nature neuroscience

Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions

Szabolcs Káli[1] & Peter Dayan[2]

During sleep, neural activity in the hippocampus and neocortex seems to recapitulate aspects of its earlier, awake form. This replay may be a substrate for the consolidation of long-term declarative memories, whereby they become independent of the hippocampus and are stored in neocortex. In contrast to storage, other crucial facets of competent long-term memory, such as maintenance of access to stored traces and preservation of their correct interpretation, have received little attention. We investigate long-term episodic and semantic memory in a theoretical model of neocortical-hippocampal interaction. We find that, in the absence of regular hippocampal reactivation, even supposedly consolidated episodic memories are fragile in the face of cortical semantic plasticity. Replay allows access to episodes stored in the hippocampus to be maintained, by keeping them in appropriate register with changing neocortical representations. Hippocampal storage and replay also has a constructive role in the recall of structured, semantic information.

Ongoing learning renders cortical representations ceaselessly plastic. This poses two problems for memory. First, the patterns of activity that should inspire the recall of a particular memory are nonstationary. Second, the information recalled from memory has to be interpreted (decoded) in an ever-mutating code. Although these stability-plasticity[1] issues of input and output accessibility have been suggested to underlie the phenomenon of infantile amnesia[2], they are neglected by most current theories of long-term memory. Here, we study the replay of hippocampal-neocortical patterns during sleep and quiet wakefulness, and suggest that it is crucial to ensuring that old memories can be appropriately retrieved and understood in the current representational coordinates.

We investigated storage, access and decoding of episodic memory (autobiographical event memory) and semantic memory (here, the storage and retrieval of structured information not specific to one particular episode). These two subtypes of declarative memory are known to depend on the hippocampus and adjacent cortical areas of the medial temporal lobes (MTL)[3]. Damage to these regions results in amnesia, whereby the acquisition of new declarative memories is impaired (anterograde amnesia) and some memories acquired before the damage are lost (retrograde amnesia). However, the detailed characteristics of these deficits are controversial.

Two issues dominate the debate. First, it is unclear whether episodic and semantic memory are affected similarly in amnesia and, generally, whether different subtypes of declarative memory are processed in the same way by the hippocampus and neocortex[4,5]. Second, it is controversial whether the hippocampus has a temporary or a permanent role in episodic memory. According to the temporary view, which we call the transfer model and which has been a popular target of computational studies[6,7], memories are reorganized (or consolidated) over a potentially long period (from several days in rats to decades in humans), so that memories ultimately lose their initial dependence on MTL areas[8–10]. According to the permanent view, the MTL is always required for recalling episodes[11,12], although not necessarily for recalling semantic information.

Though there is no direct experimental evidence, transfer has been proposed to depend on hippocampally initiated reinstantiation during slow-wave sleep (and perhaps also rapid eye movement sleep and/or quiet wakefulness) of distributed cortical activity patterns characterizing previous active behavioral states[13–17]. Indeed, sleep between training and testing can significantly enhance performance in various learning tasks[18,19], with slow-wave sleep seeming to have a particular effect on declarative memory[20].

Here, we built a combined hippocampal-neocortical model of episodic and semantic information, and studied the relationship between hippocampally initiated replay and storage, access and decoding of declarative memory in the face of representational change. First, we showed that replay does not establish in neocortex durable episodic memories that are independent of the hippocampus. Next, we demonstrated a possible role for replay as helping maintain access to episodes in the presence of the hippocampus. Finally, we examined the acquisition, consolidation and maintenance of statistically rich general semantic information and compared it with episodic memory.

RESULTS

Transfer

Our model (**Fig. 1**) follows a widely accepted abstraction scheme. The neocortex represents sensory inputs as patterns of activity distributed over large neuronal populations. Memories recalled, either

[1]Institute of Experimental Medicine, Hungarian Academy of Sciences, P.O. Box 67, Budapest 1450, Hungary. [2]Gatsby Computational Neuroscience Unit, University College London, 17 Queen Square, London WC1N 3AR, UK. Correspondence should be addressed to S.K. (kali@koki.hu).

Published online 22 February 2004; doi:10.1038/nn1202

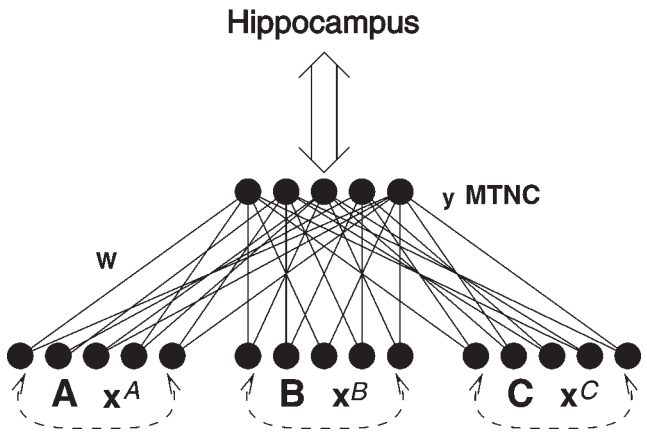

Figure 1 Model architecture. All units in neocortical areas A, B and C are connected to all units in area MTNC through bidirectional, symmetric weights (W). Connections between units in the input layer are restricted to the same cortical area and are treated as weak local attractors (dashed arrows). $\mathbf{x}^A$,$\mathbf{x}^B$,$\mathbf{x}^C$ and **y** denote activity vectors in the corresponding neocortical areas. The hippocampus is not directly implemented, but it can influence and store the patterns in MTNC. All communication between the hippocampus and the input areas is through area MTNC.

directly or via the hippocampus, should be represented by the same patterns of activation over the same populations. Distributed neocortical representations are acquired and continually adjusted through an unsupervised learning process that alters synaptic connections within and between neocortical areas to reflect the (ever-changing) statistical structure of the input[21]. Plasticity affecting the representation of one stimulus inevitably also affects the representation of many other stimuli.

