\tau)$ (Fig. 1a). Using Eq. (4), the interference noise equals $\Delta^2 \approx f\tau/(2N)$. For $\tau > \tau_0$, where $\tau_0 = 2N/(fa^2(f))$ (see Eq. (5)), A_c increases above unity (the initial efficacy) and no memory will be retrievable. This *global catastrophic forgetting* is similar to the behavior of the Hopfield model after reaching memory capacity, where the interference effect is too strong and all memory states lose their stability[5,6]. If $\tau < \tau_0$, recent memories are retrievable, while memories older than a critical age $t_0 = \frac{\tau}{2} \log\left(\frac{\tau_0}{\tau}\right)$ are forgotten (Fig. 1b). Thus, for short decay times, this model allows for continual learning of recent memories without global catastrophic forgetting. However, it predicts an unrealistic *age-dependent catastrophic forgetting*, where all memories up to a critical age are almost perfectly retrievable, and all older memories are completely forgotten. This sharp transition happens despite the graceful exponential decay of efficacies with age, and results from the collective effects of memory stability in the network.

In what follows we will show that when stochastic rehearsals are taken into account, the behavior changes dramatically, generating more realistic memory forgetting trajectories and allowing for lifelong memories.

Nonlinear stochastic reactivation. To specify the statistics of reactivations, we revert to the continuous time version of Eq. (1) which yields the efficacies dynamics:

$$\frac{dA_l}{dt} = -\frac{1}{\tau} A_l + b R_l(t) \tag{7}$$

With $A_l(t) = 0$ for $t < l$ and $A_l(l) = 1$. The reactivations are modeled as a point process
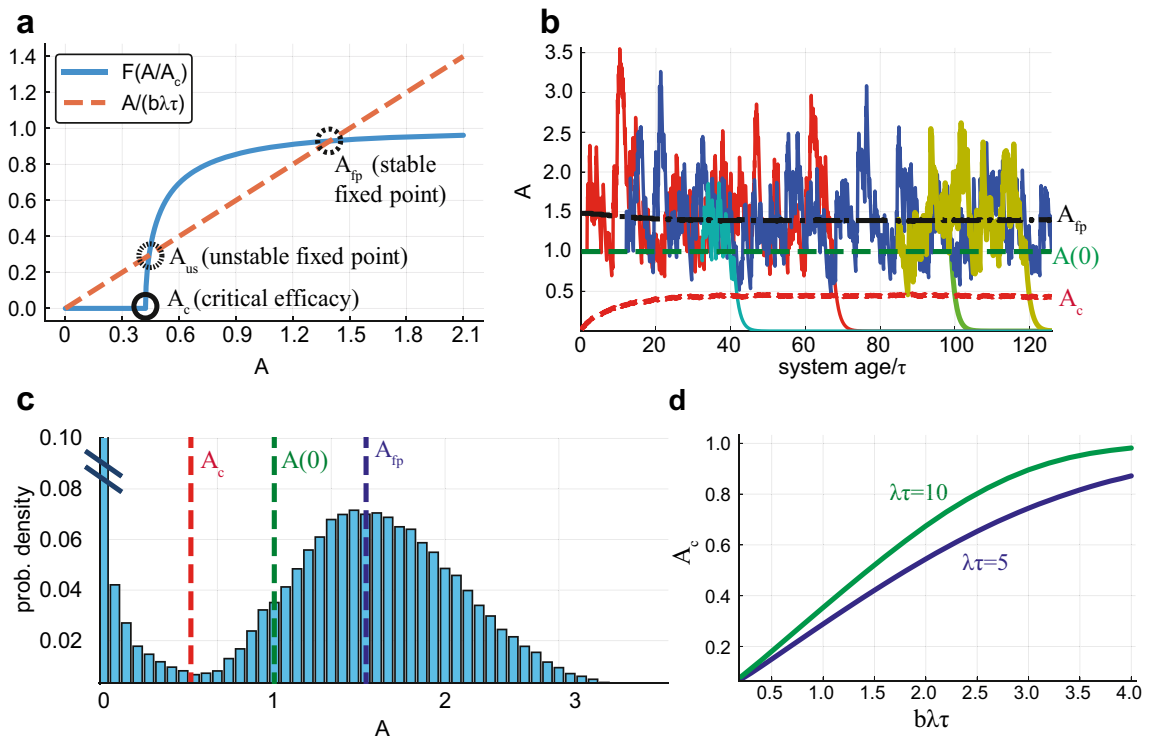
$$R_l(t) = \sum_{\{t_l\}} \delta(t - t_l), \tag{8}$$

Figure 2. Stochastic memory dynamics. (**a**) Blue: basin of attraction size F as a function of memory efficacy A. Orange dashed: A/τ, the negative of the deterministic decay term in Eq. (7). Importantly, F is zero for $A < A_c$. Here $\lambda\tau = 5$, $A_c = 0.4$. (**b**) Example memory efficacies vs. age of the system. Memories enter with efficacy $A(0) = 1$, rehearsal efficacy $b = 0.3$. Most of them increase towards $A_{fp} \approx b\lambda\tau \approx 1.5$, and fluctuate around it. Large enough fluctuations can take efficacies below A_c (e.g, cyan curve at $age/\tau \approx 40$, yellow curve at $age/\tau \approx 120$). Some memories are alive for a very short time (e.g., green curve) and some for very long (e.g., red, blue curves). (**c**) Distribution of memory efficacies after saturation of A_c. (**d**) Equilibrium values of A_c as a function of $b\lambda\tau$ for different $\lambda\tau$ values. Here $\tau = 160$, $N = 8000$.

where t_l are the times at which memory l was rehearsed. To specify the rate of the reactivation process, we hypothesize that this process is more likely to yield a Hebbian strengthening of memories with a large basin of attraction. The rationale is that during reactivation periods, the system is more likely to visit memories with large basins of attraction and stay there for a significant period of time triggering their Hebbian strengthening . In particular, memories that at some point in time lost their stability and are not attractors of the dynamics (i.e., have vanishing basin of attraction) will not be reactivated, will experience fast pure decay, and will be forgotten. Hence we model reactivation events as inhomogeneous Poisson processes, with mean rate $r_l(t) \equiv \langle R_l(t) \rangle$, which is proportional to the memory's basin of attraction size:

$$r_l(t) = \lambda F(A_l(t)/A_c(t)) \tag{9}$$

where λ denotes the maximal reactivation rate. As in Eqs. (4) and (5), at all times $A_c(t) = a(f) \cdot \Delta(t)$. The nonlinear function F denotes the size of the basin of a memory and depends on the ratio of the memory efficacy over the critical capacity A_c (Fig. 2a). We derive the function F by numerical calculation of memories' basin of attraction size for different values of A/A_c (see "Methods" section and SI for details). At any given time, only memories with non-zero basin size (i.e., $A_l(t) > A_c \rightarrow F > 0$) are retrievable and might be reactivated. Note that since the interference $\Delta(t)$ depends on the efficacies of all memories (Eq. (4)), the reactivation rates of all memories are coupled in Eq. (9) via A_c.

The approach to steady state of memory consolidation. It is useful to first consider the average dynamics, replacing the reactivation point process by its mean rate, Eq. (9). For a given A_c, the resulting self-consistent equation for the steady state efficacies,

$$A_{fp} = b\lambda\tau F(A_{fp}/A_c), \tag{10}$$

possesses two stable fixed points: one at zero and another one when the two competing processes, decay and reactivation, balance each other (A_{fp}, Fig. 2a). Due to the rapid saturation of the function F, for most of the parameter regime $A_{fp} \sim b\lambda\tau$.

To fully understand the system's behavior we need to consider the dynamics of A_c itself as well as the stochastic nature of the process. Initially, when the first memories enter the system, $A_c \propto \Delta$ is very small and the memory efficacies consolidate around the value $A_{fp} \sim b\lambda\tau$. As more memories are encoded, the interference grows and so
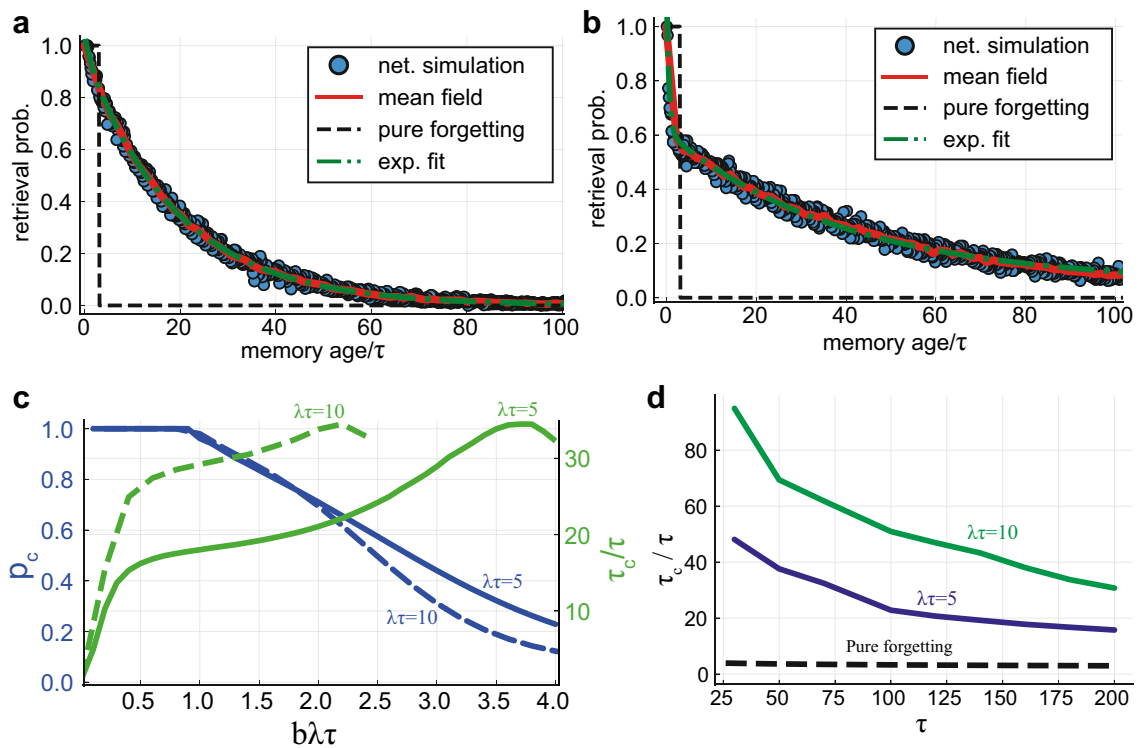
Figure 3. (**a**, **b**) The forgetting curve. The probability of retrieval as a function of memory age. Blue dots: full network simulation results (see "Methods"). Red solid lines: results of a mean field approximation ("Methods"). An exponential fit with characteristic decay time $\approx 18\tau$ is shown in green (dash-dot line) in (**a**), and a double exponential fit with characteristic decay times $\approx \tau$ and $\approx 38\tau$ in (**b**). The retrieval probability for pure forgetting is shown in black (dashed line). In (**a**) $N = 8000$, $\tau = 160$, $\lambda\tau = 5$, $b = 0.3$. In (**b**) same parameters as (**a**) except $b = 0.25$, $\lambda\tau = 10$. (**c**) Blue (left y axis): Consolidation probability vs. $b\lambda\tau$ for different $\lambda\tau$ values. Green (right y axis): Consolidation time τ_c normalized by synaptic decay time τ vs. $b\lambda\tau$ for different $\lambda\tau$ values. Here $\tau = 160$. (**d**) Consolidation time τ_c normalized by synaptic decay time τ vs. τ for different $\lambda\tau$ values. Blue curve: $\lambda\tau = 5$, $b = 0.3$. Green curve: $\lambda\tau = 10$, $b = 0.25$.

