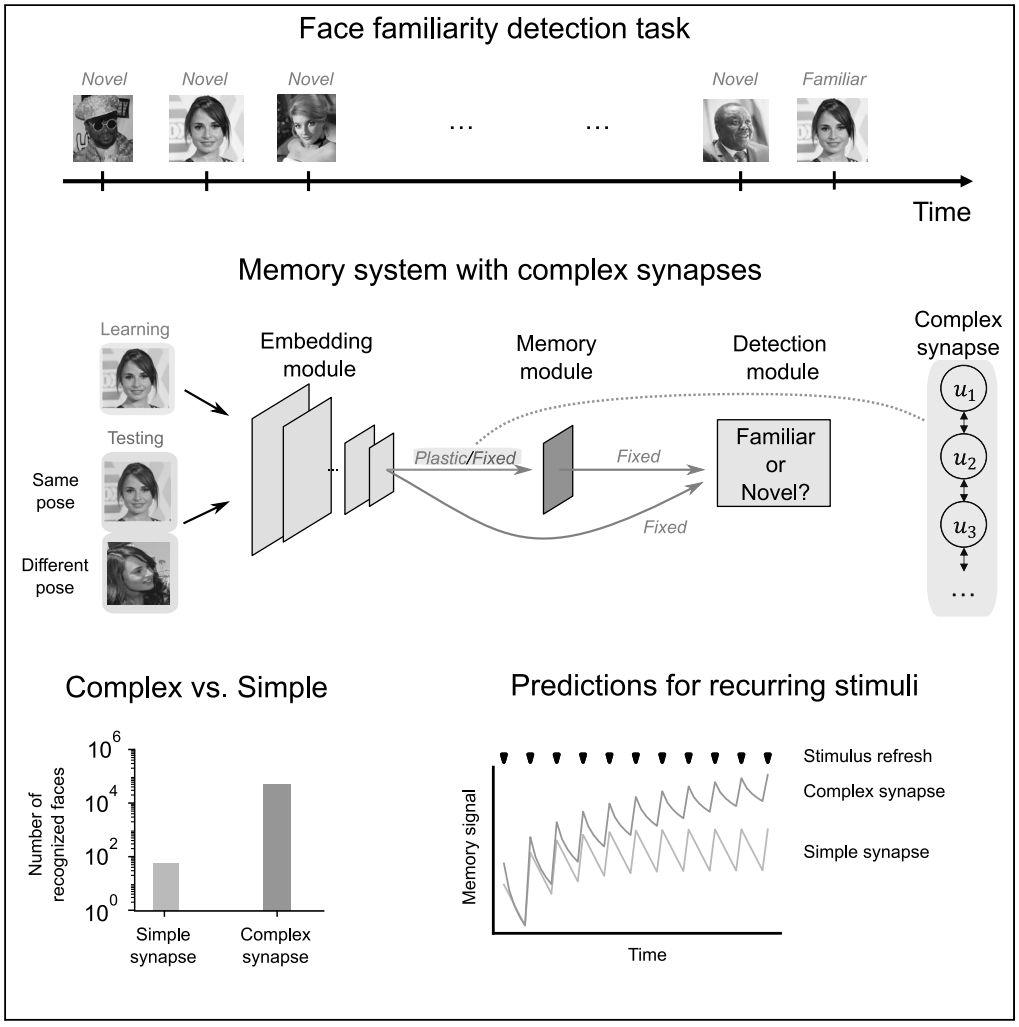
iScience

CellPress
OPEN ACCESS

Article

Face familiarity detection with complex synapses



Li Ji-An, Fabio Stefanini, Marcus K. Benna, Stefano Fusi

mbenna@ucsd.edu (M.K.B.)
sf2237@columbia.edu (S.F.)

Highlights

A memory system with complex synapses can recognize a large number of faces

The number of recognizable faces grows almost as the square of the number of neurons

Memory systems with complex synapses outperform those with simple synapses

Complex synapses have distinctive signatures that are testable in experiments

Ji-An et al., iScience 26, 105856
January 20, 2023
https://doi.org/10.1016/j.isci.2022.105856

iScience

CellPress
OPEN ACCESS

Article

Face familiarity detection with complex synapses

Li Ji-An,[1,2] Fabio Stefanini,[1] Marcus K. Benna,[1,3,4,*] and Stefano Fusi[1,4,5,*]

SUMMARY

Synaptic plasticity is a complex phenomenon involving multiple biochemical processes that operate on different timescales. Complexity can greatly increase memory capacity when the variables characterizing the synaptic dynamics have limited precision, as shown in simple memory retrieval problems involving random patterns. Here we turn to a real-world problem, face familiarity detection, and we show that synaptic complexity can be harnessed to store in memory a large number of faces that can be recognized at a later time. The number of recognizable faces grows almost linearly with the number of synapses and quadratically with the number of neurons. Complex synapses outperform simple ones characterized by a single variable, even when the total number of dynamical variables is matched. Complex and simple synapses have distinct signatures that are testable in experiments. Our results indicate that a system with complex synapses can be used in real-world tasks such as face familiarity detection.

INTRODUCTION

Synaptic memory is a complex phenomenon, which involves intricate networks of diverse biochemical processes that operate on different timescales. We recently showed that this complexity can be harnessed to greatly increase the memory capacity[1,2] in situations in which the synaptic weights are stored with limited precision. More specifically, we proposed a complex synaptic model which is characterized by m dynamical variables. These variables might correspond to different biochemical processes that operate on different timescales. If the interactions between these processes are properly tuned, the memory capacity of a population of synapses, estimated by an ideal observer who has access to all the synaptic weights, can increase almost linearly with its size (i.e., the number of synapses N_{syn}), even when both m and the number of states of each variable grow no faster than logarithmically with N_{syn}. This is the optimal scaling under some conditions (see[2]) and significantly better than what can be achieved by employing simple synapses characterized by a single variable.[3–5]

These previous studies on complex synapses focused on the problem of storing a large number of random and uncorrelated memories. Only recently, complex synapses started to be employed in more realistic problems (e.g., see[6]) in which memories are structured and correlated. Here we show that synaptic complexity can be important also in a real-world problem, face familiarity detection. The task is particularly difficult because we require that each face is presented only once (one-shot learning) and it has to remain recognizable for a long time. Moreover, each face is required to be recognizable even when a different pose is used as a memory cue. This is a typical situation in which a proper pre-processing of the visual stimuli combined with the complexity of the synapses can lead to a significant advantage in terms of memory capacity. The images of the faces that we used in our simulations are pre-processed by a simulated visual system which has been trained to report the identity of the person portrayed in the image. We then extracted the principal components (which can also be implemented by a neural network, see e.g.[7]) and binarized these representations. The pre-processed representations of different faces are approximately decorrelated, although a downstream readout can still retain the ability to generalize to different poses (i.e., different poses of the same face have similar representations). The decorrelation step is important to make the representations suitable for the memory system that stores the information about face familiarity. Modeling this process of "recoding" is of fundamental importance and it has been the subject of several studies which started with the work of David Marr in the 70s[8] and continued in the 80s and in the 90s with the first memory models of the hippocampus.[9–13] In these models the representations of memories are first orthogonalized to become more separable and hence facilitate the storage and reconstruction of memories. This orthogonalization process can be explicitly modeled as a process of

[1]Zuckerman Institute, Columbia University, New York, NY 10027, USA

[2]Neurosciences Graduate Program, University of California San Diego, La Jolla, CA 92093, USA

[3]Department of Neurobiology, School of Biological Sciences, University of California San Diego, La Jolla, CA 92093, USA

[4]Senior author

[5]Lead contact

*Correspondence: mbenna@ucsd.edu (M.K.B.), sf2237@columbia.edu (S.F.)

https://doi.org/10.1016/j.isci.2022.105856

compression[10,14–19], which leads to the most efficient decorrelated representations for memory storage. Compression is also an important underlying principle of several computational processes.[20,21]

The pre-processed representations are then stored in a neural circuit that contains the complex synapses proposed in.[2] These are characterized by dynamical variables that operate on multiple timescales. The fast ones can rapidly store information about a new visual stimulus such as a face, even when the stimulus is shown only once. This information is then progressively transferred to the slow variables, which can retain it for a long time. Because of these slow variables, which influence the synaptic efficacy, the older memories are protected from overwriting due to the storage of new faces. Synapses that are described by a single dynamical variable can either learn quickly if they are fast, but then they also forget quickly, or they can retain memories for a long time if they are slow, but then they cannot learn in one shot and require multiple exposures to the same face. This plasticity-rigidity dilemma concerns a very broad class of realistic synaptic models whose dynamical variables have a limited precision.[3,5,22]

Our memory benchmark for the complex synapses, familiarity detection (sometimes called familiarity discrimination or novelty detection), is an important component of recognition memory, which has been widely studied in humans and in animals. In particular, familiarity detection refers to the ability to rapidly memorize new items and report at a later time whether we have encountered them or not. In the case of faces, we would report that a face of a person is familiar if we experience the sense that we have already encountered that person in the past. The second component of recognition memory is recollection, which corresponds to the retrieval of the details of the individual (e.g. the name) and the episodic memories associated with that person. We can often experience a sense of familiarity without being able to recollect the details about an encountered individual. Familiarity detection, which is the focus of this article, has been studied in the famous and remarkable experiment by Standing,[23] in which he showed that it is possible to recognize a surprisingly large proportion of 10,000 images that are flashed on a screen only once and for a brief time. The subjects were asked whether they had seen an image or not, which is one way of assessing the familiarity of an image. Although familiarity detection is only one component of recognition memory, in the article we will use the verb 'recognize' to indicate the ability of a subject to report whether a visual stimulus had already been seen or not. The result of the Standing experiment is even more remarkable when one considers that more recent studies proved that subjects could memorize many details about each image.[24,25]

The neural substrate of recognition memory is unknown, although multiple lesion studies indicate that the hippocampus and perirhinal cortex play an important role.[26–28] The role of each area is controversial as for some investigators both the hippocampus and the perirhinal cortex contribute to recollection (memory retrieval) and familiarity[28,29] and for others the hippocampus supports recollection only, and perirhinal cortex supports familiarity.[26,27] One of the problems in the interpretation of these studies is that it is difficult to separate the contribution that each area gives to familiarity and recollection because when a memory can be recollected it can always be recognized. Another problem is that the role of these two areas differs depending on the nature of the memories (e.g., recognition of novel faces is intact in patients with lesioned hippocampus at a short retention interval, instead recognition memory for words, buildings, inverted faces, and famous faces is impaired[30]), on the length of the retention interval (for intervals of a few minutes or longer the hippocampus is certainly important for familiarity[28,31]) and on whether the memory is presented in a particular context or in isolation (perirhinal cortex is more important for the recognition of items in isolation whereas the hippocampus is more important when there is a contextual or associational component[26]). In the Discussion we will describe a possible interpretation of our model.

There are several biology-inspired computational models studying different aspects of recognition memory: some neural network models following the complementary learning systems approach were proposed to tease apart the hippocampal and neocortical contributions to recognition memory[32,33]; other models were concerned with the synaptic plasticity (learning) rules in the perirhinal cortex.[34] Finally, there are models that stress the distinct roles of familiarity and recollection in retrieving memories.[35]

Analytical estimates of familiarity memory capacity showed that in the case of random uncorrelated patterns, the number of memories that can be correctly recognized as familiar can scale quadratically with the number of neurons N in a recurrent network.[36] Not too surprisingly, this is a much better scaling than the linear scaling of the Hopfield model,[37] in which random memories are actually reconstructed

(see also the Discussion). The scaling for memory reconstruction is markedly worse and can be as low as $\sqrt{N}$ when the patterns representing the memories are correlated.[34] These computational models can replicate some interesting aspects of experiments on the capacity of human recognition memory.[38]

We constructed a model for recognition memory that incorporates complex synapses characterized by variables that have limited dynamical range (number of distinguishable states). We show that a simple neural circuit designed to reconstruct the memorized face can take advantage of the complexity of synapses and can efficiently store a large number of faces. In particular, we show that the number of faces that can be successfully recognized as familiar scales approximately quadratically with the number of neurons, or linearly with the number of synapses. This is the same scaling achieved in[36], in which synaptic weights could be stored with unlimited precision. Moreover, this scaling is similar to the one predicted for random patterns in[2], despite the fact that our pre-processing system does not completely decorrelate the patterns that represent different faces. Importantly, the network can recognize a face even when it is presented in a different pose, and the scaling is only slightly worse than in the case in which the exact same picture of the face is presented for familiarity testing. This ability to generalize is a distinctive feature of recognition memory, it is observed in experiments and it plays an essential role in any machine learning system that relies on novelty signals to speed up learning.[39] We then compared the performance of the recognition system with complex synapses to one with the same architecture but with simple synapses characterized by a single dynamical variable. The number of synapses is chosen so that the total number of synaptic variables would be the same in the two systems and of course the pre-processing system is exactly the same in the two cases. We show that the system with complex synapses outperforms the one with simple synapses, indicating that complexity provides the neural system with a clear computational advantage.

RESULTS

Face familiarity detection system

Our face familiarity detection system consists of three modules: an input (embedding) module, a memory module, and a readout (detection) module (see Figure 1A and model details in "face familiarity detection system" in STAR Methods and Table 1 summarizing the notations).