Our neocortical model has two layers. The lower layer consists of higher-order association areas from different (sub-)modalities (labeled A, B and C), representing highly processed sensory inputs. A given pattern of activities of neurons in one of the lower-level areas encodes a particular stimulus; in principle, we may encounter any combination of the possible patterns in each input area. The upper layer models areas in medial temporal neocortex (MTNC): entorhinal, perirhinal and parahippocampal cortices. MTNC is separated because of the cross-modal integrative nature of these areas and because they form the exclusive cortical conduit to the hippocampal formation. The two layers are connected in a reciprocal and hierarchical manner[22,23]. MTNC activity can inspire hippocampal recall, which in turn affects MTNC activity and thus activity in areas A–C. Each neocortical area contains a large number of abstract neuron-like units, which form independent cleanup networks[24] in the input areas. These are not explicitly simulated, but (in the absence of feed-forward activation of the area) have the effect of converting top-down input that closely resembles one of the valid input patterns in that area into an exact version of that pattern.

The neocortex acts as a probabilistic generative model[25]. Unsupervised learning[26,27] extracts categories, tendencies and correlations from the statistics of the input into the weights W between area MTNC and the input areas. We refer to this information as semantic knowledge or semantic memory, even though it probably best corresponds to general semantics, which is not always considered part of declarative memory despite sharing important characteristics with, for instance, fact memory. After training, the network is capable of recognizing, completing and cleaning up novel, partial and noisy inputs by statistical sampling (see Methods for details). If some input units (say, those in areas A and B) are clamped, the samples produced in area C represent the network's inferences about probable completions of the pattern in A and B.

We first examined the long-term storage and recall of episodic memories within this framework. In our scheme, episodes are specific items—that is, completely specified patterns of activity over the units in areas A, B and C. The quality of episodic recall is judged according to whether the pattern in one area (say C) for an episode can be recalled through successive sampling iterations (possibly involving the hippocampus) starting from the correct activities for that episode in areas A and B but completely corrupt (random) activities in area C. Recall stops after a fixed number of iterations (usually 20), or if the pattern in area C comes within the local basin of attraction (taken to be a Hamming distance of 5) of a valid input pattern in that area.

Under appropriate conditions, the hippocampus is assumed to (i) store a representation of a hitherto unfamiliar current MTNC pattern so that (ii) it can be autoassociatively reinstated when MTNC activity inspired by a new cue is sufficiently similar[28–34]. We used an arbitrary threshold of 20 on the bitwise Hamming distance to judge similarity. Once the mapping between the input areas and MTNC has been established, reinstating the original set of MTNC activities suffices to reinstate the whole cortical pattern, provided that the neocortical and hippocampal representations are appropriately in register. This allows sampling in the neocortex to converge instantaneously. Further, (iii) the patterns stored in the hippocampus are assumed to be activated intrinsically and randomly (for instance during sleep[13,14,16]) and this, via the top-down generative model, permits replay.

For comparison with experimental data[9,35–37] and earlier modeling[6,7], we first confirmed that our model could capture basic phenomena of hippocampal-dependent consolidation as in the transfer model[3,10,38]. We modeled consolidation by alternating learning between 'experience' and 'replay'. The first corresponded to continued exposure to regular input stimuli. The second started from hippocampal reactivation in MTNC of a random stored memory pat-

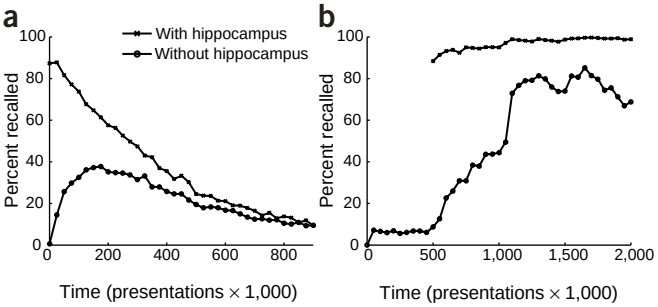

Figure 2 Replay transfers episodic memories to neocortex. The graphs show average recall performance (as percentage of successful recall) on episodic patterns as a function of time. The upper curves (with crosses) represent normal controls, and the lower curves (with circles) are for the case when the hippocampal module is inactivated for testing. (**a**) Episodic patterns were stored sequentially, and their strength decayed exponentially. The curves are averages over all episodic patterns used, aligned at the initial presentations (also the start of consolidation) of the particular patterns. (**b**) The set of input patterns changed in time according to the four phases of training described in the text. All episodic patterns were stored in the hippocampus (and thereby became eligible for replay) simultaneously at the end of phase 1 (which is why the upper curve starts there). Hippocampal patterns did not decay in these simulations.

Table 1 Simulation summary

a

Figure(s)	Input domains	Stored patterns	Hippocampal recall	Replay (1)	Replay (2)
2b	1,2 1,2 1,2 1,2	– 10 10 10	– + + +*	– + + +	– – – –
4a (7a)	1,2 1,2 1–3 1–4	– 50 50 50	– + + +*	– – – –	– – – –
4b (7b)	1,2 1,2 1–3 1–4	– 50 50 50	– + + +*	– – – –	– + + +
4c (7c)	1,2 1,2 1–3 1–4	– 50 50 50	– + + +*	– + + +	– – – –
4d (7d)	1,2 1,2 1–3 1–4	– 50 50 50	– + + +*	– + + +	– + + +

b

Curve (Fig. 6)	Input domains	Stored patterns	Hippocampal recall	Replay (1)	Replay (2)
Open circles	1,2 1,2 1,2	– 50 50	– – –	– – –	– – –
Crosses	1,2 1,2 1,2	– 50 50	– + +	– – –	– – –
Filled circles	1,2 1,2 1,2	– 50 50	– – –	– + –	– – –
Triangles	1,2 1,2 1,2	– 50 50	– + +	– + –	– – –