does the critical efficacy (red line in Fig. 2b). When the critical efficacy is large enough, fluctuations in reactivation times lead some memory efficacies to drop below A_c, making these memories irretrievable. A steady state is achieved when the flux of memories arriving at the system and consolidated is balanced by the rate of memories forgetting due to the drop of their efficacy below A_c. At this stage, A_c reaches a fixed equilibrium value and so does the mean number of retrievable memories. The specific identity of the retrievable memories varies with time—some are forgotten while new ones are being consolidated. The distribution of efficacies at equilibrium (Fig. 2c) consists of two modes: The first is the contribution of the forgotten memories, below A_c, which diverges at small A as $p(A) = \tau/A$. The second, above A_c, is a mode around A_{fp} representing the retrievable memories.

A_c increases as the amplitude b and number of reactivations per decay timescale $\lambda\tau$ increase, due to increased interference (Fig. 2d). For moderate reactivation strength, A_c is well below both the encoding strength $A(0) = 1$ and the consolidation fixed point as seen in the examples in Fig. 2b,c. As reactivation strength grows, A_c increases and approaches 1, affecting adversely the consolidation process, as will be seen in the next section.

The forgetting curve. Importantly, in our model, the time of forgetting of memories at a given age is highly variable, ranging from a fraction of the decay time τ (for unfortunate memories that weren't rehearsed fast enough after learning), and up to hundreds of τ for well-rehearsed memories (Fig. 2b). Nevertheless, on average, memory retrievability decreases with memory age, and this is captured by the forgetting curve—the probability of retrieving a memory as a function of its age, after a steady state has been reached (Fig. 3a,b). This curve exhibits an exponential tail with a long time constant, denoted as the consolidation time τ_c—a direct result of the long time required for a large fluctuation in reactivation rates to form such that consolidated efficacies decrease from around A_{fp} to A_c (SI). The consolidation time enhancement factor τ_c/τ can reach several orders of magnitude, allowing memories to survive for very long times compared to the intrinsic timescales of the system (as shown in Fig. 3c,d). For fixed consolidation parameters $\lambda\tau$ and b, consolidation time normalized by τ decreases with τ due to increased interference (Fig. 3d).

For fixed τ, consolidation time increases sharply with $\lambda\tau$ and b (Fig. 3c). However, increasing these parameters may adversely affect the consolidation process. When $b\lambda\tau$ is of order 1, most of the memories experience consolidation, as is the case in Fig. 3a. However when $b\lambda\tau \gg 1$, A_c is close to the encoding efficacy (Fig. 2d). This causes a significant number of memories not to get consolidated. Therefore, the forgetting curve exhibits an initial
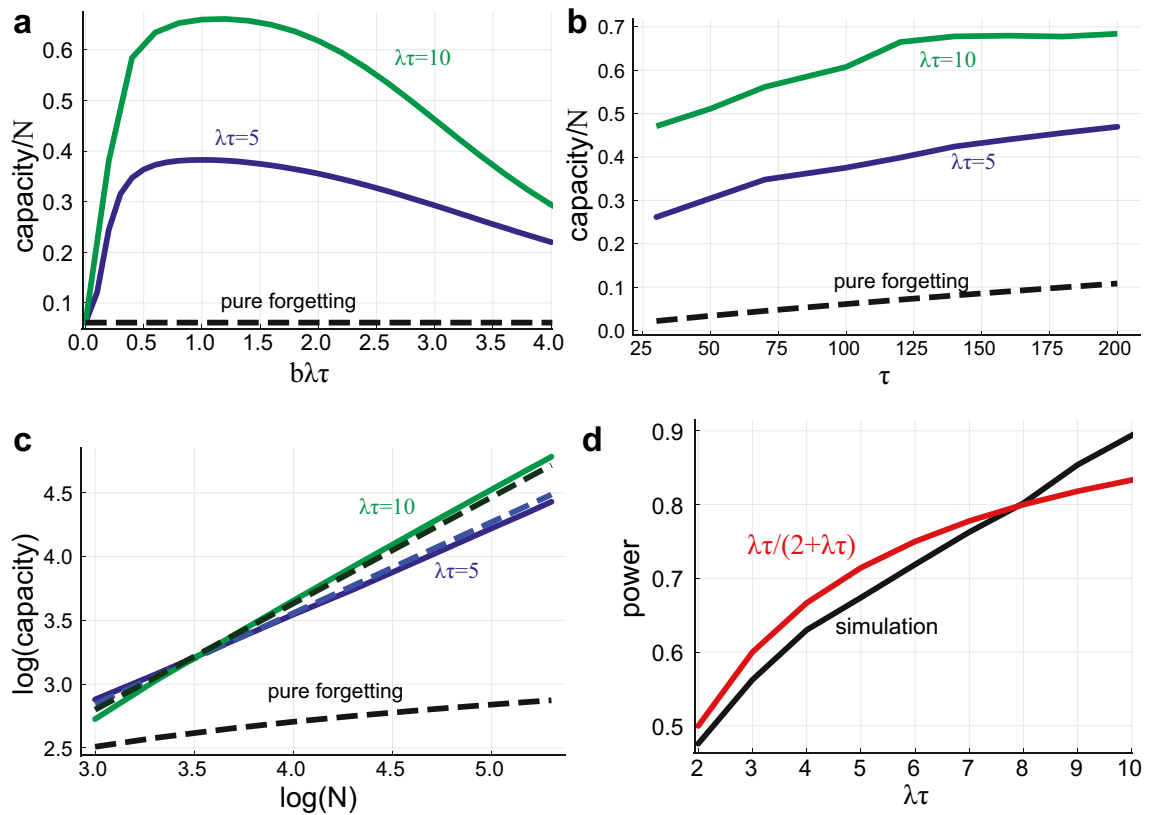
Figure 4. Memory capacity. (**a**) The number of retrievable memories divided by N as a function of $b\lambda\tau$ for different average number of rehearsals per characteristic decay time ($\lambda\tau$) values. The dashed line shows the capacity of the pure forgetting model. Here $N = 8000$, $\tau = 160$. (**b**) The number of retrievable memories divided by N as a function of τ for different $\lambda\tau$ values. (**c**) Capacity vs. N (logarithmic, base 10), solid lines show simulation results, dashed lines are analytical approximation. Here $\tau = 160$, $b = 0.3$. (**d**) The power of N vs. $\lambda\tau$ (black), and the analytical approximation (red). Here $\tau = 160$, $b = 0.3$.

fast decay with a characteristic time τ, in addition to the slow decay time τ_c (Fig. 3b). To quantify this effect, we measure the *consolidation probability* of memories, p_c, defined as the chance of a memory efficacy to reach A_{fp}, and therefore become part of long-lived memories. p_c decreases as a function of the reactivation strength from $p_c = 1$, for $b\lambda\tau < 1$, to zero for strong reactivation (Fig. 3d). The consolidation probability p_c together with τ_c are the key consolidation parameters.

Capacity increases as power law with network size. We define the network's memory capacity as the number of memories retrievable (memories with $A > A_c$) in the equilibrium phase after long encoding time (Fig. 4). The capacity can be evaluated as the area under the forgetting curve. Hence, it can be approximated as $(1 - p_c)\tau + p_c\tau_c$, where $p_c\tau_c$ is the contribution from consolidated memories and fraction of memories that are consolidated and the first term is the contribution from unconsolidated memories. As seen previously, τ_c increases with $b\lambda\tau$, while p_c decreases with $b\lambda\tau$—less memories are consolidated, but the consolidated ones live longer. Maximal capacity is achieved when $b\lambda\tau \approx A(0) = 1$, which is the maximal value that allows for 100% of the memories to get consolidated.

To assess the efficiency of information storage in the network it is important to evaluate the dependence of the capacity on the network size, N. In previous 'pure forgetting' models[8,10] the synaptic decay time was assumed to scale linearly with N, resulting in memory capacity t_0 which is proportional to N. The same holds for our model. However, this scaling results in extremely large, biologically implausible, synaptic decay times for large networks. Here we assume that τ is a property of individual synapses and is independent of network size. Under this condition, the capacity in the pure forgetting model increases only logarithmically with N, Fig. 4c.