The embedding module consists of a deep convolutional neural network (pre-trained on several face tasks), taking pre-processed face images from VGGFace2[40] as inputs (see "face data set" in STAR Methods). The activity of the penultimate layer (adjacent to the classification layer) was extracted, further decorrelated (principal component analysis), and binarized. The top N ($N \leq 2048$) binarized principal components are taken as the binary face pattern $x = [x_1, \ldots, x_N]^T$, serving as the activity of the N input neurons for the memory module. Despite the similarities between these binary face patterns and random unstructured binary patterns, we found non-trivial high-order statistics in these face patterns (see STAR Methods "statistical differences between binary face patterns and random patterns" and Figure S1).

The memory module is the only part of our network containing plastic synapses. The synapses are continuously updated by the ongoing presentation of the face patterns, whereas the weights of the input module are frozen during online learning. The memory module consists of N memory neurons, one for each input neuron in the embedding module (see Figure 1B). The j-th input neuron connects to the i-th memory neuron (for $i \neq j$) with synaptic weight (efficacy) w_{ij} and bias term b_i. There is no connection between the i-th input neuron and the i-th memory neuron for any i (i.e., $w_{ii} = 0 \forall i$). This plastic layer of synapses implements a simple feedforward memory model that can perform an approximate one-step reconstruction of a stored input pattern from a noisy cue at test time and we denote the binary memory patterns retrieved (reconstructed) in this manner as $y = [y_1, \ldots, y_N]^T$. Because the i-th memory neuron y_i is expected to reconstruct the i-th input neuron x_i, we set the value of the i-th memory neuron to be x_i during learning. The synaptic weights and biases are updated with the Hebbian learning rule with bounded dynamical ranges. For each synapse (i.e., for each weight w and bias term b), we implemented a complex synaptic model[2] with m discretized dynamical variables $u_1, \ldots, u_m$ in discrete time. Here m denotes the total number of dynamical variables per synapse (a measure of synaptic complexity), each of which operates on a different timescale.

The readout (detection) module compares the output $x = [x_1, \ldots, x_N]^T$ of the embedding module and the output $y = [y_1, \ldots, y_N]^T$ of the memory module to assess the level of familiarity of a given pattern.

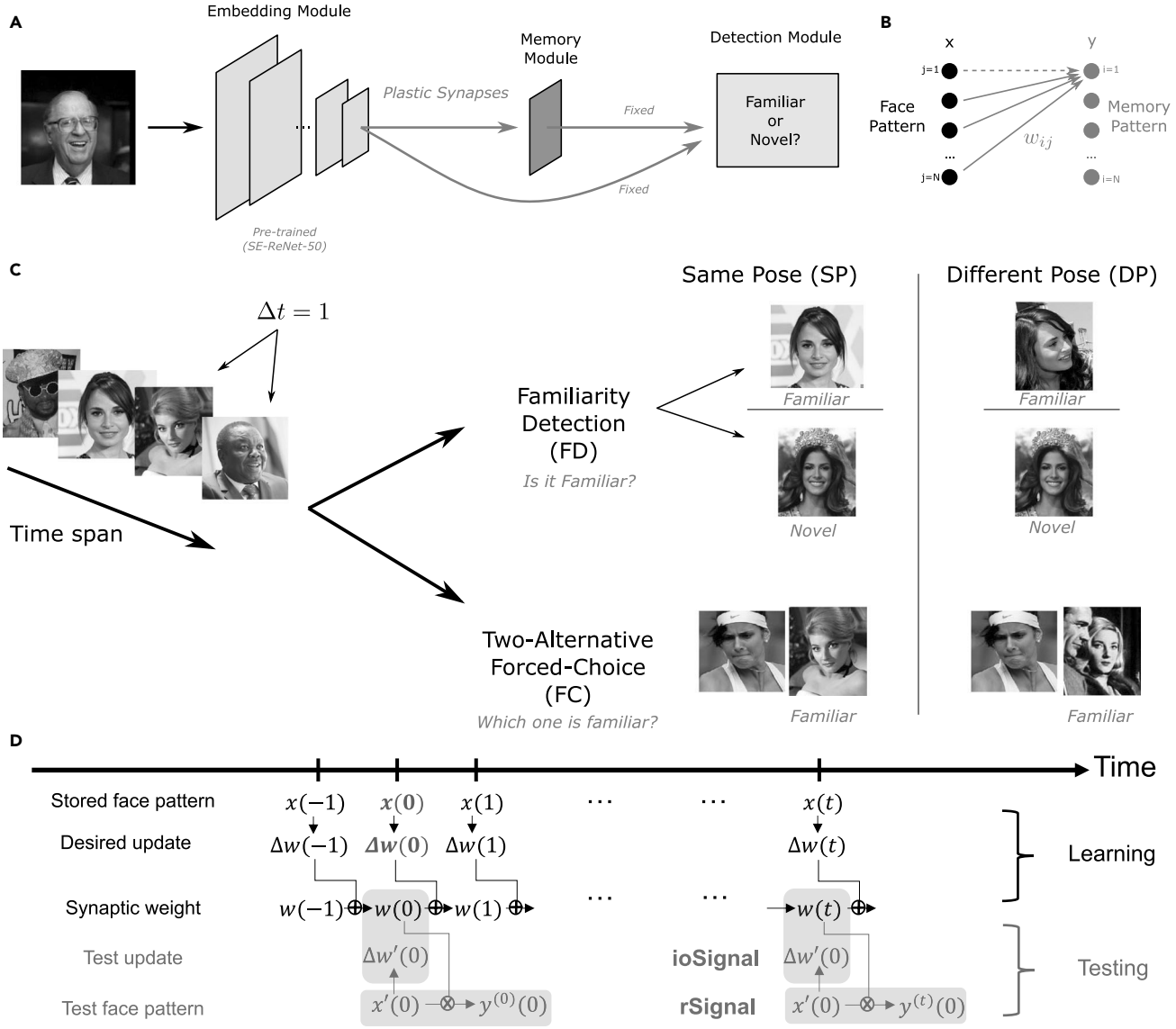


Figure 1. The architecture of our face familiarity detection system and the task diagram

(A) The neural system contains three modules: the input (embedding) module, the memory module, and the readout (detection) module. The synapses between the embedding module and the memory module (as well as the biases in the memory module) are plastic, while all other synapses are fixed (after being either set by hand or pre-trained), which requires online learning of face patterns.

(B) The plastic connections between the input neurons (encoding face patterns) in the embedding module and the memory neurons (encoding memory patterns) in the memory module.

(C) A series of face images are presented to the neural system. In each familiarity detection (FD) test, the system is required to determine whether a presented face is familiar or unseen. A face is considered familiar if the test image is identical to a previously presented one (i.e., the same pose, SP) or a new pose of a previously presented face (i.e., a different pose, DP), and is considered novel if it is an image of an unseen person's face. In each two-alternative forced-choice (FC) test, the neural system is presented with a pair of face images (exactly one familiar and one unseen), and is required to choose which one of the two is familiar.

(D) During learning, face patterns $x(\cdot)$ are stored in the synaptic weights via the desired weight update $\Delta w(\cdot)$ generated from the Hebbian learning rule. When we test the face stored at time 0, the pattern $x'(0)$ (either a noisy version of $x(0)$ in the DP case or $x(0)$ itself in the SP case) is presented to the system at time t. The ioSignal is the overlap between the synaptic weight $w(t)$ at time t and the test weight update $\Delta w'(0)$ (even though the synaptic weight is never actually changed by the test weight update). The rSignal is the overlap between the test face pattern $x'(0)$ and the corresponding memory pattern $y^{(t)}(0)$ reconstructed using the current synaptic weight $w(t)$.

To evaluate the memory performance of our familiarity detection system we studied how the number of faces that can be recognized scales with the number of neurons N. The memory capacity of neural systems always increases with the size of the network (as the number of neurons increases also the number of synapses

Table 1. Notations used in the text

Type	Notation	Explanation	Value
Hyperparameter	α	const. determining overall timescale of model dynamics	0.25
	n_0	const. related to the ratio of timescales of successive variables	2
	V	maximal value of the discrete levels of dynamical variables	31^2
	D	number of discrete levels of dynamical variables	32
	N_b	number of bits of dynamical variables	5
Model simulation	N	number of neurons	16 to 2048
	m	synaptic complexity (number of dynamical variables)	1 to 10
	q	synaptic learning rate (encoding probability)	0.01, 0.128, 1
Model parameter	w_{ij}	synaptic weight	–
	b_i	synaptic bias	–
Model variable	x	input pattern	–
	y	memory pattern	–
	I	desirable update imposed by the input pattern	–
	u_k	dynamical variable	–
Memory metric	S or $\mathcal{S}$	memory signal	–
	$\mathcal{N}$	memory noise	–
	$\mathcal{S}/\mathcal{N}$	memory signal-to-noise ratio	–
	r	idealized memory signal	–
	T	longest timescale of the synapse	–
	t^*	memory lifetime	–
Learning schedule	n	interval number	–
	γ	length of the first interval	–

increases), but the growth can vary in a wide range, from a very inefficient logarithmic scaling[3] with N to a quadratic dependence. For networks with complex synapses, the memory capacity depends also on the number of dynamical variables m per synapse (synaptic complexity), and it is important to scale m up when N increases. If m is fixed, then the memory capacity can increase rapidly with N (e.g. quadratically), but only to some value determined by m. Beyond that value, the increase is only logarithmic. Fortunately, a modest increase in m allows us to rapidly (exponentially) increase this critical value. To take advantage of a larger population of neurons, it is important to increase the longest timescale of the synapses, which is related to its complexity m. This can be achieved by choosing an m that grows logarithmically with N (such that $m = \log_2 N - 1$, as suggested in[2]). We present the results of varying m and N separately in Figures S5 and S6.

In the following simulations, all memory metrics, including the signal-to-noise ratio (SNR) and the task performance, are evaluated after the neural system reaches its steady state, i.e., when a large number of face patterns (with constant input statistics) have already been stored. In the steady state, the distribution of synaptic weights does not change any longer,[2] although synapses continue to be updated as new face images are memorized. The system is then presented with two thousand real face images from different people, interleaved with the necessary number of non-evaluated synthesized face image patterns (see "synthesizing artificial face patterns" in STAR Methods). We showed that the effect of interleaving with synthesized face patterns was indistinguishable from interleaving with real face patterns (see STAR Methods

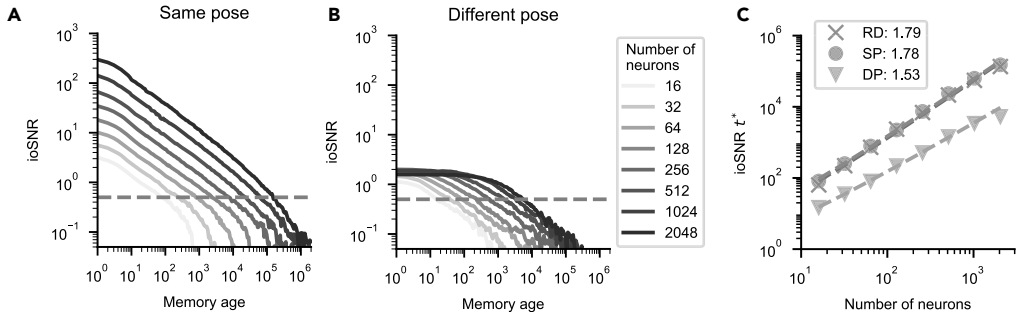


Figure 2. Ideal observer SNR (ioSNR) of the memory module as a function of face memory age and its scaling properties

(A) Doubly logarithmic plots of ioSNR versus the number of subsequently stored memories. Different curves correspond to models with a different number N of memory neurons and m of dynamical variables in the same pose (SP) case. The parameters N and m are varied by increasing N by factors of two and setting $m = \log_2 N - 1$.

(B) The same as in the previous panel, but in the different pose (DP) case.

(C) Log-Log plot of the ioSNR memory lifetime versus N in the SP, DP, and random-pattern (RD) cases. The legend indicates the best fit linear regression slopes (corresponding to the power of base N in the scaling behavior with the logarithmic correction; data points deviating from the regression line due to saturation were excluded).

"interleaving with synthesized and real face patterns" and Figure S2). Our memory metrics were evaluated only over these real face images and further averaged over independent simulations to reduce the noise floor. We also quantified the variability of our results by first computing the metrics in each independent sequence and then considering their variations across sequences (see Figure S7).