Shown is a schematic representation of some basic characteristics of the various simulations. Each column contains a separate entry for each phase of every simulation listed. (**a**) The first column refers to the figure(s) showing the basic results from a particular, four-phase simulation; the second column shows the domains from which the input patterns originate; and the third shows the number of episodic patterns stored in the hippocampus. The remaining columns indicate the presence or absence of hippocampal pattern completion, neocortical learning during replay, and updates of hippocampal-MTNC connections, respectively. *__Figures 2b__ and __7a–d__ also show the results in the absence of hippocampal recall. (**b**) This table refers to the four different conditions in __Figure 6__, which shows results from a three-phase paradigm.

tern. The resulting MTNC activation led to the reactivation of the input areas (also taking into account the effect of local attractors in each input area). The combined activity induced weight modifications in the neocortex according to the same learning algorithm used for external input-driven learning[27]. Blocks of 900 hippocampally initiated learning events alternated with blocks of 100 input-initiated (general training) events. Such a substantial degree of imbalance was required for robust episodic consolidation.

In these simulations, the inputs consisted of 20 possible random binary patterns in each input area (A, B and C), with all $20^3 = 8{,}000$ combinations initially equally likely. In a first phase of semantic training, 200,000 patterns selected at random were presented to the neocortical module, establishing the relationship between the activities in A–C and MTNC. We then simulated a common paradigm for animal studies of retrograde amnesia[9,35–37], in which several different sets of stimuli are presented at different times before the hippocampus is lesioned. In our case, 18 specific input patterns (involving all three areas) were designated as episodic patterns to be memorized. These were introduced sequentially (each separated by 50,000 pattern presentations). Recall of all stimuli was tested in the absence (and also in the presence) of the hippocampus. Normal forgetting arose through the exponential decay of hippocampal memory strength (with a time constant of 200,000 pattern presentations), affecting the probabilities of a pattern being selected for replay and of successful hippocampal pattern completion during recall.

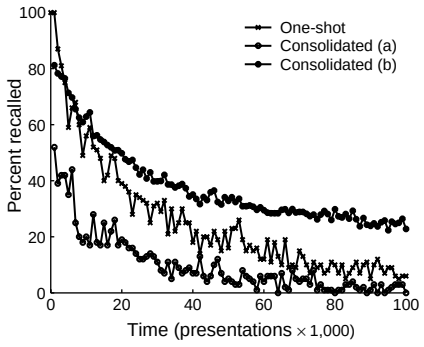

Figure 3 Episodic memories stored in neocortex are extinguished by subsequent semantic training. The curve marked with crosses ('one-shot') is for an isolated neocortical network trained to asymptotic performance on a particular episodic pattern. The other two curves show recall performance in the neocortical network as a function of time after the removal of the hippocampus. The curve marked with open circles ('consolidated (a)') is for a single pattern from among those used to construct **Figure 2a**, which has been hippocampally 'consolidated' for 250,000 presentations. The curve marked with filled circles ('consolidated (b)') is an average over the same patterns that were used in **Figure 2b**, and starts from the state of the network after 1,250,000 presentations illustrated in that figure (a time near the highest overall degree of consolidation).

The averaged time-performance curves for the full and hippocampally inactivated models (hereafter referred to as 'normals' and 'hippocampals'; **Fig. 2a**) replicated many important characteristics of the experimental data. Normals performed best directly after training and forgot gradually over time. Hippocampals performed at floor for patterns learned just before hippocampus removal, but were more proficient when more time intervened between training and lesion. The difference between hippocampals and normals became negligible for the most remote time periods. This has been taken as a signal of successful consolidation in several animal experiments and all previous models of memory consolidation[6,7,9,35–37].

The remaining simulations qualified and extended these results. They were carried out in a single standard framework differing in some respects from the simulation described above. First, we treated semantic as well as episodic memory, which required a richer collection of input patterns. We considered four different domains. In each, ten random binary patterns were generated for each input area. In all but one domain, all 1,000 possible combinations of the valid patterns in areas A, B and C were considered equally likely. The remaining, semantically 'structured' domain had inter-areal semantic structure, in that each of the ten possible patterns in area A always appeared with a (different) associated pattern in area C. Any of these A-C pairs could appear with any of the patterns in area B. All 100 possible combinations of patterns in this domain were equally likely. The unstructured domains allowed the study of episodic memory, uncontaminated by semantic knowledge, whereas