Interestingly, we find that in our model, the capacity scales as a power law of the number of neurons, with a power that approaches unity for large $\lambda\tau$ values (Fig. 4c,d). To approximate the power analytically (for the parameter range where $p_c \approx 1$), we first approximate A_c, assuming that the main contribution to the interference noise comes from consolidated, retrievable memories (SI, sec.3):

$$A_c \approx \sqrt{\frac{f \cdot a^2(f)p_c\tau_c}{N}}\, b\lambda\tau \tag{11}$$

Now, under the assumption that $p_c \approx 1$ and that the mean rehearsal rate is $\lambda\tau$, we get (SI):
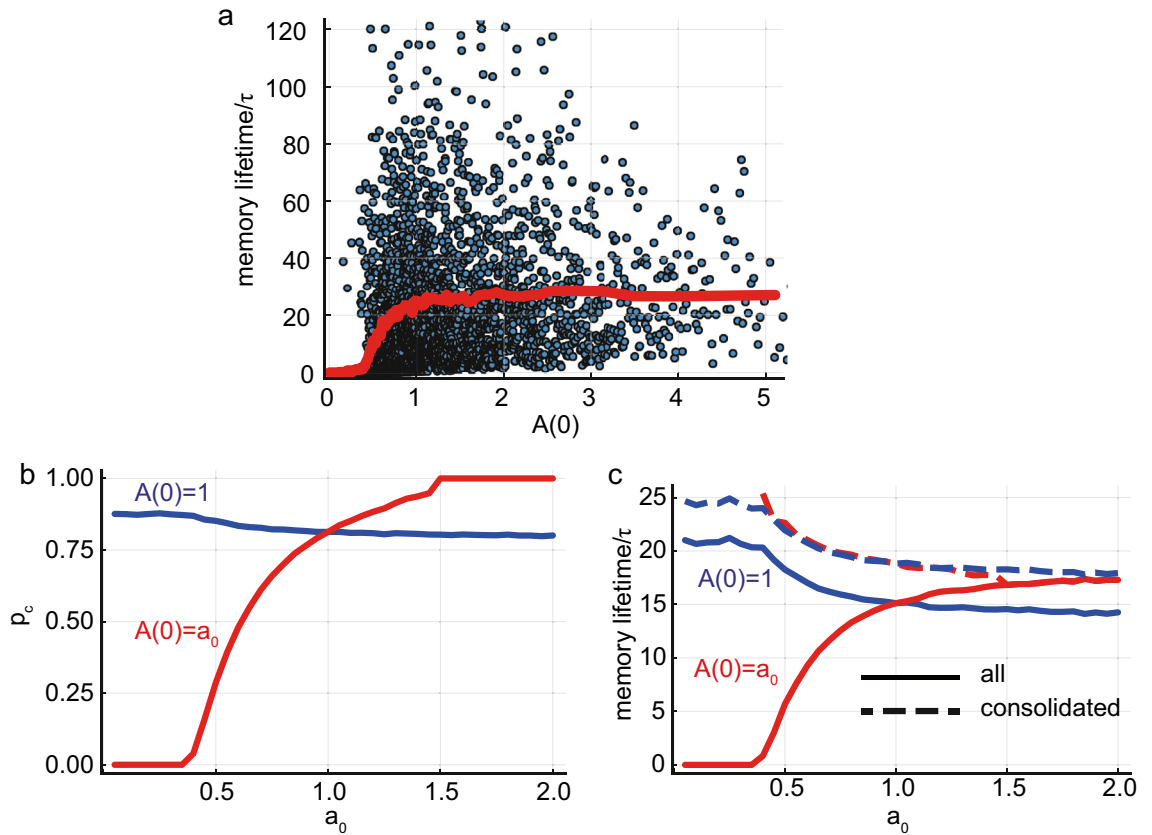
Figure 5. Initial condition distribution. (**a**) Here $A(0)$ for each memory is drawn from an exponential distribution with unity mean. Blue points are values for single memories, and the red line shows the mean. Note that the spread in lifetimes at each encoding strength is a result of the stochastic rehearsal process, which yields an exponential distribution of lifetimes and is present also in the uniform $A(0)$ case. (**b**) Consolidation probability vs. a_0, which is the initial efficacy of half of the memories (the others have $A(0) = 1$). Values for memories introduced with $A(0) = 1$ are shown in blue, and for memories introduced with $A(0) = a_0$ are shown in red. (**c**) Memory mean lifetime (time from insertion to forgetting) as a function of a_0. Same scenario and coloring as in (**a**). Dashed lines are averaged lifetimes of consolidated memories only. Parameters: $N = 8000$, $f = 0.01$, $\lambda\tau = 5$, $b = 0.3$

$$\text{capacity} \approx \tau_c \approx \tau^{\frac{1}{1+0.5\lambda\tau}} \left(\frac{N}{fa^2(f)} \right)^{\frac{\lambda\tau}{2+\lambda\tau}} \tag{12}$$

Figure 4c,d show the approximation gives a reasonable fit to the dependence of the capacity on N. Note the significant increase in capacity compared to the pure forgetting model.

Inhomogeneity in initial memory encoding. So far we have assumed that all memories are encoded initially by Hebbian plasticity with the same amplitude $A(0) = 1$ (eq. 1). In reality, memories might differ in their encoding strength, for instance, due to factors such as attention, or emotional context. Thus, it is important to explore the effect of a distribution of initial encoding strengths. As long as most of the initial efficacies are in the neighborhood of $b\lambda\tau$, the global memory properties such as A_c, forgetting curve, and capacity are not affected drastically. However, individual memories with initial efficacy below A_c are forgotten, while memories with $A(0)$ larger than $b\lambda\tau$ have slightly enhanced consolidation properties, as is confirmed in Fig. 5a for an exponential distribution of $A(0)$ with mean 1.

To better elucidate the effect of inhomogeneity in $A(0)$, we consider in Fig. 5b,c the case of a Bernoulli distribution, $A(0) \in \{1, a_0\}$ with equal probability. For small a_0 compared to $b\lambda\tau = 1.5$, the consolidation probability for memories with $A(0) = a_0$ decreases drastically and vanishes for a_0 below $A_c \approx 0.39$ (Fig. 5b). When a_0 increases above 1, consolidation probability of these memories increases until it reaches 1 for $a_0 \approx b\lambda\tau$. On the other hand, memories with $A(0) = 1$ are only moderately affected by changing a_0. The mean lifetime of memories with $A(0) = a_0 < 1$ drops considerably (Fig. 5a,c). This is, however, due to averaging the lifetime of all memories including those that did not consolidate. Importantly, in our model, memories with small a_0 that did reach the neighborhood of the fixed point have the same long lifetime as other consolidated memories, independent of the original encoding strengths as shown by the dashed lines in Fig. 5c. Note that inhomogeneous initial efficacy distribution alone will not give rise to memories with lifetime much larger than the synaptic decay time scale,

7

due to the exponential decay—for a lifetime of 10 times the decay time scale, the initial efficacy will have to be $\exp(10)$ times larger than the critical efficacy. In any initial efficacy distribution, such a memory will be extremely rare, because the critical efficacy is proportional to the square root of the distribution's second moment.

The effects of structural perturbations on memory function.   In this section we analyze the effect of damage to the circuit on memory storage and retrieval. In previous sections, we assumed for simplicity that the neuronal firing threshold is automatically adjusted to guarantee a fixed mean activation level, f (see "Methods"). Here we assume that the firing threshold is fixed since we anticipate that part of the effect of perturbation is the disruption of the level of activity. Importantly, in the case of constant threshold, the dependence of A_c on Δ is not linear. It has a non-zero value for small Δ reflecting the requirement for the encoding efficacy to be large enough for neurons to cross the threshold. Above some critical Δ, A_c rises abruptly, causing all memories to lose stability, due to over-activation of the network when the noise level is high (Fig. S4). At equilibrium Δ is below but close to the critical value (for the presented parameter range). In this scenario, the properties of the unperturbed system are similar to those of the fixed activity scenario, with a memory capacity that depends on the threshold value. For the presented results we used the threshold value 0.36 which maximizes capacity in unperturbed conditions (See SI).

Noisy synaptic dynamics.   We first consider perturbations of the synaptic learning and consolidation processes by adding white noise χ to the synaptic dynamics, for all $t \geq t_{\text{onset}}$, with a diffusion coefficient D,

$$\frac{dJ_{ij}}{dt} = -\frac{1}{\tau}J_{ij} + \sum_l \xi_i^l \xi_j^l (\delta(t-l) + bR_l(t)) + \chi_{ij}(t),$$
$$\langle \chi_{ij}(t) \rangle = 0,$$
$$\langle \chi_{ij}(t)\chi_{kl}(t') \rangle = D^2 \delta(t-t')\delta_{ik}\delta_{jl} \tag{13}$$

The effect of this noise is approximately an additive contribution to the total variance of local fields

$$\Delta^2(t) \approx \frac{1}{N}\sum_l A_l^2(t) + \frac{\tau D^2}{2N}\left(1 - \exp(-2(t-t_{\text{onset}})/\tau)\right) \tag{14}$$

After the noise onset, Δ increases rapidly above its critical value, causing a sharp increase in A_c, and the blocking of rehearsals for all memories. This in turn causes a rapid decrease in magnitude of stored memory efficacies, leading to a decrease in Δ below the critical value and a decrease in A_c to a value which is between the value before the onset and the value just after the noise onset. This new equilibrium value of A_c, with reduced capacity, occurs over $\sim \tau$ (Fig. 6a). Although overall reduction in capacity may be mild for moderate D values, there is a large reduction in the retrieval probability of memories that were encoded around the perturbation onset time, due to the sharp transient increase in A_c. In contrast, memories that have already been consolidated suffer only a mild reduction in survival probability (relative to unperturbed memories of the same age). Likewise, newly entered memories have a high probability of consolidation, since they experience the equilibrium value of A_c and their retrieval probability is similar to the unperturbed case (Fig. 6b).