SNR analysis and memory performance

To measure the strength of a memory we took the perspective of an ideal observer, who has direct access to all the synaptic weights[1,2,22] (see also "evaluating the memory signal and noise" in STAR Methods) and can compare them to the synaptic modifications Δw induced by the particular memory that we are tracking. The more similar (correlated) the current weights are to the Δw, the stronger the memory signal. The way this similarity is computed is illustrated in Figure 1D: say that the memory that we intend to track is memorized at time 0. At that time a pattern $x(0)$ is imposed on the input, determining through a simple Hebbian rule the $\Delta w(0)$, which is then used to update the synapses, leading to $w(1)$. Note that $\Delta w(0)$ is only the desired update as the final new synaptic state will depend on the complex internal dynamics of the synapse (i.e., $w(1)$ is not necessarily $w(0) + \Delta w(0)$). At a later time, say time t, we can test the memory stored at time 0. As a test face, we considered both faces in the same pose (SP) as those stored in memory and faces in a different pose (DP) (see Figures 1C and 2). In the first case, we would simply compare $w(t)$ to $\Delta w(0)$. In the DP case, $\Delta w'(0)$ is computed from a face pattern $x'(0)$ that is not in the same pose as the memorized face. The similarity (correlation) between $\Delta w'(0)$ and $w(t)$ is defined as the ideal observer signal (ioSignal). We computed the average of the signal $\mathcal{S}_{io}$ and the noise $\mathcal{N}_{io}$ (SD of the signal) across the full temporal series of the different faces. The ideal observer signal-to-noise ratio (ioSNR) $\mathcal{S}_{io}/\mathcal{N}_{io}(\Delta t)$ is our first measure of memory strength and it depends on the age of the memory Δt.

We also estimated the ability to reconstruct a memorized face by computing the rSignal, which is the overlap between the test face pattern $x'(0)$ and the corresponding memory pattern $y^{(t)}(0)$ (reconstructed from $x'(0)$ using the current synaptic weight $w(t)$). For $t \leq 0$, ioSignal and rSignal are approximately zero because $w(t)$ has not been updated by $\Delta w(0)$ yet. The ioSignal and rSignal will reach their maximum at $t = 1$, and gradually decrease as time elapses.

The ioSNR critically depends on the number of memories that are stored after the tracked face pattern, i.e., the memory age. Different curves in Figures 2A and 2B correspond to synaptic models with different numbers of input neurons (and memory neurons) N and dynamical variables m. The curves are plotted on a log-log scale, for which a straight line represents a power-law dependence.

In the SP case, the ioSNR curves decay as a power-law over a time interval T corresponding to the longest timescale of the synapse before the decay becomes exponential. The ioSNR decays as slowly as the inverse

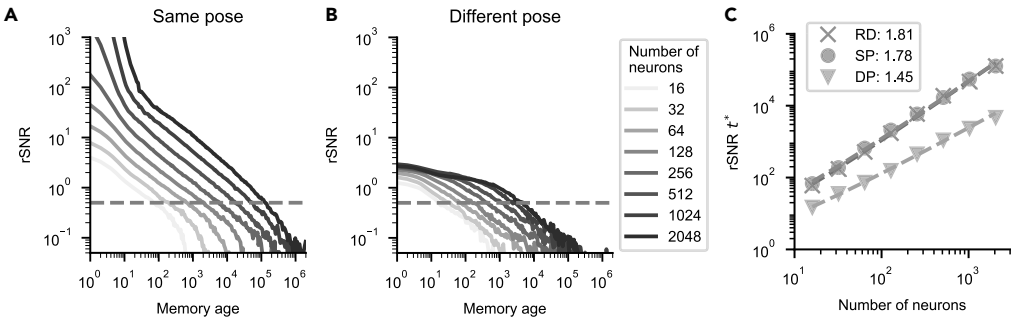


Figure 3. Readout SNR (rSNR) of the memory module as a function of face memory age and its scaling properties

(A) Doubly logarithmic plots of rSNR versus the number of subsequently stored memories. Different curves correspond to models with a different number N of memory neurons and m of dynamical variables in the same pose (SP) case. The parameters N and m are varied by increasing N by factors of two and setting $m = \log_2 N - 1$.

(B) The same as in the previous panel, but in the different pose (DP) case.

(C) Log-Log plot of the rSNR memory lifetime versus N in the SP, DP, and random-pattern (RD) cases. The legend indicates the best fit linear regression slopes (corresponding to the power of base N in the scaling behavior with the logarithmic correction; data points deviating from the regression line due to saturation were excluded).

square root of the memory age in the power-law regime. Changing N shifts the ioSNR curves in the log-log plot vertically, while increasing m primarily extends the power-law regime (i.e., increases T; see Figure 2A). We determined the scaling of the familiarity memory lifetime with N (and m), where the lifetime t^*_{ioSNR} is represented by the memory age at which the ioSNR first drops below a given threshold. A value of 1 corresponds to a situation where the signal and the noise are of the same intensity. We chose a threshold of 0.5, though its precise value does not affect the scaling behavior much. We found that the familiarity memory lifetime scales approximately as N^2 (see Figure 2C, in which the linear regression slope on a log-log scale is about 1.78 for the SP case, compared to 1.79 for random patterns (RD)). This scaling is very close to the theoretical result for optimal storage of random unstructured patterns.[2] Because m increases together with N (logarithmically), the familiarity memory lifetime scales exponentially with m (with the same linear regression slope on a $\log_2$-linear plot of t^*_{ioSNR} versus m).

For the DP case (see Figure 2B), the ioSNR curves are lower than those in the SP case, due to the differences between the memorized and the tested face patterns. When there are more memory neurons, the shape of its initial decay with memory age becomes flatter. The initial ioSNR increases slowly with N for $N < 512$, and then drops a little for larger N, because compared with the earlier features, the later features are much less correlated between poses of the same person. Nevertheless, the familiarity memory capacity still scales as a power of N: the regression slope is 1.53 (the model with 2048 neurons is removed from linear regression due to saturation).

As mentioned above, we also considered another measure of the memory signal that is more directly related to the ability of the system to reconstruct the stored memory, the rSignal. We then studied the readout signal-to-noise ratio (rSNR), defined similarly to the ioSNR (see Figure 3). We found that the rSNR behaves similarly to the ioSNR at long time lags, but deviates from it for small memory ages, reflecting the effect of the neuronal nonlinearity (i.e., the nonlinear activation function). This nonlinear effect, which becomes more significant for larger N or smaller m (see also Figure S8), leads to larger initial rSNR values, but does not substantially affect the memory lifetime t^*_{rSNR} (similarly defined as the memory age at which the rSNR first drops below a given threshold) compared to the ioSNR measure. The initial SNR enhancement quickly attenuates, leading to a similar scaling for t^*_{rSNR}. These results further validate the ideal observer approach.

Task protocol and performance

To evaluate the task performance of our system, we considered two tests in which we presented a series of pre-processed face images to the neural system and tested its memory on randomly chosen faces (see Figure 1C). These tasks are made particularly challenging by the fact that the familiar faces are presented only once.

In the first familiarity detection (FD) test, the neural system is required to determine whether the face image presented at test time is familiar (either SP or DP of a previously presented face image) or unseen (an image

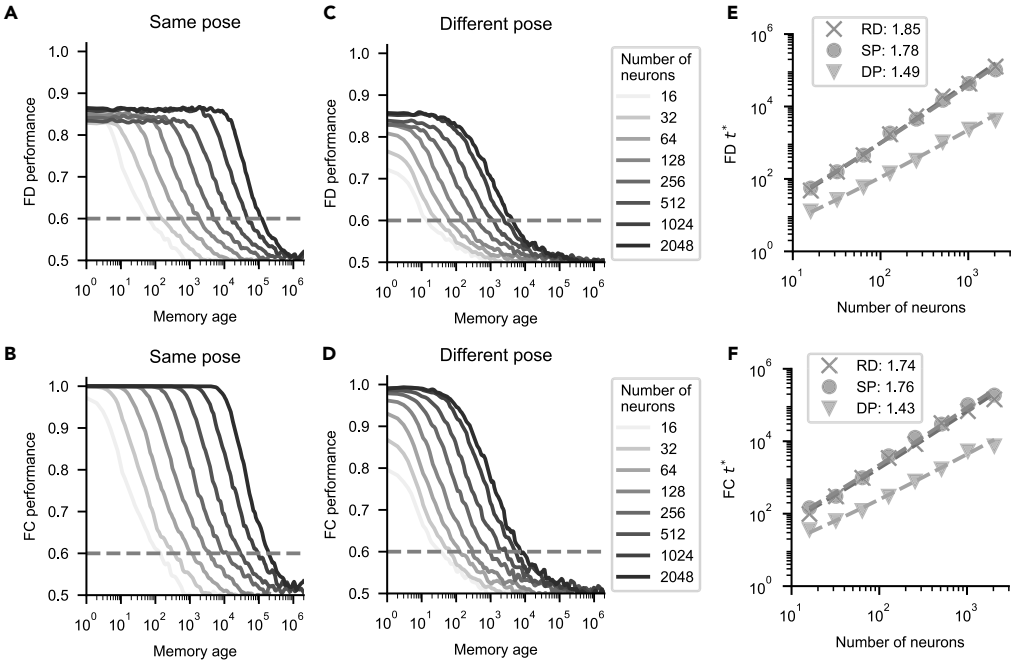


Figure 4. Familiarity detection (FD) and two-alternative forced-choice (FC) test performance of our system and their scaling properties

(A and B) Task performance as a function of the memory age. Different curves correspond to models with a different number N of memory neurons (and number m of dynamical variables such that $m = \log_2 N - 1$) in the same pose (SP) case.

(C and D) As in the previous panels, but for the different pose (DP) case.

(E and F) FD and FC memory lifetimes versus N in the SP, DP, and random-pattern (RD) cases. The legend indicates the best fit linear regression slopes (corresponding to the power of base N in the scaling behavior with the logarithmic correction; data points deviating from the regression line due to saturation were excluded). Also see Figures S3 and S4.

of an unseen person) by comparing the output of the detection module to a threshold. Here, the face images presented to the system are balanced, i.e., familiar faces previously presented within a certain age-range and unseen faces appear at test time with equal probability. The threshold on the overlap did not depend on the age of face, which is assumed to be unknown at test time, and it was optimized to best separate familiar from novel faces for all ages shorter than the memory lifetime t^*_{ioSNR} (for details see STAR Methods "choosing the optimal threshold in the FD task" and Figure S3). Note that t^*_{ioSNR}, and hence the threshold, will depend on the number of neurons and the complexity of the synapses.

We now define t^*_{FD} as the age at which the FD test performance drops below some threshold. In Figure 4, we plot the task performance in the FD test as a function of memory age. In the SP case, increasing N and m leads to a substantial extension of the task-relevant familiarity memory lifetime t^*_{FD} (see Figure 4A). The memory lifetime was estimated assuming a performance threshold of 60% (this value was chosen to keep the initial task performance of all the simulations above the threshold). The power-law scaling behavior of the familiarity memory lifetime is revealed by plotting t^*_{FD} versus N on a log-log scale (linear regression slope 1.78; see Figure 4E), which shows a very similar growth also in the RD case (linear regression slope 1.85). The initial task performance cannot reach 100% because each model optimized for the model-specific age-range has a non-zero constant error rate for unseen faces, even if the true positive rate (accuracy for familiar faces) saturates at 100% (also see Figure S4). As expected, in the DP case the task performance is worse than in the SP case (see Figure 4C). However, we still found a reasonable power-law scaling with N (regression slope 1.49).

In the second two-alternative forced-choice (FC) test, the neural system is presented with a pair of face images containing one familiar (either SP or DP) and one unseen face, and is required to choose which one of the two is familiar by comparing the output of the detection module for the two faces. The task performance is defined as the probability of correctly choosing the familiar face (over the unseen one) for face