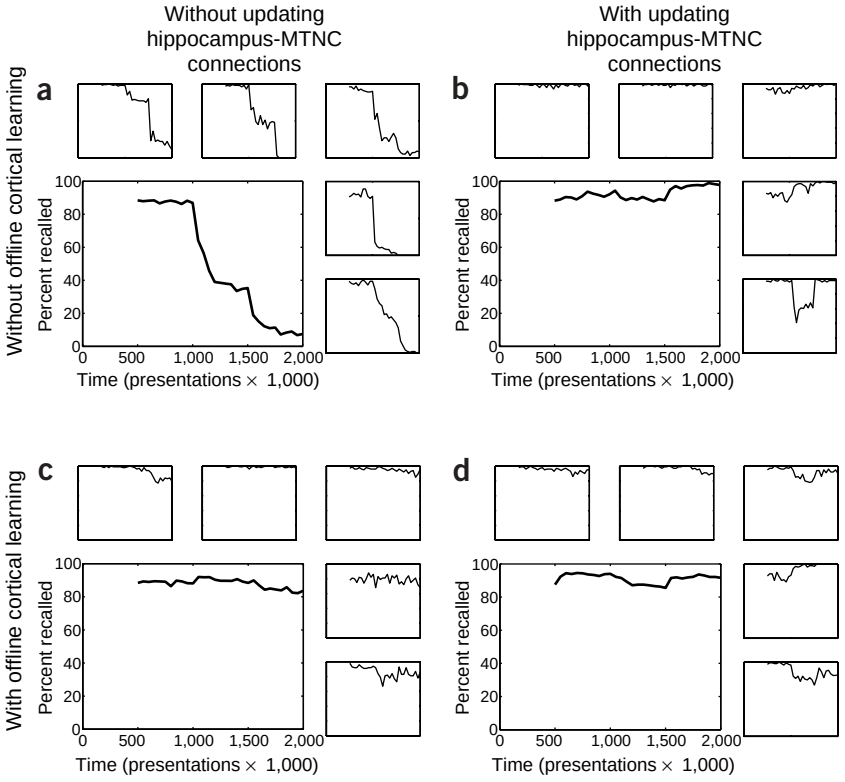

Figure 4 Replay protects episodic memories against representational change. (**a**–**d**) Effects of changes in neocortical representations on the recall of previously stored episodes (**a**) in the absence of hippocampally initiated replay; (**b**) if the correspondence between hippocampal and MTNC representations of the episode is updated during off-line replay; (**c**) if neocortical connections are updated during replay episodes but hippocampal-MTNC connections are not; and (**d**) if both neocortical and hippocampal-MTNC connections are updated. Within these panels, the larger graphs are averages over all stored episodes and the smaller graphs are examples of individual episodes.

the structured domain was used in later sections to study semantic memory. Finally, because we were studying the effects of neocortical rather than hippocampal plasticity, we ignored the decay of hippocampal memory traces.

To study the effects of changing the input statistics, we separated learning and recall into four separate phases, each of 500,000 presentations (**Table 1**). In general, patterns from domains 1 and 2 were used for general semantic training in phase 1. The same patterns were presented in phase 2, typically after episodic experience. Patterns from domain 3 and 4 were added in phases 3 and 4, respectively. Different experiments involved different aspects of storage, recall and replay in the hippocampus. To provide a fair test of semantic representation, patterns from the structured domain (domain 2) always comprised 1% of the patterns presented during experience epochs; the remaining patterns were uniformly drawn from all the other domains present during a phase.

After general semantic training with the patterns of domains 1 and 2, ten episodic patterns from domain 1 were simultaneously stored in the hippocampus (with nondecaying traces). Recall of these was monitored with constant hippocampal replay (as above) and either with or without hippocampally aided recall just at the time of test (**Fig. 2b**). This allowed us to determine how well replay instructs the neocortex about the episodic patterns and thereby obviates the need for itself in recall. The upper curve shows that, as might be expected, with hippocampal help, recall was exemplary. The lower curve, for which the hippocampus was inactivated during test, shows that hippocampal replay could indeed embed the memories in the neocortex. In the model, this result again required a rather extreme (90:10) imbalance between replay and experience epochs (data not shown). In fact, the marked improvement in performance at the start of phase 3 occurred because the addition of domain 3 to awake experiences decreased the frequency of domain 1 experiences, and thereby indirectly increased the relative frequency and impact (within domain 1) of the episodes that participated in replay (because the proportions of experience and replay epochs remained fixed).

Such consolidation does not, by itself, lead to long-term stability of episodic memory traces in neocortex (**Fig. 3**). Recall probabilities dropped rapidly in the absence of the hippocampus for episodes that were stored in neocortex, either directly (through repeated presentations of the patterns to the input layer) or through the consolidation process described above (for 'consolidated' patterns taken either from **Fig. 2a** or from **Fig. 2b**). Comparing the curves shows that recall performance on the 'consolidated' patterns of **Figure 2** decayed when the hippocampus was switched off in the model and the network was subjected to general training on all valid input patterns, and did so at speeds compara-

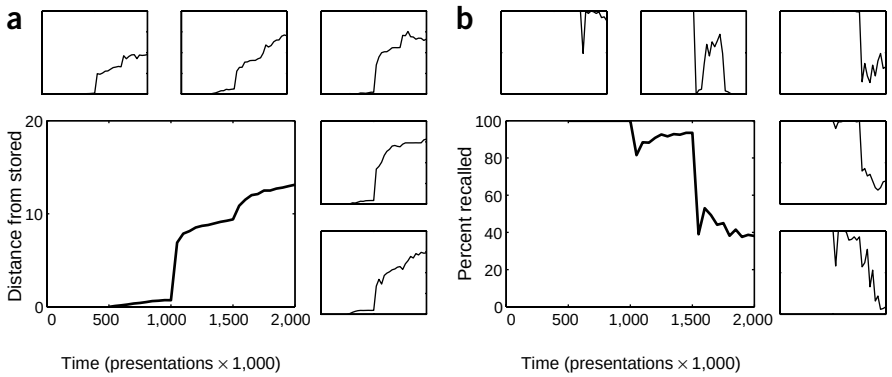

Figure 5 Analysis of the reasons why episodic recall breaks down in **Figure 4a**. (**a**) Distance between the MTNC pattern currently associated with the representation of the episode in the input areas and the MTNC pattern associated with the stored hippocampal memory trace. (**b**) Percentage of correct recall in the input areas if we start the recall process from the stored MTNC representation of the episode.

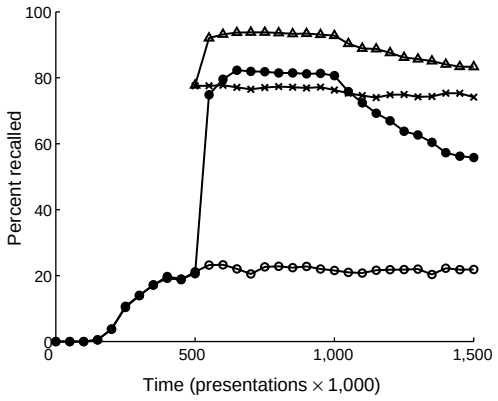

Figure 6 Hippocampal replay and recall aid the acquisition and consolidation of semantic information. The curves show the percentage of correct pattern completion in area C for patterns from domain 2. Input patterns are drawn from domains 1 and 2 throughout. Forty episodic patterns from domain 2 (and ten from domain 1) are stored after 500,000 presentations, and there is hippocampally initiated replay of episodic patterns between time points 500,000 and 1,000,000 in some conditions. The different curves represent different training conditions, and are marked by different symbols: open circles represent pure neocortical learning; crosses stand for neocortical learning supplemented by hippocampal learning and recall of episodes (but no replay); filled circles mark the condition with replay, but without hippocampal involvement in recall; and the line with triangles is for simulations with both hippocampal replay and recall.

ble to that for a nonconsolidated pattern. Indeed, forgetting was much faster than for normals (**Fig. 2a**; note the different time scales), indicating that the decay rate of hippocampal traces determined the normal forgetting rate in our complete model.