Random synaptic silencing.   Another perturbation we consider is the death of a fraction of the synapses. We model the effect of the synaptic death by multiplying the connectivity matrix J_{ij} by a binary random dilution $\{0, 1\}$ matrix:

$$J_{ij} \rightarrow C_{ij}J_{ij}, \quad C_{ij} = \begin{cases} 1 \text{ with prob. } 1-p \\ 0 \text{ with prob. } p \end{cases} \tag{15}$$

Unlike the additive noise considered above, synaptic dilution process is multiplicative, reducing both the effective efficacy of each memory (by a factor $1-p$), and the interference noise Δ (by factor $\sqrt{1-p}$), and in general reduces the signal to noise ratio ("Methods"). After the dilution onset, A_c barely changes (due to the weak dependence of A_c on Δ in the constant threshold scenario), while all the efficacies are reduced, causing a reduction of retrievability that spreads over the entire age range (Fig. 6c,d) and a new equilibrium is achieved slowly. Interestingly, neural adaptation (modeled here as a decrease in the neural activation threshold) can reduce the memory loss due to silencing (also reported in[44–46]), by reducing the minimum efficacy required for activation, i.e., A_c, thereby recovering some of the gap between memory efficacies and A_c (Fig. 7). Thus, our model predicts a qualitative difference in the effects of the two types of perturbations: synaptic dilution affects memories of all ages, causing a reduction in capacity that develops over a long time and can be partially compensated for by threshold adaptation, while additive synaptic noise results in a deficit largely confined to the time of the perturbation onset, and fast convergence to a new equilibrium.

Distribution of synaptic decay times.   Experiments showing that the time scale of synaptic and spine turnover is variable[43], and observations of power law memory retention curves in some memory studies[1–4,47] encourage consideration of the properties of synaptic dynamics with heterogeneous decay time constants, yielding,
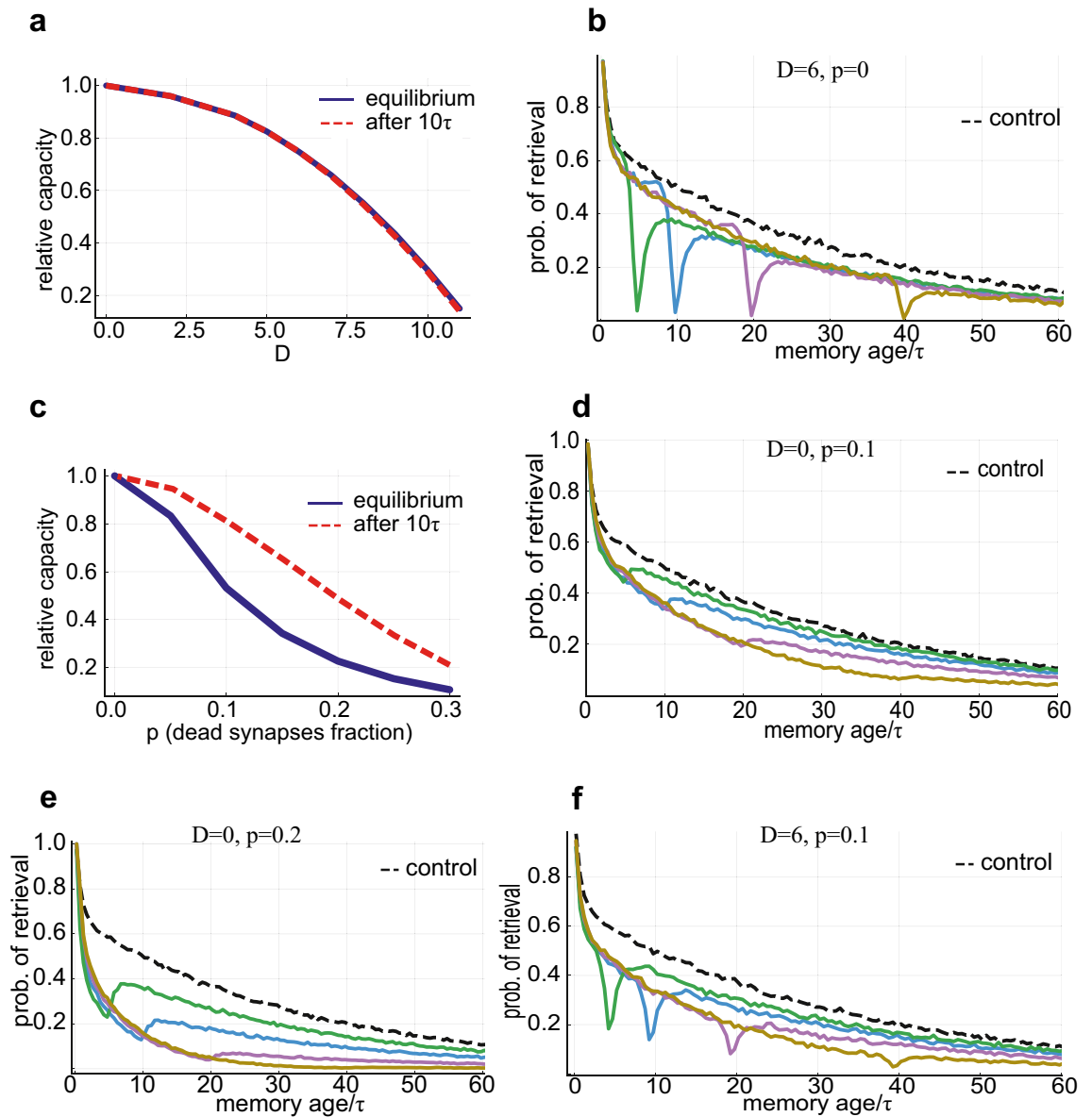
Figure 6. Perturbations and memory deficits. (**a**) The ratio between the capacity with and without injected noise vs. the diffusion coefficient *D*. (**b**) Retrieval probability vs. memory age with noisy synaptic dynamics ($D = 6$). Noise onset was before: 5τ (green), 10τ (blue), 20τ (purple), 40τ (brown). The control (black) is simulated with noiseless dynamics. (**c**) The ratio between the capacity with and without synaptic dilution vs. the silenced synapses fraction *p*. (**d**) Retrieval probability vs. memory age for random synaptic dilution ($p = 0.1$). Coloring as in (**b**). (**e**) Same as (**c**), but with $p = 0.2$. Memories of all ages are affected, with some non-monotonicity caused by the small efficacies of newly learned memories, dropping more easily below A_c. (**f**) Combination of synaptic dilution and noisy synaptic dynamics, $D = 6$ and $p = 0.1$. Coloring as in (**b**). Parameters: $N = 8000$, $\tau = 160$, $\lambda\tau = 10$, $b = 0.25$

$$J_{ij}(t) = \frac{1}{Nf(1-f)} \sum_l A_l^{ij}(t)\,(\xi_i^l - f)(\xi_j^l - f) \qquad (16)$$

where the efficacies $A_l^{ij}(t)$ contributed by each synapse obey

$$\frac{dA_l^{ij}}{dt} = -\frac{1}{\tau_{ij}}A_l^{ij} + R_l(t) \qquad (17)$$

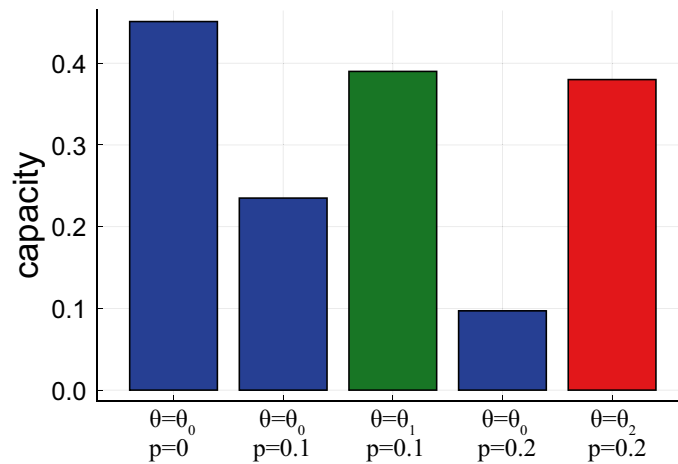The mean efficacy of each memory is the average over these contributions,

Figure 7. Effect of threshold adaptation. Blue bars show capacity with threshold optimized for the noiseless case ($\theta_0 = 0.36$). Green bar shows capacity with threshold optimized for low dilution (p=0.1, $\theta_1 = 0.31$). Red bar shows capacity for threshold optimized for high dilution (p=0.2, $\theta_2 = 0.29$). Parameters: $N = 8000, \ \tau = 160, \ \lambda\tau = 10, \ b = 0.25$

$$A_l(t) = \left\langle A_l^{ij}(t) \right\rangle_{\tau_{ij}} \tag{18}$$

where $\langle .. \rangle_{\tau_{ij}}$ denotes the average over the distribution of synaptic time constants. Likewise, the noise term is proportional to the sum of second moments of the efficacies:

$$\Delta^2(t) = \frac{f}{N} \sum_n \left\langle (A_n^{ij}(t))^2 \right\rangle_{\tau_{ij}} \tag{19}$$

As an example, we show the case where the decay time constants are power law (Pareto) distributed , i.e., $P(\tau) \propto \tau^{-(\alpha+1)}$; $\tau \geq \tau_0, \alpha > 0$ ("Methods"). In the absence of rehearsals (pure decay), there will be a global catastrophic forgetting for $\alpha \leq 1$, where the mean of the decay rate, and therefore the interference noise, diverges. For $\alpha > 1$ there will be a catastrophic age dependent forgetting, as in the case with a uniform decay time scale (See[13] SI). With stochastic nonlinear rehearsals, for large α the forgetting curve is approximately exponential, similar to the single τ case (Fig. 8b). This is because the dominant contribution comes from the shortest time τ_0. Interestingly, for intermediate values ($1 < \alpha < 1.8$), the forgetting curves have an approximately power-law decay (Fig. 8a). In this regime, the retrieval probability is affected by contributions from a broad range of time constants: neither the minimal τ nor outliers with very large values are dominant.