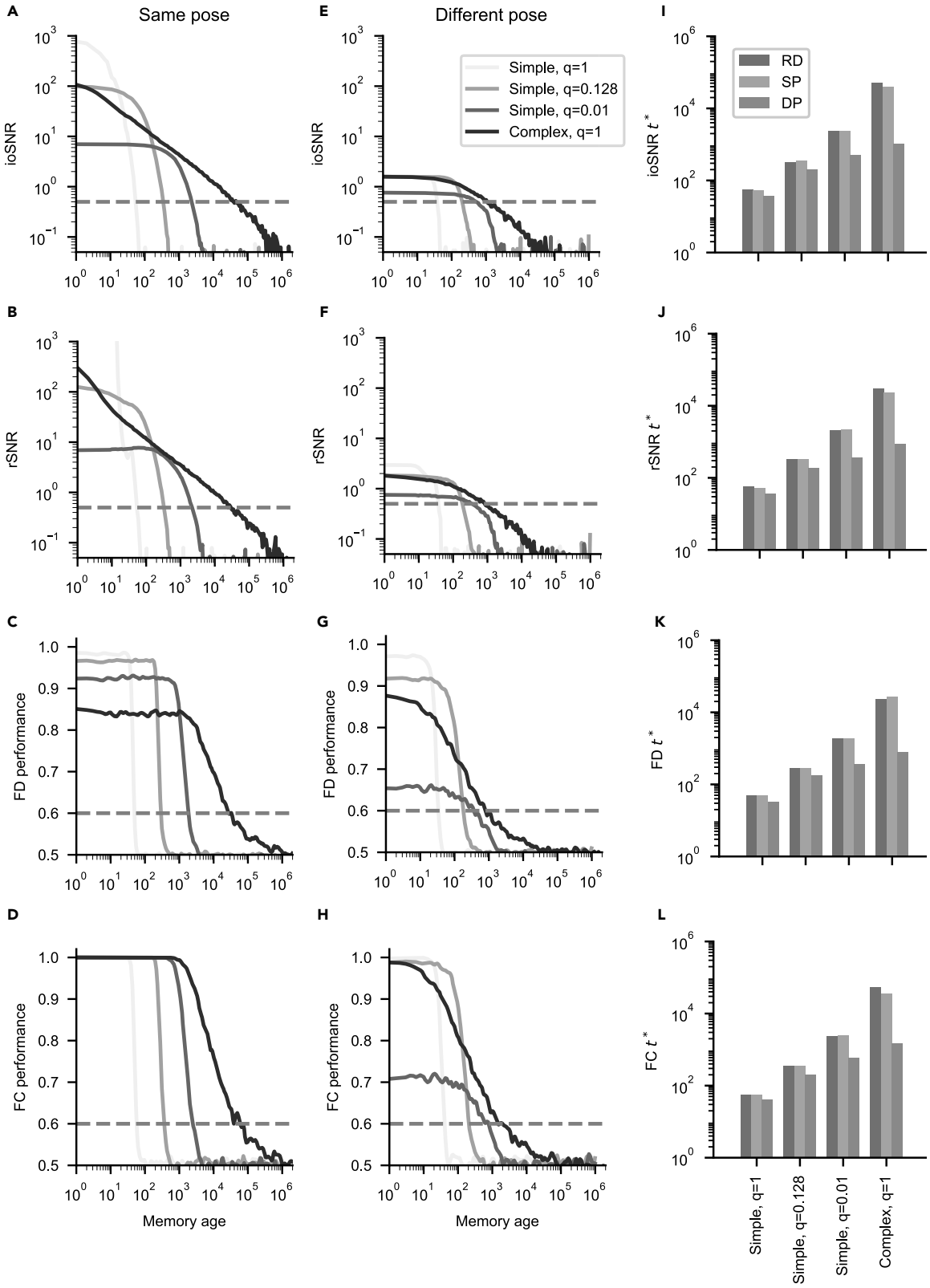

A
Same pose
E
Different pose
I
ioSNR
Simple, q=1
Simple, q=0.128
Simple, q=0.01
Complex, q=1
RD
SP
DP
ioSNR t^*
B
F
J
rSNR
rSNR t^*
C
G
K
FD performance
FD t^*
D
H
L
FC performance
FC t^*
Memory age
Simple, q=1
Simple, q=0.128
Simple, q=0.01
Complex, q=1

Figure 5. Complex and simple synapses

Comparison between models with simple synapses ($N = 2048$, $m = 1$) and different learning rates ($q = 1$, 0.128, and 0.01, respectively) and a complex model ($N = 2048$, $m = 10$, $q = 1$)that has the same total number of memory neurons and plastic variables, by keeping only 10% percent of the synapses of the fully connected simple models (i.e., 90% of the complex synapses are pruned).

(A–D) Comparisons between models for the same pose (SP) case in terms of ioSNR, rSNR, familiarity detection (FD) performance, and two-alternative forced-choice (FC) performance.

(E–H) Similar comparisons between models in the different pose (DP) case.

(I–L) Comparisons between models in terms of different measures of familiarity memory lifetime (t^*_{ioSNR}, t^*_{rSNR}, t^*_{FD}, and t^*_{FC}, respectively) in the SP, DP, and random-pattern (RD) cases. Also see Figure S8.

memories of different ages (see Figures 4B, 4D and 4F). The regression slope of the memory lifetime t^*_{FC} (defined as the age at which the FC test performance drops below some threshold) versus N on a log-log scale is 1.76 for the SP case and 1.43 for the DP case.

Complex versus simple synapses

To obtain a fair comparison between complex synapses[2] and the well-studied, simple (multi-state) synapses,[3–5] we evaluated the familiarity memory performance of a neural system with complex synapses and three models with simple synapses in which we matched the total number of dynamic variables. As the complex synapse has 10 times more dynamical variables than the simple synapse, we randomly pruned 90% of the complex synapses. In this way each memory neuron in the complex model has on average 204 incoming synapses (randomly sampled 10% from 2047 presynaptic input neurons) and 1 bias (i.e., $2048 * (204 + 1) * 10 = 4198400$ variables in total), whereas in the three simple models each neuron has 2047 incoming synapses (1 dynamical variable) and 1 bias (i.e., $2048 * 2047 * 1 = 4192256 \approx 4198400$ variables in total). The simple synapses follow essentially the same model dynamics as the previously studied hard-bounded multi-state synapses.[4] They differ in their level of plasticity: the synapses in the first model are updated every time an input pattern is stored, while the synapses in the second and third ones are changed stochastically according to a learning rate (encoding probability q) less than one, and thus are more "rigid".[3,41] Small learning rates lead to lower initial ioSNR values, but also to longer memory lifetimes. We choose $q = 0.128$ for the second model so that its initial ioSNR is comparable to the complex synapse system in the SP and DP cases. For the third model, we picked $q = 0.01$ to obtain the longest memory lifetimes possible for a system of simple synapses of this size, with an initial SNR just above the threshold (in the DP case).

Each variable in all of these models has the same number of discrete levels, and the total numbers of variables are approximately matched in the simple and the complex system. These simulations show that the complex system has a substantially better familiarity memory performance than the simpler systems (see Figure 5), despite the smaller number of synapses. For the SP and RD cases, the memory lifetime of the system with complex synapses is $\sim 400 - 900$ times longer; while for the DP case, the improvement factor is $\sim 20 - 40$. Slower simple synapses (with smaller q) can greatly extend the familiarity memory lifetime, but at the expense of the initial SNR and thus the generalization ability. Even so, they are far from matching the memory lifetime of the complex system. This clear advantage is further confirmed by another comparison, where we matched the number of dynamic variables using a different approach: We considered a larger network for the simple synapses (see Figure S8). We can conclude that the memory model with complex synapses performs at least two orders of magnitude better in terms of familiarity memory capacity, and we expect the gap between simple and complex systems to grow even wider in networks with a larger number of neurons because of the different scaling behaviors.

Testable predictions for simple and complex synapses

Simple and complex synapses exhibit quantitatively different SNR decays and memory performance. We now show that it is possible to design an experiment with a specific learning schedule that would reveal whether the synapses are complex or simple. The main idea is that memories can be repeatedly refreshed in such a way that the asymptotic minimum memory strength remains constant. Ideally, this could be implemented by monitoring the memory signal of a specific stimulus, and refreshing the memory by presenting the same stimulus again as soon as the signal drops below some threshold. Using this procedure, we obtain a refresh schedule, which can be described by specifying the intervals that separate two consecutive presentations of the same stimulus. Depending on the synaptic model, the length of these intervals will be different, and, most importantly, will change over time in a different way.

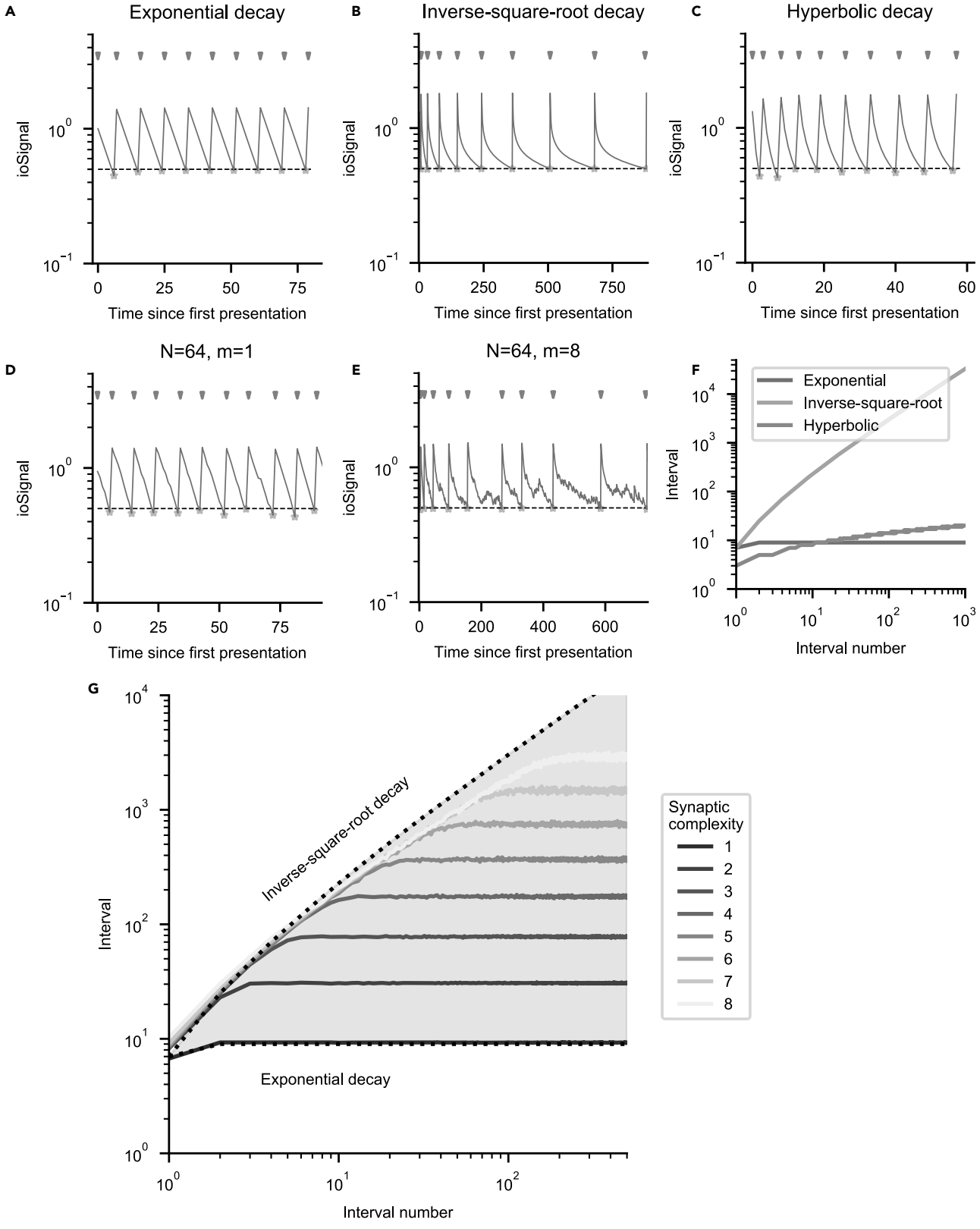

A
Exponential decay
B
Inverse-square-root decay
C
Hyperbolic decay
ioSignal
Time since first presentation
D
N=64, m=1
E
N=64, m=8
F
Exponential
Inverse-square-root
Hyperbolic
Interval
Interval number
G
Inverse-square-root decay
Exponential decay
Synaptic complexity
1
2
3
4
5
6
7
8

Figure 6. The optimal learning schedule, in which the pattern is presented whenever the monitored ioSignal drops below a pre-specified threshold (0.5 shown)

(A–C) The idealized exponential decay, inverse-square-root decay, hyperbolic decay models of ioSignal under the optimal learning schedule. The ioSignal decreases over time and is enhanced by pattern presentations indicated by red arrows. The signal strength immediately before each presentation is marked by orange stars.

(D) A typical noisy ioSignal trajectory from the model with simple synapses ($N = 64$, $m = 1$). The length of the interval between consecutive presentations is approximately constant, similar to the idealized exponential decay.

(E) A typical noisy ioSignal trajectory from the model with complex synapses ($N = 64$, $m = 8$). The length of the interval approximately increases linearly with interval numbers, similar to the idealized inverse-square-root decay.

(F) The length of the interval as a function of interval numbers for three idealized decay models.

(G) The length of the interval increases as a function of interval numbers under the optimal schedule, averaged over noisy ioSignal trajectories. Different colored curves correspond to models with a different synaptic complexity m. The shaded region is bounded by interval curves of the idealized exponential decay and inverse-square-root decay models. Also see Figure S9.

We illustrate this with three idealized models of memory traces through simulations: the exponential decay model (in which the signal decays as $e^{-t/\tau}$, where τ is the time constant), the inverse-square-root ($1/\sqrt{t}$) power-law decay, and the hyperbolic ($1/t$) power-law decay (which is achievable by a heterogeneous population of simple synapses[42]). See Figures 6A–6C and 6F. We set the threshold to $\theta = 0.5$, but its precise value does not affect the scaling behavior of the intervals. For the idealized exponential decay model, the length of the interval remains constant after the second presentation (the constant interval is $\tau \ln(1 + C/\theta)$, where C is the initial signal strength and θ is the pre-specified threshold). Of interest, the length of the interval increases asymptotically linearly for the idealized inverse-square-root decay model (the coefficient of this linear asymptotic growth is $\pi^2 C^2/2\theta^2$, see mathematical details in STAR Methods "asymptotic behavior of the optimal learning schedules for idealized synaptic models with specific decay kernels"). The situation for the hyperbolic decay is intermediate between the exponential and inverse-square-root decay models, showing an approximately logarithmic increase of the length of the interval.