True consolidation of episodes was prevented by a previously unknown type of interference between episodic and (general) semantic memories coming from the ongoing semantic plasticity in neocortex. This interference could be countered by frequent hippocampal reactivation of the episodes. It was asymmetric in that storing new episodes had less effect on previously stored general information (data not shown), probably because the episodes conformed to the statistical structure observed by the network during semantic training. These results contrast sharply with standard forms of catastrophic interference[7,39], which concern representational competition between different semantic memories. Consolidation there is required to integrate new information with previously acquired, richly structured semantic knowledge[7]. However, the computational goal of episodic learning is storing individual events rather than discovering statistical structure, seemingly rendering consolidation inappropriate. If initial hippocampal storage of the episode already ensures that it can later be recalled episodically, then, barring practical advantages such as storage capacity (or perhaps efficiency), there seems little point in duplicating this capacity in neocortex.

Index maintenance

One might conclude from the previous section that, provided the hippocampus stores the essence of episodes permanently, as suggested in the multiple trace model[12], episodic memory will be unaffected by neocortical plasticity. However, this scenario was actually overly optimistic, as the statistics of the general neocortical patterns remained constant, with no refinement of the existing semantic representation, change in input statistics or acquisition of new semantic domains. All these can occur, to some extent, even in the face of hippocampal insult[4,40–44]. Such plasticity will change the cortical representations associated with past episodes. Wherever these episodes are stored, be it in the hippocampus, as discussed above, or the neocortex, semantic cortical plasticity will erode the relationship between inputs coded in the current representation and episodes coded in past representations.

Successful recall of an episode stored in the hippocampus depends in two ways on the correspondence between low- and high-level cortical areas embodied by the neocortical

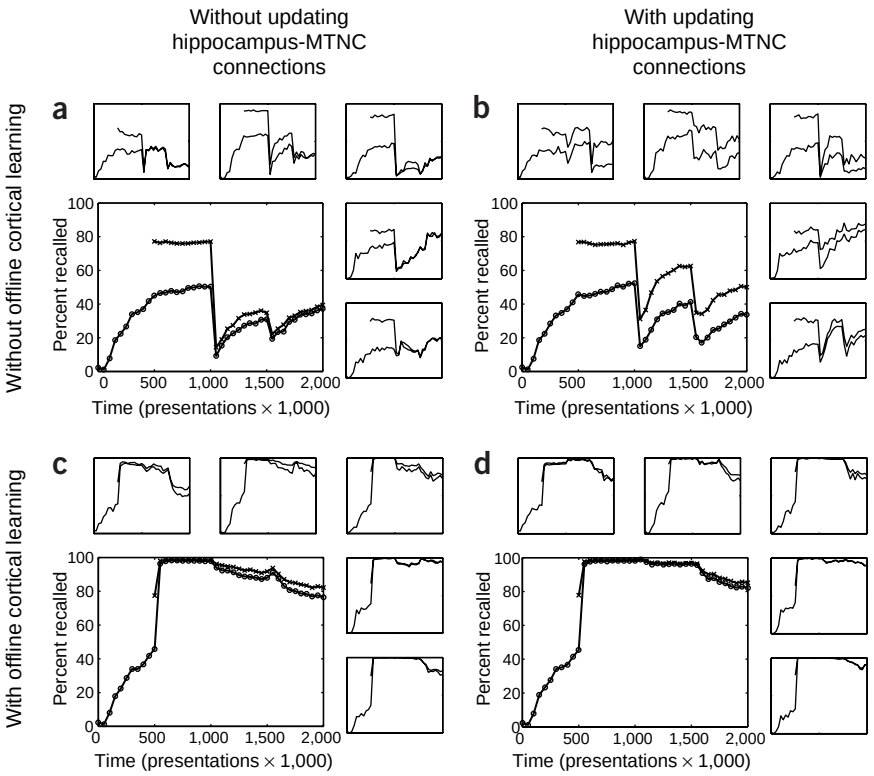

Figure 7 Neocortical plasticity during replay maintains semantic memory. (**a**–**d**) Graphs show the percentage of correct recall of the area C pattern associated with the recall cue in area A, averaged over all possible recall cues in area B. All testing was done within input domain 2. The larger graphs are averages over all possible area A patterns; the small graphs are for individual area A patterns. In all plots, the upper trace (marked with crosses in the main plots) is with hippocampal involvement in recall; the lower trace (marked with circles) is purely neocortical recall. Otherwise, training protocol, replay conditions and figure layout are as in **Figure 4**.

network. First, the high-level (MTNC) representation of the recall cue needs to be effective in activating the correct hippocampal memory trace; second, the high-level representation activated by hippocampal recall should effect the recall of the appropriate components of the corresponding episode in lower-level areas. Replay may prevent the damaging consequences of continued neocortical plasticity by maintaining the proper correspondence between hippocampal and neocortical representations.

To assess the effect of neocortical learning on the recall of previously stored episodes, we modified the presentation paradigm for patterns that we used in **Figure 2b**. For the remaining simulations, at the end of phase 1, 50 episodic patterns were stored in the hippocampus. Ten of these patterns were from the unstructured domain 1, and were used in all tests of episodic memory. The other 40 patterns included each of the 10 valid A-C pairings from the structured domain 2, presented together with 4 of the possible area B patterns—these allowed us later to test the effects of episodic storage and replay on semantic memory within the same setup. Changes in input statistics were implemented by appending phases 3 and 4 to the simulation. We first did this without replay, to establish a baseline, and then included replay (**Fig. 4**). Unlike the extreme biasing toward replay that was necessary to show (temporary) transfer (**Fig. 2b**), replay and experience were equally balanced. In all cases, the quality of recall for the stored episodes was monitored throughout.