Discussion

Consolidation time scale. We have proposed a stochastic self-amplified memory consolidation mechanism and showed that it leads to smooth forgetting curves that extend much longer than the synaptic decay time. Our model provides estimates for the global long-term memory properties such as the capacity of the network, the shape of the forgetting curve and the average lifetime of memories. Translating synaptic decay time to realistic times is hard. In rodents, spine turnover time is estimated to be of the order of several weeks in the hippocampus and up to a year in the cortex[43,48–51]. In humans, these times may be longer given the lower metabolic rate[52,53]; however, there are no direct experimental evidence. In addition, the model synaptic decay time τ is in units of the mean inverse rate of encoding of episodic memories, which is hard to estimate, but it is likely to be of the order of weeks. Thus, assuming a human spine turnover time of the order of months yields τ of the order of tens of months, which could lead to mean memory lifetime of several years. At present, these estimates are speculative.

Memory deficiencies. We have considered two types of perturbations to the memory circuit: synaptic death and increased synaptic noise. Both types of damage result in reduced retrievability of memories introduced prior to the damage onset, a phenomenon known in the literature as retrograde amnesia[16–19]. Due to the consolidation effect in our model, the amnesia is temporally graded: memories learned just before the noise onset are more severely affected than older ones, because they didn't have enough time to consolidate, and were more fragile at the onset time. This effect is more prominent in the case of increased synaptic noise than synaptic dilution, due to the sharper drop in basins of attraction size after the noise addition. Perturbations cause a drop in retrieval performance of new memories entering after the perturbation onset, a manifestation of anterograde amnesia. This effect is temporally graded as well, being more severe for memories introduced just after the onset, and is especially prominent deficit in the additive noise case. Another interesting difference is the approach to a new equilibrium, which is fast in the case of additive noise but slow in the dilution. In addition, threshold adaptation can compensate part of the memory dysfunction caused by dilution, but not by additive noise.
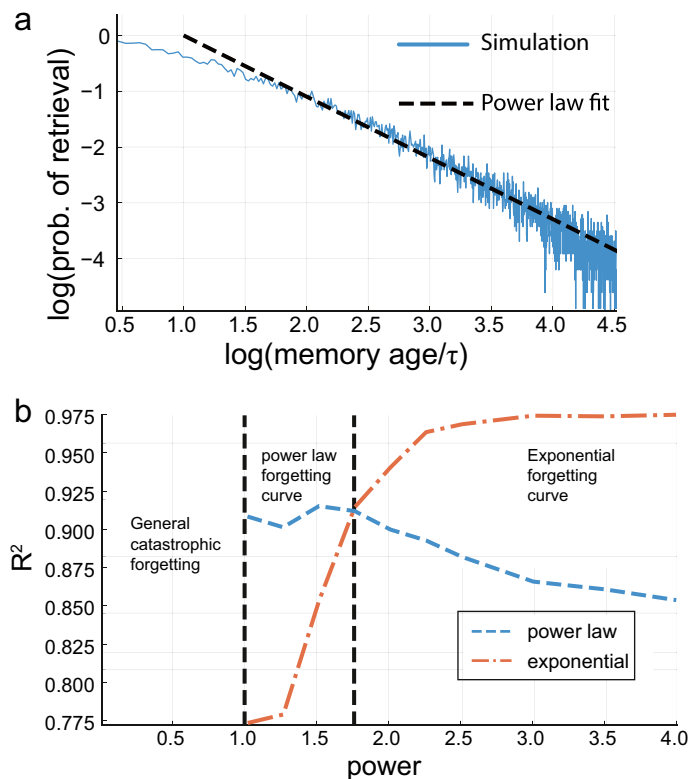
Figure 8. Power-law synaptic decay characteristic time distribution. (**a**) Retrieval probability as a function of memory age on log-log scale (blue), with power-law fit in black (dashed line, slope≈ 1). Here $\lambda\tau = 5$ for the empirical mean τ, $b = 0.25$, the power $\alpha = 1.5$. The minimal τ is 20. (**b**) Goodness of fit (R^2) for the forgetting curve using an exponential function (orange, dash-dot) and power-law function (blue), as a function of the power parameter of the characteristic time distribution.

Our model also allows for exploration of transient perturbations (SI), where the damage lasts for a finite time window[54]. In this case there is again a temporally-graded retrograde amnesia. Interestingly, new memories introduced after the end of the event not only regain retrievability, but can even improve their retrievability compared to control. This is due to the increased forgetting rate during the event, which results in lower interference noise and increased rehearsal rate after the event end.

The predictions of our model should be contrasted with the pure decay model where similar perturbations reduce the capacity (maximal age for retrievable memories), but don't introduce any non-monotonicity in the forgetting curve, which is still a step function, but with a reduced width.

Relation to previous models. In the classical theory of systems memory consolidation[16–18,20], the interaction between the hippocampus (HC) and the cortex plays a central role, with HC storing memories for short periods of time, and following rehearsals, memories are transmitted to the cortex for long-term storage. In the recent Multiple Trace Theory (MTT)[20,29] autobiographic memories are stored for long term memory in both HC and cortex and consolidated through rehearsals that establish multiple memory traces in HC. This model shares some key elements with our theory, such as ongoing, life-long consolidation of memories and rehearsals which make memories more robust to perturbations. However, it is unclear how MMT can scale to large numbers of stored memories. In addition, in[29] the rehearsal statistics (new trace formation) don't depend on the robustness of the memories, nor does the model take into account interference between memories.

In the pseudo-rehearsal model[28], new memories are learned in batches. After each batch is learned, there is a series of presudo-rehearsals, which are learning of fixed point states that are found by random initialization of the network's state. This model implements stochastic rehearsals explicitly, and it reduces loss of old information by the pseudo-rehearsals. While in our model we use Hebbian, one-shot learning of new memories, the pseudo-rehearsal model relies on multi step gradient descent of each memory batch, which is generally a less biologically plausible learning algorithm. In addition, the authors present results only for a single, small size network and it is not clear how the results scale with the size of the network, nor how the increase of memory life time depends on the different properties of the model.

A few studies used neural network models where rehearsals are modeled as random visits of learned memories[31,33,36], or implicit rehearsals (via memory traces embedded in the noise correlations[30]). In[33] the authors shows the effects of rehearsals on a large number of learned memories. Yet they do not explicitly model the "Hippocampus" part of the model where rehearsals occur, and they do not provide quantitative relations between the

model's performance and the different parameters, such as number of neurons in the "neocortex" part. In[30,31,36] The authors study the effect of rehearsals in networks with specific size, small number of memories, that don't scale with the network size. It is not clear how their results scale with the network size, and how does the enhancement in memory lifetime depends on the different network properties. The authors of[30,36] consider rehearsals of a finite batch of previously encoded memories rather than with life-long learning as in our study. Benna and Fusi[13] studied memory storage with complex synapses, where a consolidation process is implemented in the dynamics of synapses, with a cascade of synaptic characteristic times. They show that their mechanism gives rise to a power-law decay of the signal-to-noise ratio (SNR, equivalent to $A_l(t)/\Delta(t)$ in our model) with age. However, this model still exhibits a deterministic catastrophic age-dependent forgetting, such that all memories older than a critical age are non-retrievable, whereas all newer memories are almost perfectly retrievable. A recent phenomenological model[47] derives a power-law form for memory retention curves with a power of 1 or smaller. A power close to 1 for intermediate ages is consistent with our result for a power law distribution of synaptic decay time. However, at present, it is unclear whether the experimental paradigms and time scales in which a power law is observed are relevant to life long episodic memory.

Fiebig and Lansner[35] proposed a three component model, each with different synaptic decay rate, which performs continual learning with self-generated rehearsals. Similar to our work, they study the effects of perturbations and show similarities to human data. However, this work does not provide an analysis of the model, and does not explore the dependence on the different parameters such as network size and synaptic decay time and rehearsal rates. Comparison with our results is hampered also by the fact that synaptic decay in their model is an active process, dependent on memory arrivals among other factors. In general, none of the past models provide quantitative analysis of the memory capacity and the memory lifetime statistics, while enabling lifelong learning and avoiding global and critical-age catastrophic forgetting.

Catastrophic forgetting in memory models vs. machine learning models.

There is a fundamental difference between the catastrophic forgetting nature in long term memory network models, which we address here, and what is called catastrophic forgetting in machine learning, and deep learning especially[55–57]. In the deep learning literature, it is assumed that if all the data was available all the time the model was able to learn from it and successfully solve the relevant task. In other words, the model is in a regime below its capacity limits, and the problem is the online, incremental presentation of the data, which causes stability-plasticity issues. In contrast, when modeling long term memory we assume that there is too much information to be stored- even if the data to be stored was available all the time and learning wasn't online, still we would encounter catastrophic forgetting due to crossing of the model capacity limit.

Limitations and future work.

In this paper we don't explicitly model the rehearsals process—how the system moves between activation states and visits different attractors. Possible mechanisms could be destabilization of attractors by adaptation[36,58] or transitions induced by random initialization processes[31]. These mechanisms will generate a rate of rehearsals per memory that depends on its basin of attraction size, as in our model, but whether the rate is simply proportional to the basin's size as we assume is yet to be tested.

Our model can be extended in a variety of ways, including more biologically plausible neuronal and synaptic integration. For example, detailed neuron models with rich morphologies and dendritic structure might influence the memory capacity and lifetime, in addition to the number of neurons. In addition, our model does not obey Dale's law, and restricting neurons to be either excitatory or inhibitory, while also allowing different long term and short term plasticity dynamics for the different populations, will likely introduce richer memory properties.

Studying more synaptic decay characteristic time distributions, such as bimodal distributions where one synapse population decays much slower than another population, could also give rise to interesting memory properties. A relevant example is an inhibitory population with slow synaptic decay, interacting with an excitatory population with faster synaptic decay. We limit ourselves in this work to the mechanism of synaptic decay, while other mechanisms, such as bounded, discrete synapses[14,15] can also prevent global catastrophic forgetting. Due to the qualitative similarity of the behavior of such models to synaptic decay models (palimpsestic behavior with critical age catastrophic forgetting), we expect that this model will respond similarly to the introduction of stochastic rehearsals. Additionally, our framework allows for analyzing the effect of other types of perturbations, such as post-traumatic stress disorder amnesia[59,60].