The idealized exponential decay and inverse-square-root decay models represent very different degrees of synaptic complexity. The memory signal of complex synapses decays as an inverse-square-root power-law over the longest timescale of the synapse before the decay becomes exponential. Example memory signal trajectories under the optimal learning schedule for simulated models with different synaptic complexity are shown in Figures 6D and 6E. The length of the interval is approximately constant for the model with simple synapses ($m = 1$), similar to the idealized exponential decay, but increases linearly for the model with complex synapses ($m = 8$), similar to the idealized inverse-square-root decay. Averaged over multiple such noisy trajectories, the interval curves are plotted on a log-log scale as a function of interval numbers (see Figure 6G). Increasing the synaptic complexity m effectively extends the linear growth regime (corresponding to the inverse-square-root power-law decay regime of the ioSNR) and postpones the gradual transition into the constant interval regime (corresponding to the exponential decay regime of the ioSNR). The idealized inverse-square-root decay thus approximates the envelope of the interval curves of models of different synaptic complexity m. We further study the scaling properties of the length of the interval (Figure S9).

However, it would not be feasible in experiments to monitor the memory signal in real time. Indeed, to measure the signal we need to expose the subject to the memory we intend to test, and hence we are going to modify the memory signal we want to estimate. We propose that we can simply use either a constant interval or a presentation schedule with a linearly increasing one without monitoring the memory signal (between presentations). Both protocols will be parameterized by a single variable. Under the constant schedule, a specific memory is refreshed after an interval of a fixed length γ, whereas under the linear schedule, a specific memory is refreshed after a linearly increasing interval. In particular, the interval is equal to γn, where n is the interval number, and γ is the length of the first interval (see Figures 7A and 7B). These refreshes serve the dual roles of evaluating familiarity detection performance (e.g., by querying the subject whether the presented stimulus is familiar) and boosting memory strength (corresponding to the increase of the signal strength immediately after the refresh). Under the constant learning schedule, the performance of the exponential decay model will remain constant, consistent with the conclusion drawn from the optimal schedule. The inverse-square-root decay and the hyperbolic decay models will exhibit gradually improving performance. Under the linear learning schedule, the performance of exponential decay and hyperbolic decay models will quickly drop to chance level, but the inverse-square-root decay model will maintain its performance, as predicted by the optimal schedule.

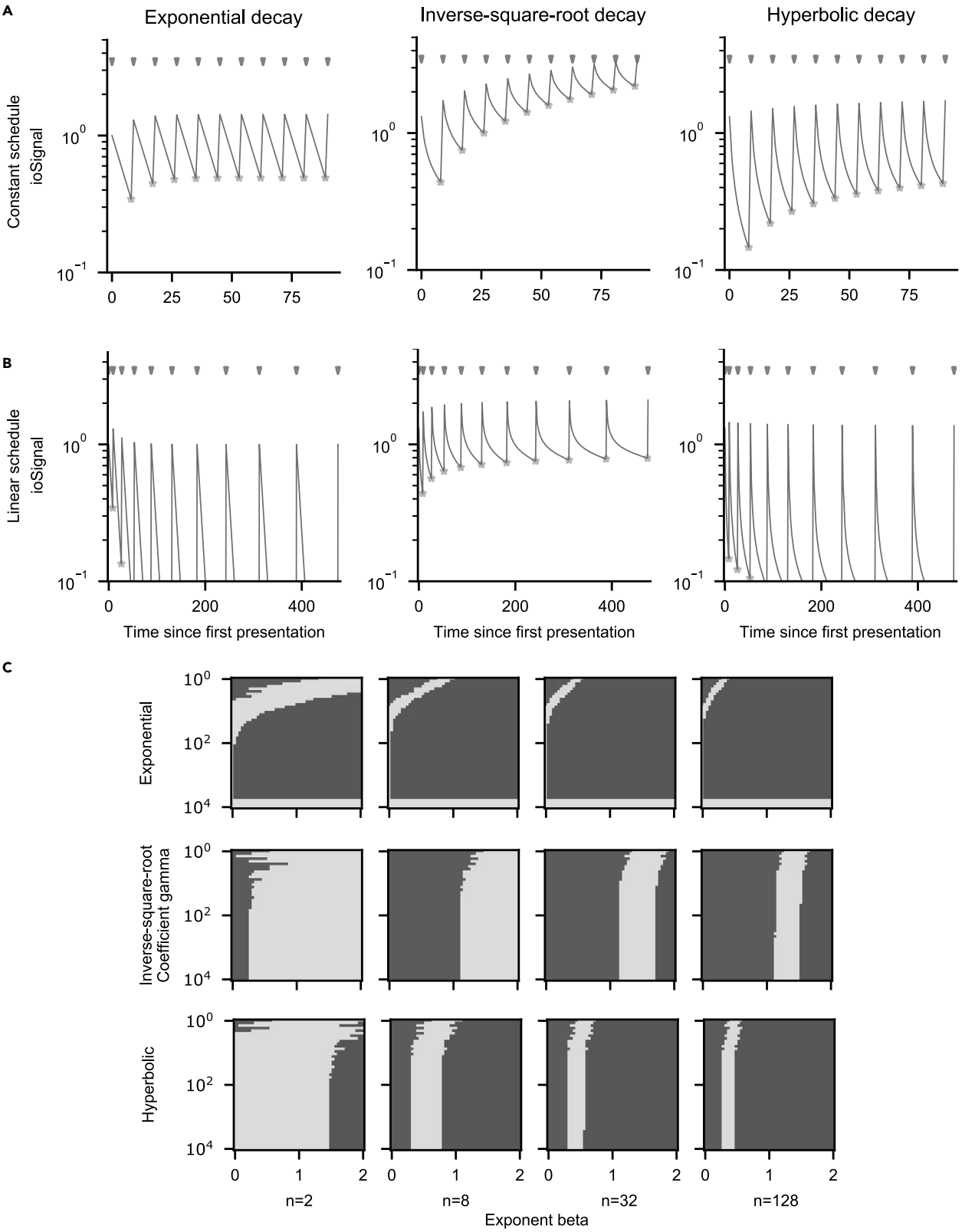

A
Exponential decay
Inverse-square-root decay
Hyperbolic decay
Constant schedule
ioSignal
B
Linear schedule
ioSignal
Time since first presentation
C
Exponential
Inverse-square-root
Coefficient gamma
Hyperbolic
n=2
n=8
n=32
n=128
Exponent beta

Figure 7. Pre-determined learning schedules

(A) The idealized exponential decay, inverse-square-root decay, hyperbolic decay models of ioSignal under the constant learning schedule, in which the pattern is presented each time after an interval of a pre-determined constant length. The ioSignal decreases over time and is enhanced by pattern presentations indicated by red arrows. The signal strength immediately before each presentation is marked by orange stars, reflecting familiarity task performance. In the following presentations, the task performance of the exponential decay remains constant, while the two power-law decay models' performance gradually increases.

(B) The idealized exponential decay, inverse-square-root decay, hyperbolic decay models of ioSignal under the linear learning schedule, in which the pattern is presented each time after an interval of linearly increasing length. In the following presentations, the inverse-square-root decay model maintains its performance, but the performance of exponential decay and hyperbolic decay models quickly drops to chance level (orange stars not shown due to extremely low signal strength).

(C) The signal gain as a function of γ (length of the first interval) and β (exponent of length increase) for the three idealized decay models under the general pre-determined learning schedule, where the length of the interval equals γn^{β} (n denotes the interval number, $10^0 \leq \gamma \leq 10^4$ on a log scale, $0 \leq \beta \leq 2$ on a linear scale). Red regime is shown for positive gain (>0.2), blue for negative gain (< -0.2), and gray for marginal gain (between -0.2 and 0.2). The three idealized decay models exhibit qualitatively different behaviors. Also see Figure S10.

We define the signal gain for interval number n as the logarithmic ratio of the ioSignal after the n-th interval (immediately before the $(n+1)$-th presentation) relative to the ioSignal after the first interval (immediately before the second presentation). Positive signal gains correspond to better familiarity detection performance, and negative ones indicate worse performance after the following presentations. The three idealized decay models demonstrate qualitatively different signal gains under the constant and linear schedules with varying γ (see Figure S10), offering testable predictions for experiments.

To better discriminate between complex synapses with a square root decay and simple heterogeneous synapses with a hyperbolic decay, we introduced a more general learning schedule (see Figure 7C). Here the length of the interval takes the form of γn^{β}, where n is the interval number, and γ is the length of the first interval. $\beta = 0$ corresponds to the constant schedule, and $\beta = 1$ to the linear schedule. In the parameter space spanned by the parameters γ and β, as the interval number increases, the positive gain regime (red) shrinks quickly for any positive β in the idealized exponential decay model. This regime exists in the inverse-square-root decay model for $\beta > 1$ and in the hyperbolic decay model for smaller but positive β. Differentiating between these two power-law decay models requires the examination of the sign of the signal gain around $\beta = 1$: positive for the inverse-square-root decay and negative for the hyperbolic decay. This general learning schedule thus provides experimental predictions for behavioral signatures that differ between the three idealized decay models, and allow us to discriminate between memory networks of various degrees of complexity.

DISCUSSION

We have presented a modular memory system that can solve a real-world problem such as face familiarity detection, which involves the ability to store in memory in one shot a large number of visual inputs. Thanks to the pre-processing and the interactions between fast and slow variables of the complex synaptic model, the familiarity memory capacity grows almost linearly with the number of plastic synapses or quadratically with the number of neurons of the memory module. The scaling of the system with simple synapses is only logarithmic with the number of synapses,[3,41] although the memory performance can significantly increase when the learning rate q becomes small, or when the number of states per variable increases.[3,4] However, even when the parameter q is properly tuned, the linear scaling cannot be achieved with a small number of states, and the system with complex synapses outperforms the one with simple synapses in all cases, even when the total number of dynamical variables is the same for the two systems.

The advantage of complex synapses comes from two important properties: the first one is that they involve multiple timescales, enabling the system to learn quickly using the fast components, and forget slowly due to the slow components. The second one is that the dynamical components operating on different timescales can interact to transfer information from one component to another. In the case of our specific model the information diffuses from the fast components to the slow ones, and back (see[2] for more details). These two properties are important for any memory system that involves a process of consolidation, whether the process is synaptic or requires communication across multiple brain areas (memory consolidation at the systems level, see e.g.[42]).

Our previous work[2] systematically studied the scaling properties, the memory capacity, and the robustness of a broad class of complex synaptic models for random and uncorrelated synaptic modifications. One of

the situations in which the synaptic modifications are random and uncorrelated is when the patterns of activity that represent the memories are also random and uncorrelated, which is what was assumed in all the early works on memory capacity (e.g.[37]). One of the reasons behind this assumption is that it allowed theorists to perform analytic calculations. However, it is a reasonable assumption even when more complex memories are considered. Indeed, storage of new memories is likely to exploit similarities with previously stored information. Hence, the information contained in a memory is likely to be pre-processed, so that only those components that are not correlated with previously stored memories are actually stored. In other words, it is more efficient to store only the information that is not already present in our memory. As a consequence, it is not unreasonable to consider memories that are unstructured (random) and do not have any correlations with previously stored information (uncorrelated). Unfortunately, these processes that lead to uncorrelated representations are rarely modeled explicitly (but see[18]) and we currently do not have a general theory for dealing with more realistic, highly structured memories. In our model, the face stimuli, which are highly structured and correlated, are pre-processed by a simulated visual system, whose intermediate representations are then used as inputs to our memory module. Though non-trivial higher-order statistics remain in those intermediate representations (see Figure S1), this pre-processing seems to be sufficient to achieve approximately the same scaling properties predicted for random patterns.

Another important difference between our previous and present work is related to the nature of the memory problem to be solved. In our previous work, we were dealing either with a classification task with randomly chosen labels (a typical perceptron problem with only one output unit) or with a reconstruction memory problem in which a recurrent network would learn to reproduce a previously seen input at the time of memory retrieval. In this work, we considered familiarity detection, which is a recognition memory problem. To reconstruct each individual binary feature of a memorized pattern, we would employ $N-1$ synapses. Here we have designed a system in which N such output neurons are combined and readout to report a one-bit response, which is familiarity. We are using all $N(N-1)$ plastic synapses that are available to output only one bit of information. Hence it is not surprising that in the case of reconstruction memory, the number of memories that can be retrieved (reconstructed) scales linearly with the number of neurons N, while in the case of familiarity detection, the memory lifetime scales quadratically with N.

We also studied the generalization performance of the system by considering different poses of presented faces as retrieval cues (the DP case), using probe patterns that differ from the originally stored ones. Although the task performance for this DP case is worse than in the SP case, the power-law scaling properties are similar, and the drop in performance could be compensated by introducing more memory neurons and possibly increasing the synaptic complexity. The ability to generalize to different poses is presumably helped by the complexity of the synapses. Indeed, in the case of random patterns, generalization is related to the memory SNR.[2] In future studies, we will determine whether there is a similar relationship between the SNR and the ability to generalize to different poses.