Even though the episode remained perfectly stored in the hippocampus throughout and the MTNC-hippocampus pattern-matching process was just as described in the previous section, on average, neocortical learning came to erase the route to recall (**Fig. 4a**). There were two primary reasons for this. First, continued semantic learning after the storage of the episode caused its MTNC representation to move away from the version with which the stored hippocampal trace was associated (**Fig. 5a**). This drift was markedly enhanced by the substantial changes in neocortical representations that followed the expansion in the input patterns (at 1 million and 1.5 million presentations), whereupon even the full original episode may have failed to activate the corresponding hippocampal memory trace. The effect of representational change on hippocampally directed recall in the input areas was also considerable (**Fig. 5b**). In phase 4, even if the correct hippocampal trace became activated, the full episode could be successfully recalled only about 40% of the time, with a large variability between patterns.

Episodic memories were clearly fragile. To test how hippocampally initiated replay might help, input-driven training was interleaved with epochs of replay, assumed to take place during sleep. Replay happened just as described in the previous section (including the use of the cleanup connections within each lower cortical area, which effectively restricted activity to legitimate input patterns). With plasticity just in the neocortex, replay prevented degradation (**Fig. 4c**; compare **Fig. 4a**). Detailed analysis of the representational changes (analogous to the results in **Fig. 5**; data not shown) indicated that replay in our model worked by preventing the MTNC representations of episodes from changing, thereby assuring a continued perfect fit with the stored hippocampal trace.

However, it is very stringent, and potentially deleterious to neocortical capabilities, to constrain new neocortical learning such that the internal representation of all hippocampally stored episodes remains fixed. Therefore, we introduced a different sort of plasticity during replay. Here, once the stored representation of the episode was reactivated in MTNC and the representation of the episode was reconstructed in the lower-level areas, the present feed-forward mapping between the input areas and MTNC was used to determine the up-to-date MTNC representation of the episode. This MTNC pattern was then associated with the stored hippocampal episode that initiated the replay, so that the hippocampal and input-level representations of the episode were again in register. This also maintained episodic recall at a high level (**Fig. 4b**), despite substantial changes in the neocortical network. For replay to work, it was essential that the episodic patterns be representationally refreshed sufficiently frequently so that the hippocampus and MTNC remained tied.

Combining both forms of replay also resulted in good preservation of old episodes in the face of neocortical representational change (**Fig. 4d**). However, there was no obvious gain over the previous cases.

Acquisition and consolidation of semantic information

It is hotly debated whether different subtypes of declarative memory depend in similar ways on the hippocampus. We argued that episodic and semantic memory present quite different computational challenges: the rapid, interference-free learning capabilities of hippocampus seem especially relevant for episodic learning, whereas the slower, integrative plasticity of neocortex could by itself be sufficient to support semantic learning. However, substantial evidence from amnesic patients indicates that semantic memory can be strongly affected by hippocampal lesions (although generally less so than episodic memory)[10,42–45]. We consider two (nonexclusive) possibilities for the contribution of the hippocampus to semantic memory. First, the hippocampus might aid semantic recall through the episodic storage and retrieval of examples. Second, off-line replay of stored examples might facilitate the acquisition of (hippocampal-independent) semantic memories.

Semantic memory was measured by testing pattern completion in the structured domain 2. Recall that, within this domain, every pattern in area A has a single associated pattern in area C, independent of the pattern in area B. To test the degree to which the model acquired an internal representation of this statistical regularity (our simple model of semantic knowledge), all possible combinations of patterns in areas A and B (from within domain 2) were presented multiple times with random initial activity in area C, and we measured the percentage in which the semantically correct A-C pairing was recovered by the usual pattern-completion process (with hippocampal involvement when appropriate).

First, we examined the acquisition and consolidation of semantic memories in a paradigm similar to that of **Figures 2b** and **3**. In the baseline simulation, general neocortical training on patterns from domains 1 and 2 continued throughout, without any hippocampal involvement. We also tested three other conditions, with the hippocampus being involved either in recall, replay or both, for 50 episodic patterns from domains 1 and 2 stored at the end of phase 1. We monitored semantic recall during phases 1 and 2 and then during a subsequent 500,000 presentations of patterns just in domains 1 and 2 (**Fig. 6**). During the latter there was no hippocampal replay (to test the permanence of semantic consolidation—analogous to the procedure illustrated in **Fig. 3**).

Under these conditions semantic learning without hippocampal involvement was very slow and never reached high levels. Episodic storage and recall of examples in the hippocampus brought an immediate increase in semantic recall performance, going beyond the level expected from simply getting right those queries involving the exact patterns stored. Hippocampal replay of these examples resulted in a less immediate but equally pronounced gain in performance, even if the hippocampus was then disabled for testing. Enabling hippocampal pattern completion during recall led to a further substantial performance increase. Finally, in the replay con-

ditions, there was a slow, moderate decrease in performance once replay stopped. This was surprising because, unlike in the episodic case earlier, the pattern-completion task for the network remains statistically constant, so decay occurs despite the absence of contradictory input and the presence of some reinforcement. We believe that decay resulted from indirect competition from patterns in other domains (in our case, the more frequent 'background' patterns) for the internal representational resources of the neocortical network. However, this is not a direct capacity issue, because doubling the number of units in area MTNC did not affect the qualitative behavior (data not shown).

We also examined the effect of neocortical representational change on semantic memories. Semantic maintenance was measured in the same simulations (consisting of four phases) that were used to demonstrate episodic maintenance, but here we tested semantic memory on patterns from structured domain 2 in the way described above.