Conclusions.

The stochastic nonlinear rehearsal mechanism proposed in our work is, to the best of our knowledge, the first large-scale memory model that gives rise to realistic gracefully decaying forgetting probability curves, with exponential or power law tails depending on synaptic decay rate distribution. Our model's capacity scales as a power law of the number of neurons, with a power that approaches unity for a large mean number of rehearsal events per synaptic decay time. Our model's capacity interpolates between two extreme cases—for low rehearsal rate it approaches the pure forgetting case (logarithmic scaling of the capacity with the number of neurons), where learning is incremental and synapses decay, and for high rehearsal rate it approaches linear scaling of capacity with N, as in the static case without synaptic decay, where memories are stored (not necessarily incremental, they can be available forever) until capacity is reached and then all memories become irretrievable. Our model predicts that the onset of perturbation to the circuit, in the form of synaptic noise, leads to non-monotonic memory deficits affecting more strongly memories encoded around perturbation onset time, which have not yet a chance to consolidate. The richness of the model behavior in normal and diseased conditions provides a theoretical framework for predictions and testing against empirical data on human memory.

Methods

Network model.   As described in section II, memories are modeled as sparse, uncorrelated N dimensional activation patterns (N is the number of neurons), and the synaptic dynamics are governed by three processes: deterministic synaptic decay with rate $1/\tau$, Hebbian learning of new memories, and rehearsal of old memories, which is the central novelty of our model. In continuous time, Eq. (1) becomes:

$$\frac{dJ_{ij}}{dt} = -\frac{1}{\tau}J_{ij} + \sum_l \xi_i^l \xi_j^l \delta(t-l) + b\sum_k \xi_i^k \xi_j^k R_k(t) = -\frac{1}{\tau}J_{ij} + \sum_l \xi_i^l \xi_j^l (\delta(t-l) + bR_l(t)) \tag{20}$$

The rehearsals are modeled as a point process

$$R_l(t) = \sum_{\{t_l^n\}} \delta(t - t_l^n) \tag{21}$$

Inserting the ansatz (3), we get that A_l obeys the differential equation:

$$\frac{dA_l}{dt} = -\frac{1}{\tau}A_l + bR_l(t) \tag{22}$$

With $A_l(t) = 0$ for $t < l$ and $A_l(l) = 1$.

The single neuron dynamics are binary, and given by:

$$\sigma_i(t+dt) = \Theta(h_i(t) - \theta) \tag{23}$$

where $\sigma_i(t)$ is the state of neuron i at time t, $\Theta(x)$ is the Heaviside step function, $h_i(t)$ is the local field (total input received by neuron i at time t):

$$h_i(t) = \sum_j J_{ij}\sigma_j(t) \tag{24}$$

and θ is a threshold, set at every time step such that the total activation of the network is maintained and equal to fN. This can be thought of as the effect of an inhibitory population, regulating the total population activity (practically, in full simulations, at each time step we choose the fN neurons with the largest local fields and set their state to one, and all the others are set to zero).

Mean field equations, basins of attraction.   We would like to find the relation between memory stability, measured by the memory pattern's basin of attraction size, and the efficacy of the memory and all other memories in the system. First, we define two useful quantities: f_+^l is the probability of a neuron to be active in the current state given that it is active in the memory pattern ξ^l. f_-^l is the probability of a neuron to be active in the current state given that it is not active in the memory pattern ξ^l. In other words, f_+^l is the fraction out of the neurons active in memory state l that are active in the current state. f_-^l is the fraction out of the neurons not active in memory state l that are active in the current state. We will omit the l dependence of $f_\pm$ from now on. In terms of these quantities,

$$f = \frac{1}{N}\sum_{i=1}^N \sigma_i(t) = ff_+ + (1-f)f_- \tag{25}$$

For clarity, in this section instead of the memory patterns definition we use above (Eq. (2)) we define the patterns in an equivalent, more explicit way:

$$\xi_i^l = \begin{cases} 1 \text{ with prob. } f \\ 0 \text{ with prob. } 1-f \end{cases} \tag{26}$$

, and we normalize the connectivity matrix accordingly.

The overlap between memory pattern l and the system's state σ

$$M_l \equiv \frac{1}{Nf(1-f)}\sum_{j=1}^N (\xi_j^l - f)\sigma_j = \frac{1}{Nf(1-f)}\left(\sum_{j=1}^N \xi_j^l \sigma_j(t) - f\sum_{j=1}^N \sigma_j(t)\right)$$
$$= \frac{1}{Nf(1-f)}(Nff_+ - Nf^2) = \frac{1}{(1-f)}(f_+ - (ff_+ + (1-f)f_-) = f_+ - f_- \tag{27}$$

Now, assume that the current state is close to the memory state ξ^l. The input to neuron i which is active in memory pattern l $(\xi_i^l = 1)$:

$$h_i^+ = \frac{1}{Nf(1-f)} \sum_{j=1; \, j\neq i}^{N} \sum_{n=1}^{T} A_n(\xi_i^n - f)(\xi_j^n - f)\sigma_j(t) =$$

$$= A_l(1-f)M_l + \frac{1}{Nf(1-f)} \sum_{j=1; \, j\neq i}^{N} \sum_{n=1; \, n\neq l}^{T} A_n(\xi_i^n - f)(\xi_j^n - f)\sigma_j(t) \tag{28}$$

Averaging h_i^+ over memories realizations gives $A_l(1-f)M_l$, and the variance:

$$\Delta^2 = \frac{1}{N^2 f^2 (1-f)^2} \sum_{j,j'=1; \, j,j'\neq i}^{N} \sum_{n,n'=1; \, n,n'\neq l}^{T} A_n A_{n'} \langle (\xi_i^n - f)(\xi_i^{n'} - f)(\xi_j^n - f)(\xi_{j'}^{n'} - f)\sigma_j \sigma_{j'} \rangle \tag{29}$$

Now, because the system state is close to the memory state ξ^l, we can assume it is uncorrelated with all other memory states. Hence, there are contributions only from terms with $j = j'$, $n = n'$:

$$\Delta^2 = f \cdot \frac{1}{N} \sum_{n=1; \, n\neq l}^{T} A_n^2 \tag{30}$$

Applying the central limit theorem, we approximate h_i^+ by a Gaussian variable with mean $A_l(1-f)M_l$ and variance Δ^2. Now, f_+ is the probability for h_i^+ to be larger than θ, which is given by the complimentary error function, $H(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-0.5t^2)dt$:

$$f_+ = H\left(\frac{\theta - A_l(1-f)M_l}{\Delta} \right) \tag{31}$$

In a similar way (same noise term, mean equals $-A_l M_l f$) we find for f_- (for $f \ll 1$):

$$f_- = H\left(\frac{\theta + A_l M_l f}{\Delta} \right) \tag{32}$$

Note that we didn't set the threshold θ, but instead demanded a constant population activation f. Equations (25), (27), (31) and (32) allow us to write an equation for the overlap dynamics:

$$M_l(t+1) = G(M_l(t), A_l(t)/\Delta(t)) \tag{33}$$

where

$$G(M_l, A_l/\Delta) = H\left(\theta^*(M_l, A_l/\Delta) - \frac{A_l(1-f)M_l}{\Delta} \right) - H\left(\theta^*(M_l, A_l/\Delta) + \frac{A_l M_l f}{\Delta} \right)$$
$$= H\left(H^{-1}(f(1-M_l)) - \frac{A_l M_l}{\Delta} \right) - f(1-M_l) \tag{34}$$

and

$$\theta^*(M_l, A_l/\Delta) = H^{-1}(f(1-M_l)) - f \cdot M_l \cdot A_l/\Delta \tag{35}$$

Therefore, the equation for the overlap fixed points is:

$$M_l = H\left(H^{-1}(f(1-M_l)) - \frac{A_l M_l}{\Delta} \right) - f(1-M_l) \tag{36}$$

Now, we numerically find the fixed points for M at a given A/Δ by running the dynamics described by Eq. (33). Typically (for large enough A/Δ), there will be a stable fixed point at $M = 0$, a stable fixed point $0 < M_s \leq 1$ and an unstable fixed point $0 < M_{us} < M_s$. We approximate the basin of attraction size as the distance $M_s - M_{us}$. This way, we obtain the basin size as a function of A/Δ, $F(A/\Delta)$. We check the validity of our approximations by simulating a full neural network model and checking numerically the basin of attraction sizes, and find good agreement (SI), which is the basis for the good agreement in retention curves between the mean field simulations and the full network simulation (Fig. 3a,b). We define the critical efficacy A_c as the efficacy for which $M_s = M_{us}$ (meaning, the non-zero overlap solution loses stability). This happens approximately when $M_s = M_{us} \approx 0.85$. As one can see from Eq. (36), the fixed points depends only on the ratio A/Δ and on f, and therefore A_c/Δ is only a function of f, and we can write $A_c = a(f)\Delta$. For $f = 0.01$ (the typical value we use throughout the manuscript) we find numerically that $a(f) \approx 4.7$. Analytical approximation for $a(f)$ is given in the SI.

Numerical simulations. In our simulations, we first numerically solve the coupled stochastic differential equations for the efficacies (Eq. (7)). Theoretically our model considers infinite number of memories. However, practically we solve the equations for a finite but large number of memory efficacies, typically $200\tau - 1000\tau$, chosen such that A_c saturates to its steady state value. We measure time in units of the lag between the introduction of two consecutive memories (assumed constant). At every integration time step dt (small compared to all characteristic timescales of the system, typically $dt = 0.05/\lambda$), a rehearsal event might occur for each memory with $A_l \geq A_c$. We generate a uniform random number between 0 and 1 and compare it to $\lambda \cdot F(A_l(t)/\Delta(t)) \cdot dt$. A rehearsal event of memory l will happen at time t if the uniform random number is

smaller than $\lambda \cdot F(A_l(t)/\Delta(t) \cdot dt$. This approximates the statistics of a non-homogeneous Poisson process. By averaging over many such realizations (typically 500), we calculate the efficacy histogram, capacity (counting how many efficacies are above A_c) and retrieval probabilities (by checking the probability for the efficacy of a memory introduced l time units into the past to be retrievable now). These calculations are referred to as "mean field simulations", and they don't include generation of random memories and building the connectivity matrix.