Biological interpretation

We hypothesize that the embedding module represents the ventral stream of visual cortex, where faces are clearly represented in dedicated patches, which are present in the inferior temporal cortex[43,44] and in the perirhinal cortex.[45] The memory module could be mapped onto the hippocampus, containing synapses that can be significantly more plastic than in the cortex. These highly plastic components would support one-shot online learning. This hypothesis would be compatible with the models that see the hippocampus as a memory device that compresses correlated memories before they are stored.[10,17,18] This compression process is often achieved by modeling the hippocampus as a sparse auto-encoder with one input layer, containing the representation of the memory to be compressed, an intermediate layer and an output reconstruction layer. The weights are tuned to reproduce the input in the output layer. The representations in the intermediate layer are compressed because sparseness is imposed during the learning process. Comparing the input and the output layer would be equivalent to the comparison we perform in our model between the representations in the embedding module and the representations in the memory module. In our model, we did not consider an intermediate layer as the face representations are already approximately uncorrelated. However, we could easily introduce an intermediate layer to deal with other classes of visual inputs. The reconstruction layer, and hence the detection module of our model could be in the entorhinal cortex (EC), taking advantage of the architecture of the hippocampal-cortex loop[17] (the hippocampus projects back to EC, which is also the main input to the hippocampus). Alternatively, it could be that the reconstruction layer is not explicitly implemented (see e.g.[18]). In this case the compressed representations

would emerge in one of the parts of the hippocampus without the need to reconstruct the inputs. It could be in the dentate gyrus, as hypothesized in[18], or in specific parts of the hippocampus that are involved in social interactions (e.g., CA2 is known to be involved in familiarity detection in mice[46]). The absence of an explicit reconstruction layer would require a more complex readout, that probably needs to be trained because the detection module would have to compare two different representations. This problem could be solved by adopting a different strategy to detect novelty, as suggested in.[47]

Perirhinal cortex is bidirectionally connected with EC and hence with the hippocampus. It could certainly represent familiarity even if we hypothesize that the hippocampus is the main locus of the memory module. This familiarity signal could then be broadcast to the rest of the cortex and explain why familiarity can be decoded also in other areas like infero-temporal cortex.[48]

Biological complexity at the systems level

In this article, we discussed how to take advantage of the biological complexity of individual synapses to achieve an elevated memory capacity. Complex synapses are characterized by multiple dynamical variables that operate on different timescales with interactions among them. The same computational principles could be applied to memory consolidation mechanisms implemented at the systems level: for example we could assume that the synapses are simple (e.g. binary) but heterogeneous, each characterized by a different learning rate. This is a scenario proposed in[42] where not only the synapses had different time constants, but they could also communicate through replay activity, effectively implementing a mechanism of information transfer that is similar to the one that occurs in the complex synapses. In these scenarios it is possible to obtain a power law decay of the memory SNR, however in the case studied in[42], the slowest decay would scale as $1/t$, whereas with the complex synapses of[2] we can achieve approximately $1/\sqrt{t}$. It is possible to choose a distribution of timescales of simple synapses that allow the SNR to decay as in the case of complex synapses (see[2], supplementary information 9). However, such a distribution would be strongly skewed toward slow synapses, making the initial SNR very small.

Although we currently do not have an efficient model of heterogeneous simple synapses that has the same performance as the model with complex synapses that we studied here, we cannot rule out the possibility that the brain is using both mechanisms of memory consolidation (synaptic and systems level). With the experiments that we proposed and that are discussed in the next subsection we cannot really separate the contributions of the two mechanisms.

Predictions for familiarity detection experiments

Using different learning schedules (including the constant, the linear, and a more general learning schedule), we demonstrate that the exponential decay, the inverse-square-root decay, and the hyperbolic decay models lead to distinct and testable predictions for the familiarity task performance. This can be directly tested in human (and animal) experiments. Within a series of images used in such a familiarity experiment, face images of different identities are ordered so that the same face image is repeatedly presented and at the same time evaluated (by testing familiarity detection) after an interval of a pre-determined length following a specific schedule. When designing the experiment, a large γ (the first inter-refresh interval) can lead to almost chance-level initial task performance, and a small one will cause saturated initial performance. In practice, γ should be chosen based on preliminary experiments to find an initial performance which is sensitive to manipulation (e.g., 60–90%). A recent study showed that the two-alternative forced-choice task performance for images (sketches) is around 85% when there are 100 interposed items between the first presentation and the test (with a speed of 1s for each stimulus and 0.5s for the inter-item interval).[49] We thus take $\gamma \approx 100$ as an educated guess, which could be even smaller for a shorter stimulus duration (i.e., less than 1s). For β, the relevant range in which it should vary would be between zero and one to estimate the complexity of synapses (e.g. by choosing β uniformly for different images in one experiment). Taking $\gamma = 100$ and $\beta = 1$, the interval lengths would be $\gamma n^{\beta} = 100, 200, 300, 400, 500, 600$ for $n = 1, \ldots, 6$. Roughly speaking, within a 1-h experiment containing 2400 interleaved images (about 1s for each stimulus plus 0.5s for response and inter-item interval, the same speed as in[49]), we will have test images refreshed at least six times. Then how the signal gain (and the corresponding probability of successful familiarity detection) develops as a function of the interval/presentation number will determine the temporal decay kernel of the memory signal and therefore it will allow us to infer the complexity of the memory consolidation process. We expect that biological synapses will behave similarly to the inverse-square-root model for a wide range of intervals until they gradually transition into an exponential

decay. If we could ignore the effects of systems-level consolidation and internal replay, the interval number at which this transition occurs would provide a measure of the intrinsic complexity of synapses in the hippocampus.

Limitations of the study

One limitation of our work is the assumption that the memory neurons use exactly the same representations as the input neurons. In reality, the number of memory neurons is unlikely to be precisely the same as the number of input pattern dimensions, and they would in general use a different representation of a given face from the input neurons. The detection module has to essentially compare the reconstructed memory with the representation of the current cue. This is a computation that can be performed even when the representations in the detection module are completely different from those in the input. However, it will require a smarter readout system that is trained to perform this comparison. Generalizing our system to include a more biologically plausible mapping between the embedding module and the memory module, with a corresponding readout mechanism in the detection module, is an important direction for our future work.

In our hippocampus-like memory module, there is only one feedforward layer that uses dense neural representations. However, recurrent neural computations in the hippocampus can be beneficial in some memory tasks.[50,51] In addition, sparse representations of memory patterns have long been known to harbor computational benefits such as larger memory capacity and the capability to mitigate disruptive effects of correlations.[2,3,52,53] To what extent recurrent connections and sparse coding are beneficial in our neural system for familiarity detection are questions currently under investigation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Face data set
  - Face familiarity detection system
  - Synthesizing artificial face patterns
  - Interleaving with synthesized and real face patterns
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Statistical differences between binary face patterns and random patterns
  - Evaluating the memory signal and noise
  - Choosing the optimal threshold in the FD task
  - Advanced learning schedules and idealized decay models
  - Asymptotic behavior of the optimal learning schedules for idealized synaptic models with specific decay kernels

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105856.

ACKNOWLEDGMENTS

This work was supported by NSF NeuroNex Award DBI-1707398, the Gatsby Charitable Foundation and the Swartz Foundation. M.K.B. was supported by the Kavli Institute for Brain and Mind.

AUTHOR CONTRIBUTIONS

All authors conceived the study, participated in the discussions, and wrote the paper. M.K.B. and S.F. supervised the project and acquired funding.

DECLARATION OF INTERESTS

This research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: August 20, 2022
Revised: November 30, 2022
Accepted: December 16, 2022
Published: January 20, 2023

REFERENCES

1. Fusi, S., J Drew, P., and Abbott, L.F. (2005). Cascade models of synaptically stored memories. Neuron *45*, 599–611. https://doi.org/10.1016/j.neuron.2005.02.001.

2. Benna, M.K., and Fusi, S. (2016). Computational principles of synaptic memory consolidation. Nat. Neurosci. *19*, 1697. https://doi.org/10.1038/nn.4401.

3. Amit, D.J., and Fusi, S. (1994). Learning in neural networks with material synapses. Neural Comput. *6*, 957–982. https://doi.org/10.1162/neco.1994.6.5.957.

4. Fusi, S., and Abbott, L.F. (2007). Limits on the memory storage capacity of bounded synapses. Nat. Neurosci. *10*, 485–493. https://doi.org/10.1038/nn1859.

5. Fusi, S. (2023). Memory capacity of neural network models. Human Memory (Oxford University Press) In press.

6. Kaplanis, C., Murray, S., and Clopath, C. (2018). Continual reinforcement learning with complex synapses. In International Conference on Machine Learning (PMLR), pp. 2497–2506.

7. Sanger, T.D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. Neural Netw. *2*, 459–473. https://doi.org/10.1016/0893-6080(89)90044-0.

8. Marr, D. (1971). Simple memory: a theory for archicortex. Philos. Trans. R. Soc. Lond. B Biol. Sci. *262*, 23–81. https://doi.org/10.1098/rstb.1971.0078.

9. McNaughton, B., and Morris, R. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. Trends Neurosci. *10*, 408–415. https://doi.org/10.1016/0166-2236(87)90011-7.

10. Gluck, M.A., and Myers, C.E. (1993). Hippocampal mediation of stimulus representation: a computational theory. Hippocampus *3*, 491–516. https://doi.org/10.1002/hipo.450030410.

11. O'Reilly, R.C., and McClelland, J.L. (1994). Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. Hippocampus *4*, 661–682. https://doi.org/10.1002/hipo.450040605.

12. Treves, A., and Rolls, E.T. (1994). Computational analysis of the role of the hippocampus in memory. Hippocampus *4*, 374–391. https://doi.org/10.1002/hipo.450040319.

13. McClelland, J.L., and Goddard, N.H. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. Hippocampus *6*, 654–665. https://doi.org/10.1002/(SICI)1098-1063(1996)6:6<654::AID-HIPO8>3.0.CO;2-G.

14. O'Reilly, R.C., and Frank, M.J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Comput. *18*, 283–328. https://doi.org/10.1162/089976606775093909.

15. Káli, S., and Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. Nat. Neurosci. *7*, 286–294. https://doi.org/10.1038/nn1202.

16. Battaglia, F.P., and Pennartz, C.M.A. (2011). The construction of semantic memory: grammar-based representations learned from relational episodic information. Front. Comput. Neurosci. *5*, 36. https://doi.org/10.3389/fncom.2011.00036.

17. Schapiro, A.C., Turk-Browne, N.B., Botvinick, M.M., and Norman, K.A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. Philos. Trans. R. Soc. Lond. B Biol. Sci. *372*, 20160049. https://doi.org/10.1098/rstb.2016.0049.

18. Benna, M.K., and Fusi, S. (2021). Place cells may simply be memory cells: memory compression leads to spatial tuning and history dependence. Proc. Natl. Acad. Sci. USA *118*. e2018422118. https://doi.org/10.1073/pnas.2018422118.

19. Whittington, J.C.R., Muller, T.H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T.E.J. (2020). The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. Cell *183*, 1249–1263.e23. https://doi.org/10.1016/j.cell.2020.10.024.

20. Olshausen, B.A., and Field, D.J. (2004). Sparse coding of sensory inputs. Curr. Opin. Neurobiol. *14*, 481–487. https://doi.org/10.1016/j.conb.2004.07.007.

21. Ma, Y., Tsao, D., and Shum, H.Y. (2022). On the principles of parsimony and self-consistency for the emergence of intelligence. Front. Inf. Technol. Electron. Eng. *23*, 1298–1323. https://doi.org/10.1631/FITEE.2200297.

22. Fusi, S. (2002). Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates. Biol. Cybern. *87*, 459–470. https://doi.org/10.1007/s00422-002-0356-8.

23. Standing, L. (1973). Learning 10, 000 pictures. Q. J. Exp. Psychol. *25*, 207–222. https://doi.org/10.1080/14640747308400340.

24. Brady, T.F., Konkle, T., Alvarez, G.A., and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. Proc. Natl. Acad. Sci. USA *105*, 14325–14329. https://doi.org/10.1073/pnas.0803390105.

25. Brady, T.F., Konkle, T., and Alvarez, G.A. (2011). A review of visual memory capacity: beyond individual items and toward structured representations. J. Vis. *11*, 4. https://doi.org/10.1167/11.5.4.

26. Brown, M.W., and Aggleton, J.P. (2001). Recognition memory: what are the roles of the perirhinal cortex and hippocampus? Nat. Rev. Neurosci. *2*, 51–61. https://doi.org/10.1038/35049064.

27. Eichenbaum, H., Yonelinas, A.P., and Ranganath, C. (2007). The medial temporal lobe and recognition memory. Annu. Rev. Neurosci. *30*, 123–152. https://doi.org/10.1146/annurev.neuro.30.051606.094328.

28. Squire, L.R., Wixted, J.T., and Clark, R.E. (2007). Recognition memory and the medial temporal lobe: a new perspective. Nat. Rev. Neurosci. *8*, 872–883. https://doi.org/10.1038/nrn2154.

29. Squire, L.R., and Wixted, J.T. (2011). The cognitive neuroscience of human memory since hm. Annu. Rev. Neurosci. *34*, 259–288. https://doi.org/10.1146/annurev-neuro-061010-113720.

30. Smith, C.N., Jeneson, A., Frascino, J.C., Kirwan, C.B., Hopkins, R.O., and Squire, L.R. (2014). When recognition memory is independent of hippocampal function. Proc. Natl. Acad. Sci. USA *111*, 9935–9940. https://doi.org/10.1073/pnas.1409878111.