Replay could also help maintain semantic memories in the face of changes in input statistics (**Fig. 7**). Without replay, introduction of a new 'background' semantic domain disrupted the recall of semantic information from the structured domain (**Fig. 7a**). However, the extent to which the two different types of plasticity during replay that had been introduced earlier contributed to acquisition and maintenance was different from that seen for episodes. For episodes, either kind of replay by itself resulted in a similar high level of performance. For semantics, the best performance was afforded by combining during replay the update of neocortical-hippocampal connections with neocortical processing of the patterns recalled (**Fig. 7d**). However, as indicated in previous accounts of catastrophic interference in semantic memory[7], the most important contribution came from neocortical learning on the recalled patterns (which is absent in **Fig. 7b**) rather than the regular updating of the mapping between hippocampus and neocortex (absent in **Fig. 7c**).

DISCUSSION

The problem of maintaining access to, and readout from, memory traces in the light of representational change, as well as our replay-based solution, apply wherever the episodic memory is ultimately located. Partly because of the extreme imbalance of replay and experience epochs required to achieve adequate transfer in our model, we had episodes reside permanently in the hippocampus, consistent with one standard set of views of amnesia[11]. Nevertheless, our results may also be compatible with the view that episodic memories consolidate just like semantic ones[38], provided that neocortically stored episodes can be prevented from degrading by a non-hippocampal mechanism, such as active maintenance, 'freezing' certain neocortical synapses, or (as a result of MTL lesion) a pathological slowdown of neocortical plasticity. However, none of these mechanisms by itself alleviates the indexing problem (**Fig. 4a**).

Our model is consistent with evidence showing that the hippocampus can have an important role in normal semantic learning[44,45] and a temporary role in semantic recall[46,47]. As in previous accounts[7], replay allows the slow integration of information from episodes into neocortical memory systems. Semantic knowledge that remains consistent with the observed statistical structure of the world is not greatly affected by subsequent neocortical plasticity; this is in contrast with episodes, whose instability arises because they correspond to extremely peaked probability distributions that are not statistically representative. Thus, our model predicts that remote memories in patients with hippocampal lesions (and possibly also normal individuals) are semantic—even those that concern personally experienced events. This is broadly consistent with the available data[12,48]. In early phases of acquisition, the hippocampus can temporarily aid semantic memory through recall of episodes that provide possible answers to general semantic queries[34]. These instances of recall could even provide further opportunities for neocortical learning.

Replay may also serve other purposes—for instance, extending to semantically related stimuli the set of cues that can directly elicit the retrieval of an episode[49]. Hippocampal recall depends on the similarity between MTNC codes for cues and stored episodes. However, semantic relationships are implicit in neocortical synaptic weights, and only a fraction is directly reflected in MTNC codes. During replay, semantic cousins of existing episodes can be generated, and associations made between their MTNC representations and the hippocampal traces. Then, in normal operation, the episodes can subsequently be retrieved when the semantic associates are presented.

Different aspects of replay impose different requirements on the coordination between hippocampus and neocortex, and it is tempting, though highly speculative, to relate these to systematic differences found in different sleep phases. For instance, index extension requires extensive stochastic exploration of cortical semantic knowledge, which it is tempting to associate with rapid eye movement sleep. By contrast, autonomous reactivation of hippocampal memory traces leading to reactivation of the corresponding neocortical representation, which is a common element in all of our replay-based algorithms, may preferentially take place during slow-wave sleep. If so, slow-wave sleep may be required for the long-term maintenance of episodic memories and may (at least partially) underlie the hippocampal-dependent enhancement of semantic learning.

One direct experimental prediction is that episodic recall will degrade given continued cortical plasticity and no replay, irrespective of the state of neocortical consolidation of those episodes. Unfortunately, post-lesion forgetting rates for retrograde memories are rarely measured experimentally. Quantitative studies on the speed of neocortical learning in the face of hippocampal insult are also important to assess whether neocortical learning is slowed. Even without hippocampal insult, episodic recall should be fragile in the face of reversible blockade of activity or plasticity in the hippocampus for a prolonged period after acquisition, or selective blockade of hippocampally initiated replay (perhaps through sleep manipulations). Indeed, inactivation of AMPA and kainate glutamate receptors in the hippocampus disrupts the retention of spatial memory[50]. We predict that the degree of impairment should correlate with the degree of neocortical learning during the blockade and that, paradoxically, concurrent blockade of neocortical plasticity should help. Physiologically, we predict that replay should result in (perhaps rather subtle) changes in the connections between the hippocampus and entorhinal cortex, particularly the perforant path.

METHODS

Network architecture. Each of the four neocortical areas in our model (A, B, C and MTNC; see **Fig. 1**) had 100 binary units. Connections between areas in adjacent layers were all-to-all and symmetric. For each of A, B and C, 20 random binary vectors (denoted $\mathbf{x}^A{}_1$–$\mathbf{x}^A{}_{20}$, $\mathbf{x}^B{}_1$–$\mathbf{x}^B{}_{20}$ and $\mathbf{x}^C{}_1$–$\mathbf{x}^C{}_{20}$, each bit of which is turned on with probability 0.5) were generated to represent possible stimuli in the modality of that area. Activity in MTNC is denoted by $\mathbf{y}$.

Neocortical dynamics. This model network functions as a Boltzmann machine[25]. The Boltzmann machine uses weights and biases $W = \{\mathbf{W},\mathbf{w}\}$ to parameterize a probability distribution $P[\mathbf{x}^A,\mathbf{x}^B,\mathbf{x}^C; W]$ over the inputs, in such a way that the Monte Carlo sampling method called Gibbs sampling can be

used to perform inferences such as pattern completion. Because the within-area connections are not explicitly simulated, for the purposes of learning, the network becomes a restricted Boltzmann machine, for which units within each layer can be updated synchronously, so the dynamics of activity in the network consists of updates alternating between the two layers. Unit x_i in the input layer is set as

$$x_i = \begin{cases} 1 \text{ with probability} & \sigma\left(\sum_j W_{ij} y_j + w_i\right) \\ 0 \text{ with probablity} & 1 - \sigma\left(\sum_j W_{ij} y_j + w_i\right) \end{cases} \quad (1)$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the standard logistic sigmoid function; an analogous rule is used for units in area MTNC. Our model of the local cleanup connections specifies that, in the absence of feed-forward input, if the local activity pattern is within a bitwise Hamming distance of five of one of the previously experienced input patterns in that area, the activation pattern is changed (locally) to that pattern.