Full network simulation.　When simulating the full network model, after generating the efficacies, we randomly generate memory patterns (binary vectors of dimension N) to be stored, and build the connectivity matrix according to Eq. (3). Then, to measure retrievability, we initialize the network's state at a memory pattern, and let the binary neurons dynamics run until they settle to a steady state. Then, we measure the overlap between the pattern and the steady-state activity. We say a memory is retrievable if the overlap is ≥ 0.85. For measuring the basin of attraction sizes of the memory patterns, we generate an initial state by randomly flipping the state of units in the memory pattern (conserving the total activation fN), and run the dynamics until convergence. Then we measure the overlap between the final state and the memory pattern. We keep increasing the number of flipped units until the final state has an overlap smaller than ≥ 0.85 with the memory state. We define the normalized basin size as the maximal number of flips allowing for a large overlap divided by $2fN$, the maximal number of flips. Results are shown in the SI.

Noisy synaptic dynamics.　The synaptic dynamics in the presence of Gaussian noise is presented in Eq. (13). It is straightforward to show that a solution to the equation can be written as:

$$J_{ij}(t) = \sum_l \xi_i^l \xi_j^l A_l(t) + \int_0^t e^{-(t-t')/\tau} \chi_{ij}(t')dt' \tag{37}$$

with $A_l(t)$ obeying Eq. (7) as before. The nonlinear effect of the noise arises through the self consistent requirement, that the rehearsal rate of memory l is proportional to the basin of attraction size of this memory, which depends on A_l and on Δ. We calculate Δ with the injected noise (Eq. (14)) the same way we calculated Δ without noise above. Here there is a non trivial mixed term involving the average $\langle A_l(t)\chi_{ij}(t)\rangle$, which we found numerically to be negligible for the parameter range we are interested in.

Synaptic dilution.　The random silencing is done by multiplying the connectivity matrix J_{ij} by a binary matrix:

$$J_{ij} \rightarrow C_{ij}J_{ij}, \quad C_{ij} = \begin{cases} 1 \text{ with prob. } 1-p \\ 0 \text{ with prob. } p \end{cases} \tag{38}$$

We would like to calculate the effect of the dilution on the memory efficacies dynamics, and for that we need to find the effect on $A_l(t)$ and on $\Delta(t)$. Let us calculate the local field near memory l as before:

$$h_i^{l+} = \sum_{j=1; \, j\neq i}^N C_{ij}J_{ij}\sigma_j(t) \tag{39}$$

Taking the mean over memories and over C_{ij} realizations (denoted by []) we get:

$$\left[h_i^{l+}\right] = (1-p)\left\langle \sum_{j=1; \, j\neq i}^N J_{ij}\sigma_j(t) \right\rangle = (1-p)(1-f)M_lA_l \tag{40}$$

Here $\langle \rangle$ denotes average over memories realizations. As one can see, the efficacies are scaled by a factor of $1-p$. We assumed here we can neglect correlations between A_l and C_{ij}. The second moment:

$$\left[\left(h_i^{l+}\right)^2\right] = \sum_{j,j'\neq i} \left[J_{ij}J_{ij'}C_{ij}C_{ij'}\sigma_j\sigma_j'\right] = \sum_{j,j'\neq i} \left((1-p)\delta_{jj'} + (1-p)^2(1-\delta_{jj'})\right)\left\langle J_{ij}J_{ij'}\sigma_j\sigma_j'\right\rangle \tag{41}$$

There are two contributions arising from the C_{ij} randomness. Now, when calculating the local field variance, the term proportional to $(1-p)^2$ is exactly canceled by the squared mean, and we are left with:

$$\left[\left(h_i^{l+} - \left[h_i^{l+}\right]\right)^2\right] = (1-p)\sum_{j,j'\neq i} \left\langle J_{ij}^2\sigma_j^2\right\rangle = (1-p)\tilde{\Delta}^2 \tag{42}$$

where $\tilde{\Delta}^2$ is the local field variance without dilution.

Now, we obtain the efficacies modified dynamics by using these expressions for the signal and noise to calculate the basins of attraction sizes as before.

Non-uniform characteristic decay time.　Assuming synapse J_{ij} has a decay rate ϵ_{ij}, and all memories have unit initial efficacy. Memory l appears for the first time at time l. The learning dynamics is:

15

$$J_{ij}(t + \Delta t) = (1 - \epsilon_{ij}\Delta t)J_{ij} + \sum_l \xi_i^l \xi_j^l \delta_{t,l} + b \sum_{k,n_k} \xi_i^l \xi_j^l \delta_{t,t_k^{n_k}} \tag{43}$$

In continuous time,

$$\frac{dJ_{ij}}{dt} = -\epsilon_{ij}J_{ij} + \sum_l \xi_i^l \xi_j^l \delta(t - l) + b \sum_k \xi_i^l \xi_j^l R_k(t) = -\epsilon_{ij}J_{ij} + \sum_l \xi_i^l \xi_j^l (\delta(t - l) + R_l(t)) \tag{44}$$

$$R_k(t) = b \sum_{\{t_k^n\}} \delta(t - t_k^n) \tag{45}$$

Assuming all synapses starts at zero value, the solution can be written as:

$$J_{ij}(t) = \sum_l \xi_i^l \xi_j^l \int_0^t dt' e^{-\epsilon_{ij}(t-t')}(\delta(t - l) + R_l(t)) = \sum_l \xi_i^l \xi_j^l A_l^{ij}(t) \tag{46}$$

Let us define efficacies

$$A_l^{ij} = \Theta(t - l) \int_0^t dt' e^{-\epsilon_{ij}(t-t')}(A_0\delta(t - l) + R_l(t)) \tag{47}$$

A_l^{ij} obeys the differential equation:

$$\frac{dA_l^{ij}}{dt} = -\epsilon_{ij}A_l^{ij} + R_l(t) \tag{48}$$

With $A_l^{ij}(t) = 0$ for $t < l$ and $A_l^{ij}(l) = 1$.

Given that the decay rates have a probability density $\rho(\epsilon)$, let us define:

$$A_k(t) = \left\langle A_k^{ij}(t) \right\rangle_{\epsilon_{ij}} = \int d\epsilon \rho(\epsilon) \left(\Theta(t - l) \int_0^t dt' e^{-\epsilon(t-t')}(A_0\delta(t - l) + R_l(t)) \right) \tag{49}$$

and

$$A_k^{ij}(t) = A_k(t) + \delta A_k^{ij}(t) \tag{50}$$

$$J_{ij}(t) = \sum_l \xi_i^l \xi_j^l A_l(t) + \sum_l \xi_i^l \xi_j^l \delta A_l^{ij}(t) \tag{51}$$

Including normalization and sparseness considerations,

$$J_{ij}(t) = \frac{1}{Nf(1-f)} \left(\sum_l (\xi_i^l - f)(\xi_j^l - f)A_l(t) + \sum_l (\xi_i^l - f)(\xi_j^l - f)\delta A_l^{ij}(t) \right) \tag{52}$$

Now let us calculate the mean local field on neuron i in a state near memory state k, and assume $\xi_i^k = 1$:

$$h_i^k = \sum_{j \neq i} J_{ij}\sigma_j = (1 - f)M_k A_k + \frac{1}{Nf} \sum_{j \neq i} (\xi_j^k - f)\delta A_k^{ij}(t)\sigma_j + $$
$$+ \frac{1}{Nf(1-f)} \left(\sum_{l \neq k, j \neq i} (\xi_i^l - f)(\xi_j^l - f)\sigma_j A_l(t) + \sum_{l \neq k, j \neq i} (\xi_i^l - f)(\xi_j^l - f)\sigma_j \delta A_l^{ij}(t) \right) \tag{53}$$

Taking an average over the memories realizations and the decay rates, we get:

$$\left\langle h_i^k \right\rangle_{\xi,\epsilon} = (1 - f)M_k A_k \tag{54}$$

And the variance:

$$\left\langle \left(\delta h_i^k\right)^2 \right\rangle_{\xi,\epsilon} = \frac{f}{N} \sum_{l \neq k} \left\langle (A_l^{ij})^2 \right\rangle_\epsilon + \left(\left\langle (A_k^{ij})^2 \right\rangle_\epsilon - A_k^2 \right) \frac{1}{N^2 f^2} \sum_{j \neq i} \left\langle (\xi_j^k - f)^2 \sigma_j^2 \right\rangle_\xi \tag{55}$$

The second term does not include summation over all memories, and therefore it is negligible for large N values (the first term is $O(1)$ while the second is $O(N^{-2})$). This leads to Eq. (19).

Power law τ distribution. For each synapse we generated synaptic decay characteristic times from a power law (Pareto) distribution with density:

$$P(\tau_0) = \begin{cases} \alpha \cdot \tau_0^{-(\alpha+1)} & \tau_0 \geq 1 \\ 0 & \tau_0 < 1 \end{cases} \qquad (56)$$

In this distribution, for $\alpha < 1$ the mean diverges. We scaled the resulting τ_0^{ij} values by a uniform factor: $\tau^{ij} = 2\omega \cdot N \cdot \tau_0^{ij}$. We fixed ω value for the average number of rehearsals per mean decay time R_0, and used it to set the λ parameter by dividing R_0 by the empirical average of the generated decay times. Next we solved the stochastic differential Eq. (48) numerically. The rehearsals are generated with time dependent rates proportional to the basin of attraction size, now as a function of the average and variance of the memory efficacies over all synaptic timescales.

Data availability
All presented results were obtained by a custom code written in the Julia language. Code will be made available upon request. No experimental data was used or generated.