31. Cohen, S.J., and Stackman, R.W., Jr. (2015). Assessing rodent hippocampal involvement in the novel object recognition task. a review. Behav. Brain Res. *285*, 105–117. https://doi.org/10.1016/j.bbr.2014.08.002.

32. Norman, K.A. (2010). How hippocampus and cortex contribute to recognition memory: revisiting the complementary learning systems model. Hippocampus *20*, 1217–1227. https://doi.org/10.1002/hipo.20855.

33. Norman, K.A., and O'Reilly, R.C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. Psychol. Rev. *110*, 611–646. https://doi.org/10.1037/0033-295X.110.4.611.

34. Bogacz, R., and Brown, M.W. (2003). Comparison of computational models of familiarity discrimination in the perirhinal cortex. Hippocampus *13*, 494–524. https://doi.org/10.1002/hipo.10093.

35. Savin, C., Dayan, P., and Lengyel, M. (2011). Two is better than one: distinct roles for familiarity and recollection in retrieving palimpsest memories. Adv. Neural Inf. Process. Syst. *24*.

36. Bogacz, R., Brown, M.W., and Giraud-Carrier, C. (1999). High capacity neural networks for familiarity discrimination. In 1999 Ninth International Conference on Artificial Neural Networks ICANN 99.(Conf. Publ. No. 470), *volume 2* (IET), pp. 773–778. https://doi.org/10.1049/cp:19991205.

37. Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA *79*, 2554–2558. https://doi.org/10.1073/pnas.79.8.2554.

38. Androulidakis, Z., Lulham, A., Bogacz, R., and Brown, M.W. (2008). Computational models can replicate the capacity of human recognition memory. Network *19*, 161–182. https://doi.org/10.1080/09548980802412638.

39. Jaegle, A., Mehrpour, V., and Rust, N. (2019). Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. Curr. Opin. Neurobiol. *58*, 167–174. https://doi.org/10.1016/j.conb.2019.08.004.

40. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., and Zisserman, A. (2018). Vggface2: a dataset for recognising faces across pose and age. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (IEEE), pp. 67–74. https://doi.org/10.1109/FG.2018.00020.

41. Ostojic, S., and Fusi, S. (2013). Synaptic encoding of temporal contiguity. Front. Comput. Neurosci. *7*, 32. https://doi.org/10.3389/fncom.2013.00032.

42. Roxin, A., and Fusi, S. (2013). Efficient partitioning of memory systems and its importance for memory consolidation. PLoS Comput. Biol. *9*, e1003146. https://doi.org/10.1371/journal.pcbi.1003146.

43. Tsao, D.Y., Freiwald, W.A., Tootell, R.B.H., and Livingstone, M.S. (2006). A cortical region consisting entirely of face-selective cells. Science *311*, 670–674. https://doi.org/10.1126/science.1119983.

44. Chang, L., and Tsao, D.Y. (2017). The code for facial identity in the primate brain. Cell *169*, 1013–1028.e14. https://doi.org/10.1016/j.cell.2017.05.011.

45. Liang, S., Benna, M.K., Shi, Y., Fusi, S., and Tsao, D.Y. (2021). The neural code for face memory. Preprint at bioRxiv.

46. Hitti, F.L., and Siegelbaum, S.A. (2014). The hippocampal CA2 region is essential for social memory. Nature *508*, 88–92. https://doi.org/10.1038/nature13028.

47. Tyulmankov, D., Yang, G.R., and Abbott, L.F. (2022). Meta-learning synaptic plasticity and memory addressing for continual familiarity detection. Neuron *110*, 544–557.e8. https://doi.org/10.1016/j.neuron.2021.11.009.

48. Li, L., Miller, E.K., and Desimone, R. (1993). The representation of stimulus familiarity in anterior inferior temporal cortex. J. Neurophysiol. *69*, 1918–1929. https://doi.org/10.1152/jn.1993.69.6.1918.

49. Katkov, M., Naim, M., Georgiou, A., and Tsodyks, M. (2022). Mathematical models of human memory. J. Math. Phys. *63*, 073303. https://doi.org/10.1063/5.0088823.

50. Kumaran, D., Hassabis, D., and McClelland, J.L. (2016). What learning systems do intelligent agents need? complementary learning systems theory updated. Trends Cogn. Sci. *20*, 512–534. https://doi.org/10.1016/j.tics.2016.05.004.

51. Kumaran, D., and McClelland, J.L. (2012). Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. Psychol. Rev. *119*, 573–616. https://doi.org/10.1037/a0028681.

52. Tsodyks, M.V., and Feigel'man, M.V. (1988). The enhanced storage capacity in neural networks with low activity level. Europhys. Lett. *6*, 101. https://doi.org/10.1209/0295-5075/6/2/002.

53. Wu, X.E., and Mel, B.W. (2009). Capacity-enhancing synaptic learning rules in a medial temporal lobe online learning model. Neuron *62*, 31–41. https://doi.org/10.1016/j.neuron.2009.02.021.

54. Huang, G.B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition.

55. Sun, Y., Wang, X., and Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1891–1898. https://doi.org/10.1109/CVPR.2014.244.

56. Parkhi, O.M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. BMVC 2015 - Proceedings of the British Machine Vision Conference, 1–12.

57. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., and Brossard, E. (2016). The megaface benchmark: 1 million faces for recognition at scale. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4873–4882. https://doi.org/10.1109/CVPR.2016.527.

58. Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. (2016). Ms-celeb-1m: a dataset and benchmark for large-scale face recognition. In European Conference on Computer Vision (Springer), pp. 87–102. https://doi.org/10.1007/978-3-319-46487-9_6.

59. Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141. https://doi.org/10.1109/CVPR.2018.00745.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Vggface2	Cao et al.[40]	https://academictorrents.com/details/535113b8395832f09121bc53ac85d7bc8ef6fa5b
Software and algorithms		
Python	Open source	www.python.org
SE-ResNet-50	Cao et al.[40]	https://github.com/ox-vgg/vgg_face2
Deposited code	This study	https://github.com/jil095/cmplx-syn

RESOURCE AVAILABILITY

Lead contact

Further information and requests for data should be directed to and will be fulfilled by the lead contact, Stefano Fusi (sf2237@columbia.edu).

Materials availability

The study did not generate new reagents.

Data and code availability

The publicly available VGGFace2 face data set[40] and pre-trained face networks employed in our study can be found on the following websites: http://www.robots.ox.ac.uk/vgg/data/vgg_face2/, https://github.com/ox-vgg/vgg_face2 and https://academictorrents.com/details/535113b8395832f09121bc53ac85d7bc8ef6fa5b.

All code has been deposited and is publicly available on GitHub: https://github.com/jil095/cmplx-syn.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Face data set

We used a large-scale face data set called VGGFace2.[40] Compared to other public face data sets (such as Labelled Faces in the Wild data set,[54] CelebFaces + data set,[55] VGGFace data set,[56] MegaFace data set,[57] and Ms-Celeb-1M data set[58]), it contains a relative large number of individuals (3.31 million images of 9131 individuals) and large intra-identity variations in pose, age, illumination and background (362.6 images per person on average), with available human-verified bounding boxes around faces. For each face image, the bounding box was then enlarged by 30% to include the whole head, resized such that the shorter side was 256 pixels long, and center-cropped to 224×224 pixels to serve as the input for our neural system described below.

Face familiarity detection system

We provide Table 1 summarizing the notations used in the text.

Input (embedding) module

The embedding module consists of a deep convolutional neural network (SE-ResNet-50, SENet for short), which is a ResNet architecture integrated with Squeeze-and-Excitation (SE) blocks adaptively recalibrating channel-wise feature responses.[59] Such networks for face recognition with different architectures and different training protocols are publicly available online.[40] We used one specific version of SENet, which is pre-trained on the MS-Celeb-1M data set[58] and then fine-tuned on the VGGFace2 data set. This version

was reported to have the best generalization power on face verification and identification among architectures (e.g., SENet and ResNet-50) and training protocols (e.g., training on different data sets with or without fine-tuning) tested.[40]

The 2048 dimensional activity of the penultimate layer was extracted as the face feature vector for each face image input. Because the face feature vectors are sparse and non-negative, we took the following steps to transform them into a format that's suitable as the input to the memory module: (i) the dimensionality of the feature vector of each face was first reduced using principal component analysis (PCA); (ii) each dimension was then binarized with a threshold equal to the median (−1 for values less than the median and +1 for values larger than the median). The first N binarized principal components were taken as the binary face pattern $x = [x_1, \ldots, x_N]^T$, serving as the activity of the N input neurons of the memory module.

Memory module

The memory module consists of N memory neurons, one for each input neuron of the embedding module. The j-th input neuron connects to the i-th memory neuron (for $i \neq j$) with synaptic weight (efficacy) w_{ij} and bias term b_i. There is no connection between the i-th input neuron and the i-th memory neuron for any i (i.e., $w_{ii} = 0 \forall i$). The activity of the i-th memory neuron is

$$y_i = \text{sign}\left(b_i + \sum_{j \neq i} w_{ij} x_j\right), \quad \text{(Equation 1)}$$

and we denote the binary memory patterns retrieved in this manner as $y = [y_1, \ldots, y_N]^T$. This plastic layer of synapses implements a simple feedforward memory model that can perform an approximate one-step reconstruction of a stored input pattern from a noisy cue at test time. Because the i-th memory neuron y_i is expected to reconstruct the i-th input neuron x_i, we set the value of the i-th memory neuron to be x_i during learning.

To update the synaptic weights and biases we used the learning rule:

$$\Delta w_{ij} = x_i x_j, \quad \text{(Equation 2)}$$

$$\Delta b_i = x_i. \quad \text{(Equation 3)}$$

These equations describe the desirable plasticity steps to store each new pattern. However, simply applying these additive updates would eventually result in unbounded values of the w_{ij}. Therefore, we employed a mechanism to limit the weights to bounded dynamical ranges. For each synapse (i.e., for each weight w and bias term b), we implemented a complex synaptic model[2] with m dynamical variables $u_1, \ldots, u_m$ in discrete time. Here m denotes the total number of variables per synapse (a measure of synaptic complexity), each of which operates on a different timescale. Specifically, at each time step t the dynamical variables u_k (for $2 \leq k \leq m$) are updated as follows (the indices i and j labeling the synapses are omitted for simplicity)

$$u_k(t+1) = u_k(t) + n_0^{-2k+2}\alpha(u_{k-1}(t) - u_k(t)) - n_0^{-2k+1}\alpha(u_k(t) - u_{k+1}(t)). \quad \text{(Equation 4)}$$

For $k = m$, the last variable u_{k+1} is simply set to zero in this update equation, and for $k = 1$ we have

$$u_1(t+1) = u_1(t) + l(t) - n_0^{-1}\alpha(u_1(t) - u_2(t)). \quad \text{(Equation 5)}$$

Here $l(t)$ is the desirable update (Δw or Δb) imposed by the pattern $x(t)$, which takes a value $+1$ or -1 and is computed from Equations 2 or 3. The first variable u_1 is used as the actual value of the synaptic weight w or bias b at test time. The parameters α and n_0 determine the overall timescale of the model dynamics and the ratio of timescales of successive synaptic variables (we set $\alpha = 0.25$ and $n_0 = 2$ in our models; see[2] for additional details).

To study the situation in which variables can only be stored with limited precision, we discretized the m synaptic variables and truncated their dynamical range to a maximum and minimum value. Hence, each variable can take one of only a finite number of integer-spaced values arranged symmetrically around zero, namely $\{-V, -V+1, \ldots, V-1, V\}$, where in our simulations we chose $V = 31/2$, corresponding to 32 levels (5 bits). At every time step, if the $u_k(t+1)$ computed according to Equations 4 and 5 falls between two adjacent levels, its new value is set to one of those two levels, based on the result of a biased coin flip with an odds ratio equal to the inverse ratio of the distances from $u_k(t+1)$ to the two levels.

In the comparison between models with simple synapses and complex synapses, we further considered different learning rates q for the plastic synapses. Each synapse in the model is updated independently. A synapse (w_{ij}) is updated with probability q every time an input pattern is stored (Δw_{ij}). With probability $1 - q$, the weight update Δw_{ij} is rejected and w_{ij} remains unchanged. Synapses with smaller learning rate are considered more "rigid".[3,41]

Readout (detection) module

The readout (detection) module compares the output $x = [x_1, \ldots, x_N]^T$ of the embedding module and the output $y = [y_1, \ldots, y_N]^T$ of the memory module to assess the level of familiarity of a given pattern. This module computes the Hamming distance between x and y, and outputs "familiar" (or "unseen"/"unfamiliar"/"novel") if the distance is smaller (or larger) than some pre-set threshold. This approach is similar to the one proposed in.[36]

Synthesizing artificial face patterns

The VGGFace2 data set only contains faces from 9131 different people. To facilitate the evaluation of the memory performance of our system over multiple time scales, a larger number of independent non-evaluated patterns are required to be stored in between the face patterns whose memory signals are being tracked. Thus, we synthesized artificial face patterns matching the first and second moments of real face patterns. First, we extracted the mean values and the diagonal covariance matrix of the face feature vectors after PCA to get an estimate of the distribution of patterns generated by the faces of all the people in the data set. We then synthesized artificial face patterns by passing new samples from the corresponding multivariate normal distribution through the binarization step. Mathematically, this process is equivalent to generating unstructured, random, binary patterns. These artificial patterns were presented to the neural system to be memorized at time steps in between the storage of the real face patterns, but were not used to evaluate the memory performance.

Interleaving with synthesized and real face patterns

For all simulations in the main text, we stored synthesized face patterns (equivalent to unstructured, random, binary patterns) at time steps in between the storage of the real face patterns. Therefore, most of the patterns before and after one arbitrary stored real face pattern are synthesized face patterns. Here we compared the case of interleaving with synthesized patterns and the case of interleaving with real face patterns (different poses of face images from other non-evaluated people) for a system with N = 128 and m = 6. We found that the two cases are almost indistinguishable (see Figure S2). This indicates that our main simulations are unaffected no matter whether we are using synthesized patterns or real face patterns for the interleaving purpose.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical differences between binary face patterns and random patterns

After dimensionality reduction and binarization in the preprocessing for face image inputs, the face patterns become binary, similar to random unstructured binary patterns sampled from independent and identically distributed Bernoulli variables. Here we show that there are higher-order non-trivial statistics in the face patterns, by comparing the pattern correlation matrices and the feature correlation matrices between the face and the random data sets.

To generate the data sets we are comparing here, the outputs of the convolutional neural network's penultimate layer are extracted, consisting of 431050 samples (8621 people, each with 50 poses) with 2048 features. Features are centered first and then fed into principal component (PC) analysis. The covariance matrix of these 2048 PC features has decreasing diagonal variance and zero off-diagonal covariance.

Taking one pose from each person, we then binarized these PC features and obtain the face data set ($M = 8621$ binary patterns, each with 2048 features). In the random data set, the 8621 random patterns (with 2048 binary features) are directly sampled from a multivariate Gaussian distribution (with zero mean and diagonal covariance matrix derived from the above face PC features) and then binarized. We ran the above process with ten random seeds (sampling new poses or new random patterns).

For the top N features ($N \leq 2048$) in the face and the random data sets, we then computed the pattern correlation matrix ($M \times M$; each element representing the correlation between two patterns of N dimensions) and the feature correlation matrix ($N \times N$; each element representing the correlation between two features of M dimensions). The variance, skewness, and kurtosis of these off-diagonal elements in the pattern correlation and the feature correlation matrices are shown in Figure S1. For the off-diagonal elements in the pattern correlation matrix, the face data set and the random data set have indistinguishable variance, but different skewness and kurtosis. For the off-diagonal elements in the feature correlation matrix, the face data set also has different profiles in variance, skewness, and kurtosis from the random data set. These differences indicate that the face patterns are substantially statistically different from the random patterns also studied in the main text.

Evaluating the memory signal and noise

The ideal observer signal $\mathcal{S}_{io}$ and noise $\mathcal{N}_{io}$ are computed as follows.

For a given face memory, the signal at time t of the input pattern $x(t')$ stored at an earlier time t' is defined as the overlap (inner product) between the synaptic modification $\Delta w_{ij}(t')$ imposed at storage and the current ensemble of synaptic weights $w_{ij}(t)$:

$$S_{io}(t - t') = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \Delta w_{ij}(t') w_{ij}(t). \quad \text{(Equation 6)}$$

We can then compute the average (denoted by $\langle\rangle$) over all memories with an age of $\Delta t = t - t'$ to obtain the expected signal

$$\mathcal{S}_{io}(\Delta t) = \langle S_{io}(\Delta t) \rangle, \quad \text{(Equation 7)}$$

and the corresponding noise term

$$\mathcal{N}^2_{io}(\Delta t) = \langle (S_{io}(\Delta t) - \langle S_{io}(\Delta t) \rangle)^2 \rangle. \quad \text{(Equation 8)}$$

Similarly to the ioSNR, the readout signal $\mathcal{S}_r$ is defined as the overlap between an input pattern $x(t')$ (stored at time t') and the retrieved memory pattern $y(t, t')$, the output of the memory module when the same pattern is presented again at time t without updating the synaptic weights. We have

$$S_r(t - t') = \frac{1}{N} \sum_{i=1}^{N} x_i(t') y_i(t, t'). \quad \text{(Equation 9)}$$

As above, we can compute the expected signal $\mathcal{S}_r$ and noise $\mathcal{N}_r$ by averaging over memories of a given age, and obtain the readout signal-to-noise ratio (rSNR) $\mathcal{S}_r/\mathcal{N}_r(\Delta t)$.

Choosing the optimal threshold in the FD task

As new patterns are presented to our proposed memory system, the familiarity of any old patterns decreases with time, and therefore the distribution of the readout signal for familiar patterns approaches the one of the unseen patterns (see Figure S3A). In Figure S3B, we plot the classification accuracy of the detection module as a function of elapsed time (relative to the presentation of the face pattern) with different signal thresholds. Smaller thresholds result in higher error rates for unseen faces, but have a better performance for familiar faces. To provide an operational definition of familiarity for the detection module, we study the overall performance (the average of classification accuracy over the whole age-range) as a function of the signal threshold and age-range (see Figure S3C). We include an equal number of familiar and unfamiliar faces in these test sets, and weight false positive and false negative errors equally. For longer age-ranges, the optimal threshold gradually decreases, since the familiar faces become indistinguishable from unseen faces, although choosing the right age-range ultimately depends on the application and on the longest timescale of the synaptic model employed in the memory system.

For the evaluations in the main text, the threshold of each model is first optimized on a balanced test set with familiar faces within the model-specific age-range (we choose t^*_{ioSNR}), and then evaluated over longer time scales. Since the distribution of the readout signal for unseen faces does not change over time, the detection module with the fixed threshold detects unseen faces with constant error rate (1 - true negative rate) (see Figure S4B), while it recognizes familiar faces (true positive rate) better for more recent than for older ones (see Figure S4A). The FD classification performance is the average of the true positive and true negative rates.

Advanced learning schedules and idealized decay models

In addition to the simple schedule described above in which each face is stored only once, we further consider two types of advanced learning schedules (namely optimal and pre-determined learning schedules), where the same image pattern can be presented more than once and is evaluated during each presentation.

Under the optimal learning schedules, the ideal observer signal (ioSignal) of a specific pattern is constantly monitored after its initial presentation. Every time its ioSignal drops below the pre-specified threshold, the same pattern is presented again (refreshed) to boost its memory strength. The length of the n-th interval between the n-th and $(n + 1)$-th presentations will vary with the interval number n. Even though monitoring the memory signal in real time (without modifying it) is not feasible in experiments, we will show that this theoretical analysis of the length of the n-th interval reveals major differences between synaptic models with different complexity, which have measurable consequences.

Motivated by theoretical results on optimal learning schedules, we also propose pre-determined learning schedules, under which the length of the interval between each two consecutive presentations takes the form of γn^{β}, where n is the interval number, γ is the length of the first interval (between the first and the second presentations), and β is the exponent. When β takes value 0 or 1, this schedule degenerates into constant or linear schedules, in which a specific pattern is presented again (refreshed) after an interval of a constant length or after an interval of a linearly increasing length.

To facilitate the study of the behavior of our simulated models with simple and complex synapses under these schedules, we introduced three idealized decay models in which the memory signal decays with a pre-specified profile: exponential, inverse-square-root power-law, and hyperbolic power-law. This allows us to quickly run many numerical experiments with different learning schedules, since we do not have to simulate the internal dynamics of the complex synapses, which is inherently stochastic and would require averaging over many realizations to obtain an estimate of the expected behavior (which is instead represented directly by the specified decay function). These idealized models will allow us to demonstrate that pre-determined learning schedules can be used in experiments to discriminate different decay profiles, which are related to the complexity of a memory system, without the need to access its internal constituents.

Free parameters in the idealized decay models were chosen to match the behavior of the simulated models with simple or complex synapses:

(1) The exponential decay model, with memory signal $r(t) = C_{\mathrm{exp}} e^{-t/\tau_{\mathrm{exp}}}$, where $C_{\mathrm{exp}} = 1$ is the initial memory strength and $\tau_{\mathrm{exp}} = 7.486$ is the time constant, fit to the signal strength of simulated models with simple synapses ($m = 1$).

(2) The inverse-square-root power-law decay model, with $r(t) = C_{\mathrm{isr}}/\sqrt{t + 1}$, where $C_{\mathrm{isr}} = 1.316$ is the initial memory strength in order to fit the power-law decay regime of signal strength of models with complex synapses ($m = 8$).

(3) The hyperbolic power-law decay model, with $r(t) = C_{\mathrm{hyp}}/(t + 1)$, where C_{hyp} is equal to C_{isr}.

Asymptotic behavior of the optimal learning schedules for idealized synaptic models with specific decay kernels

Here we derive the constant length of the interval between successive presentations of the same pattern for a simple synaptic model with an exponential decay and the asymptotically linear increase of the length of the interval for the inverse-square-root decay model.

Let $r(t; t_n) = r(t - t_n)$ denote the decay kernel of the memory signal of a synaptic model at time t for a pattern presented at time t_n (the n-th presentation, $t \geq t_n$). If θ is the threshold on this memory signal (such that when the signal has dropped to this level the same pattern will be presented again), we have the following system of equations:

$$\begin{cases} r(t_2; t_1) & = \theta \\ r(t_3; t_2) + r(t_3; t_1) & = \theta \\ & \cdots \\ r(t_n; t_{n-1}) + r(t_n; t_{n-2}) + \cdots + r(t_n; t_1) & = \theta. \end{cases} \qquad \text{(Equation 10)}$$

Let the n-th interval be $\tau_n := t_{n+1} - t_n$. Then we have:

$$\begin{cases} r(\tau_1) & = \theta \\ r(\tau_2) + r(\tau_2 + \tau_1) & = \theta \\ & \cdots \\ r(\tau_n) + r(\tau_n + \tau_{n-1}) + \cdots + r(\tau_n + \tau_{n-1} + \cdots + \tau_1) & = \theta. \end{cases} \quad \text{(Equation 11)}$$

For the exponential decay with $r(t) = C\exp(-t/\tau_0)$, we can solve these equations from τ_1 to τ_n sequentially and obtain the intervals for the exponential decay:

$$\begin{cases} \tau_n & = \tau_0 \ln(C/\theta), n = 1 \\ \tau_n & = \tau_0 \ln(1 + C/\theta), n > 1. \end{cases} \quad \text{(Equation 12)}$$

For the inverse-square-root decay with $r(t) = C/\sqrt{t+1}$, we have:

$$1\Big/\sqrt{\tau_n + 1} + 1\Big/\sqrt{\tau_n + \tau_{n-1} + 1} + \cdots + 1\Big/\sqrt{\tau_n + \tau_{n-1} + \cdots + \tau_1 + 1} = \theta\Big/C, \quad \text{(Equation 13)}$$

which can be well approximated by the following integral equation when n is large enough:

$$\theta / C = \int_0^n \left[\int_{n-t}^n \tau(x)dx\right]^{-1/2} dt. \quad \text{(Equation 14)}$$

Taking the derivative of both sides with respect to n, we have:

$$\begin{aligned} 0 &= \left[\int_{n-t}^n \tau(x)dx\right]^{-1/2}_{t=n} + \int_0^n \frac{d}{dn}\left[\int_{n-t}^n \tau(x)dx\right]^{-1/2} dt \\ &= \left[\int_0^n \tau(x)dx\right]^{-1/2} - \frac{1}{2}\int_0^n \frac{\tau(n) - \tau(n-t)}{\left[\int_{n-t}^n \tau(x)dx\right]^{3/2}} dt. \end{aligned} \quad \text{(Equation 15)}$$

It can be verified that $\tau(n) = Mn$ is the solution of the above equation (where M is a constant), and thus we proved the asymptotic linearity for the inverse-square-root decay.

To calculate the asymptotic linear coefficient M, we insert $\tau_n = Mn$ into Equation 13 (and ignore $+1$ under each square root when n is large enough),

$$1\Big/\sqrt{n} + 1\Big/\sqrt{n + n - 1} + \cdots + 1\Big/\sqrt{n + n - 1 + \cdots + 1} = \sqrt{M}\theta\Big/C. \quad \text{(Equation 16)}$$

We call the left side A_n and let n go to infinity. In this limit we have

$$M = (A_\infty C/\theta)^2. \quad \text{(Equation 17)}$$

Finally, we calculate A_∞ as follows:

$$\begin{aligned} A_\infty &= \lim_{n\to\infty} \sqrt{2}\sum_{x=0}^{n-1} \frac{1}{\sqrt{(2n-x)(x+1)}} \\ &\simeq \lim_{n\to\infty} \sqrt{2}\int_0^{n-1} \frac{1}{\sqrt{2n + (2n-1)x - x^2}} dx \\ &= \lim_{n\to\infty} -\sqrt{2}\arcsin\frac{-2x + 2n - 1}{\sqrt{8n + (2n-1)^2}}\Bigg|_{x=0}^{n-1} \\ &= \frac{\pi}{\sqrt{2}}. \end{aligned} \quad \text{(Equation 18)}$$