References
1. Rubin, D. C. On the retention function for autobiographical memory. *J. Verbal Learn. Verbal Behav.* **21**, 21–38 (1982).
2. Rubin, D. C. & Schulkind, M. D. The distribution of autobiographical memories across the lifespan. *Mem. Cogn.* **25**, 859–866. https://doi.org/10.3758/BF03211330 (1997).
3. Meeter, M., Murre, J. M. & Janssen, S. M. Remembering the news: Modeling retention data from a study with 14,000 participants. *Mem. Cogn.* **33**, 793–810 (2005).
4. Averell, L. & Heathcote, A. The form of the forgetting curve and the fate of memories. *J. Math. Psychol.* **55**, 25–35 (2011).
5. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.* **79**, 2554–2558 (1982).
6. Amit, D. J., Gutfreund, H. & Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.* **55**, 1530 (1985).
7. Amit, D. J., Gutfreund, H. & Sompolinsky, H. Information storage in neural networks with low levels of activity. *Phys. Rev. A* **35**, 2293 (1987).
8. Nadal, J., Toulouse, G., Mézard, M., Changeux, J. & Dehaene, S. Neural networks: Learning and forgetting. In *Computer Simulations and Brain Science* (Cambridge University Press, 1986).
9. Parisi, G. A memory which forgets. *J. Phys. A: Math. Gen.* **19**, L617 (1986).
10. Mézard, M., Nadal, J. & Toulouse, G. Solvable models of working memories. *J. Phys.* **47**, 1457–1462 (1986).
11. Akaho, S. Capacity and error correction ability of sparsely encoded associative memory with forgetting process. In *International Conference on Artificial Neural Networks*, 707–710 (Springer, 1993).
12. Storkey, A. J. & Valabregue, R. The basins of attraction of a new hopfield learning rule. *Neural Netw.* **12**, 869–876 (1999).
13. Benna, M. K. & Fusi, S. Computational principles of synaptic memory consolidation. *Nat. Neurosci.* **19**, 1697–1706 (2016).
14. Tsodyks, M. Associative memory in neural networks with binary synapses. *Mod. Phys. Lett. B* **4**, 713–716 (1990).
15. Amit, D. J. & Fusi, S. Learning in neural networks with material synapses. *Neural Comput.* **6**, 957–982 (1994).
16. Squire, L. R. & Alvarez, P. Retrograde amnesia and memory consolidation: a neurobiological perspective. *Curr. Opin. Neurobiol.* **5**, 169–177 (1995).
17. Nadel, L. & Moscovitch, M. Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* **7**, 217–227. https://doi.org/10.1016/S0959-4388(97)80010-4 (1997).
18. Brown, A. S. Consolidation theory and retrograde amnesia in humans. *Psychon. Bull. Rev.* **9**, 403–425 (2002).
19. Frankland, P. W. & Bontempi, B. The organization of recent and remote memories. *Nat. Rev. Neurosci.* **6**, 119–130 (2005).
20. Moscovitch, M., Nadel, L., Winocur, G., Gilboa, A. & Rosenbaum, S. The cognitive neuroscience of remote episodic, semantic and spatial memory. *Curr. Opin. Neurobiol.* **16**, 179–190. https://doi.org/10.1016/j.conb.2006.03.013 (2006).
21. Axmacher, N. & Rasch, B. *Cognitive Neuroscience of Memory Consolidation* (Springer, 2017).
22. Fellner, M.-C., Waldhauser, G. T. & Axmacher, N. Tracking selective rehearsal and active inhibition of memory traces in directed forgetting. *Curr. Biol.* **30**, 2638-2644.e4. https://doi.org/10.1016/j.cub.2020.04.091 (2020).
23. Wilson, M. A. & McNaughton, B. L. Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676–679 (1994).
24. Stickgold, R. Sleep-dependent memory consolidation. *Nature* **437**, 1272–1278 (2005).
25. Rasch, B., Büchel, C., Gais, S. & Born, J. Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science* **315**, 1426–1429 (2007).
26. Diekelmann, S. & Born, J. The memory function of sleep. *Nat. Rev. Neurosci.* **11**, 114–126 (2010).
27. Carr, M. F., Jadhav, S. P. & Frank, L. M. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nat. Neurosci.* **14**, 147 (2011).
28. Robins, A. & McCALLUM, S. Catastrophic forgetting and the pseudorehearsal solution in hopfield-type networks. *Connect. Sci.* **10**, 121–135 (1998).
29. Nadel, L., Samsonovich, A., Ryan, L. & Moscovitch, M. Multiple trace theory of human memory: Computational, neuroimaging, and neuropsychological results. *Hippocampus* **10**, 352–368 (2000).
30. Wei, Y. & Koulakov, A. A. Long-term memory stabilized by noise-induced rehearsal. *J. Neurosci.* **34**, 15804–15815 (2014).
31. Meeter, M. & Murre, J. M. Tracelink: A model of consolidation and amnesia. *Cogn. Neuropsychol.* **22**, 559–587 (2005).
32. Murre, J. M. J., Chessa, A. G. & Meeter, M. A Mathematical Model of Forgetting and Amnesia. *Front. Psychol.* **4**, 76. https://doi.org/10.3389/fpsyg.2013.00076 (2013).
33. Káli, S. & Dayan, P. Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nat. Neurosci.* **7**, 286–294 (2004).
34. Roxin, A. & Fusi, S. Efficient partitioning of memory systems and its importance for memory consolidation. *PLoS Comput. Biol.* **9**, e1003146 (2013).
35. Fiebig, F. & Lansner, A. Memory consolidation from seconds to weeks: A three-stage neural network model with autonomous reinstatement dynamics. *Front. Comput. Neurosci.* **8**, 64 (2014).
36. Fauth, M. J. & Van Rossum, M. C. Self-organized reactivation maintains and reinforces memories despite synaptic turnover. *ELife* https://doi.org/10.7554/eLife.43717 (2019).
37. Tsodyks, M. V. & Feigelman, M. V. The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett. (EPL)* **6**, 101–105. https://doi.org/10.1209/0295-5075/6/2/002 (1988).
38. Tsodyks, M. Associative memory in asymmetric diluted network with low level of activity. *EPL Europhys. Lett.* **7**, 203 (1988).

39. Rolls, E. T. *Attractor Network Dynamics, Transmitters, and Memory and Cognitive Changes in Aging* (Cambridge University Press, 2019).
40. Rolls, E. T. & Treves, A. The neuronal encoding of information in the brain. *Prog. Neurobiol.* **95**, 448–490 (2011).
41. Naim, M., Katkov, M., Romani, S. & Tsodyks, M. Fundamental law of memory recall. *Phys. Rev. Lett.* **124**, 018101 (2020).
42. Hebb, D. O. *The organization of behavior: A neuropsychological theory* (Chapman & Hall, 1949).
43. Mongillo, G., Rumpel, S. & Loewenstein, Y. Intrinsic volatility of synaptic connections-a challenge to the synaptic trace theory of memory Synapses and memory. *Curr. Opin. Neurobiol.* **46**, 7–13. https://doi.org/10.1016/j.conb.2017.06.006 (2017).
44. Horn, D., Ruppin, E., Usher, M. & Herrmann, M. Neural network modeling of memory deterioration in alzheimer's disease. *Neural Comput.* **5**, 736–749 (1993).
45. Horn, D., Levy, N. & Ruppin, E. Neuronal-based synaptic compensation: a computational study in alzheimer's disease. *Neural Comput.* **8**, 1227–1243 (1996).
46. Horn, D., Levy, N. & Ruppin, E. Memory maintenance via neuronal regulation. *Neural Comput.* **10**, 1–18 (1998).
47. Georgiou, A., Katkov, M. & Tsodyks, M. Retroactive interference model of forgetting. *J. Math. Neurosci.* **11**, 1–15 (2021).
48. Grutzendler, J., Kasthuri, N. & Gan, W.-B. Long-term dendritic spine stability in the adult cortex. *Nature* **420**, 812–816 (2002).
49. Zuo, Y., Lin, A., Chang, P. & Gan, W.-B. Development of long-term dendritic spine stability in diverse regions of cerebral cortex. *Neuron* **46**, 181–189 (2005).
50. Holtmaat, A. J. *et al.* Transient and persistent dendritic spines in the neocortex in vivo. *Neuron* **45**, 279–291 (2005).
51. Attardo, A., Fitzgerald, J. E. & Schnitzer, M. J. Impermanence of dendritic spines in live adult ca1 hippocampus. *Nature* **523**, 592–596 (2015).
52. Schmidt-Nielsen, K. *Animal Physiology: Adaptation and Environment* (Cambridge University Press, 1997).
53. Eckert, R. *et al. Animal Physiology: Mechanisms and Adaptations* 3rd edn. (WH Freeman & Co., 1988).
54. Miller, J., Petersen, R. C., Metter, E., Millikan, C. & Yanagihara, T. Transient global amnesia: Clinical characteristics and prognosis. *Neurol.* **37**, 733–733 (1987).
55. Kirkpatrick, J. *et al.* Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci.* **114**, 3521–3526 (2017).
56. Hadsell, R., Rao, D. & Rusu, A. A. *& Pascanu, R* (Continual learning in deep neural networks. Trends cognitive sciences, Embracing change, 2020).
57. Hayes, T. L. *et al.* Replay in deep learning: Current approaches and missing biological elements. arXiv preprint arXiv:2104.04132 (2021).
58. Recanatesi, S., Katkov, M., Romani, S. & Tsodyks, M. Neural network model of memory retrieval. *Front. Comput. Neurosci.* **9**, 149. https://doi.org/10.3389/fncom.2015.00149 (2015).
59. Rubin, D. C., Berntsen, D. & Bohni, M. K. A memory-based model of posttraumatic stress disorder: evaluating basic assumptions underlying the ptsd diagnosis. *Psychol. Rev.* **115**, 985 (2008).
60. van Marle, H. PTSD as a memory disorder. *Eur. J. Psychotraumatol.* **6**, 27633 (2015).

Acknowledgements

Author contributions

N.S., H.S. and G.K. conceived the research, all authors took part in discussions. N.S. and H.S. performed the analytical calculations, N.S., H.S. and J.C. performed the numerical simulations and prepared the figures. N.S. and H.S. wrote the paper. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-16407-9.

Correspondence and requests for materials should be addressed to H.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.