Neocortical learning. During input-driven activity in the network, and also during replay episodes when neocortical plasticity is enabled, the weights W are updated using Hinton's recent modification[27] of the standard Boltzmann machine learning rule[25], which involves two phases. In the Hebbian phase, a complete input pattern is presented to the input layer; the corresponding activities in MTNC are determined stochastically according to the equivalent of equation (1); and the weights between the two layers are increased in proportion to the product of the activities of the nodes connected. The units in the input layer and then once more in the MTNC layer are then sampled, and, in the anti-Hebbian phase, the weights are decreased in proportion to the product of these new, internally generated activities. In this simple version of the architecture, weights are allowed to take both positive and negative real values.

ACKNOWLEDGMENTS

We thank S. Becker, N. Burgess, M. Lengyel, J.L. McClelland and A.D. Wagner for their extensive comments on earlier drafts. Funding was from the Hungarian Academy of Sciences (S.K.) and the Gatsby Charitable Foundation (S.K. and P.D.).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 4 November 2003, accepted 16 January 2004
Published online at http://www.nature.com/natureneuroscience/

1. Grossberg, S. Nonlinear neural networks: principles, mechanisms, and architectures. *Neural Netw.* **1**, 17–61 (1988).
2. Freud, S. The archaic features and infantilism of dreams. in *Introductory Lectures on Psychoanalysis* (ed. & trans. Strachey, J.) 199–212 (Norton, New York, 1966; original publication 1916).
3. Squire, L.R. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychol. Rev.* **99**, 195–231 (1992).
4. Vargha-Khadem, F. *et al.* Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277**, 376–380 (1997).
5. Manns, J.R., Hopkins, R.O. & Squire, L.R. Semantic memory and the human hippocampus. *Neuron* **38**, 127–133 (2003).
6. Alvarez, P. & Squire, L.R. Memory consolidation and the medial temporal lobe: a simple network model. *Proc. Natl. Acad. Sci. USA* **91**, 7041–7045 (1994).
7. McClelland, J.L., McNaughton, B.L. & O'Reilly, R.C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
8. Winocur, G. Anterograde and retrograde amnesia in rats with dorsal hippocampal or dorsomedial thalamic lesions. *Behav. Brain Res.* **38**, 145–154 (1990).
9. Zola-Morgan, S.M. & Squire, L.R. The primate hippocampal formation: evidence for a time-limited role in memory storage. *Science* **250**, 288–290 (1990).
10. Rempel-Clower, N.L., Zola, S.M., Squire, L.R. & Amaral, D.G. Three cases of enduring memory impairment after bilateral damage limited to the hippocampal formation. *J. Neurosci.* **16**, 5233–5255 (1996).
11. Nadel, L. & Moscovitch, M. Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* **7**, 217–227 (1997).
12. Nadel, L., Samsonovich, A., Ryan, L. & Moscovitch, M. Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. *Hippocampus* **10**, 352–368 (2000).
13. Pavlides, C. & Winson, J. Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *J. Neurosci.* **9**, 2907–2918 (1989).
14. Wilson, M.A. & McNaughton, B.L. Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676–679 (1994).
15. Siapas, A.G. & Wilson, M.A. Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep. *Neuron* **21**, 1123–1128 (1998).
16. Louie, K. & Wilson, M.A. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron* **29**, 145–156 (2001).
17. Hoffman, K.L. & McNaughton, B.L. Coordinated reactivation of distributed memory traces in primate neocortex. *Science* **297**, 2070–2073 (2002).
18. Stickgold, R. Sleep: off-line memory reprocessing. *Trends Cogn. Sci.* **2**, 484–492 (1998).
19. Hobson, J.A. & Pace-Schott, E.F. The cognitive neuroscience of sleep: neuronal systems, consciousness and learning. *Nat. Rev. Neurosci.* **3**, 679–693 (2002).
20. Plihal, W. & Born, J. Effects of early and late nocturnal sleep on declarative and procedural memory. *J. Cogn. Neurosci.* **9**, 534–547 (1997).
21. Rao, R.P.N., Olshausen, B.A. & Lewicki, M.S. (eds.). *Probabilistic Models of the Brain: Perception and Neural Function* (MIT Press, Cambridge, Massachusetts, USA, 2002).
22. Felleman, D.J. & Van Essen, D.C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
23. Lavenex, P. & Amaral, D.G. Hippocampal-neocortical interaction: a hierarchy of associativity. *Hippocampus* **10**, 420–430 (2000).
24. Plaut, D.C. *Connectionist Neuropsychology: The Breakdown and Recovery of Behavior in Lesioned Attractor Networks.* PhD thesis, Carnegie Mellon Univ. (1991). Available as Technical Report CMU-CS-91-185.
25. Hinton, G. & Sejnowski, T.J. Learning and relearning in Boltzmann machines. in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations* (eds. Rumelhart, D.E. & McClelland, J.L.) 282–317 (MIT Press, Cambridge, Massachusetts, USA, 1986).
26. Hinton, G. & Sejnowski, T.J. (eds.). *Unsupervised Learning* (MIT Press, Cambridge, Massachusetts, USA, 1999).
27. Hinton, G.E.. Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14**, 1771–1800 (2002).
28. Marr, D. Simple memory: a theory for archicortex. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **262**, 23–81 (1971).
29. McNaughton, B.L. & Morris, R.G.M. Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends Neurosci.* **10**, 408–415 (1987).
30. Treves, A. & Rolls, E.T. Computational analysis of the role of the hippocampus in memory. *Hippocampus* **4**, 374–391 (1994).
31. O'Reilly, R.C. & McClelland, J.L. Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus* **4**, 661–682 (1994).
32. Hasselmo, M.E., Wyble, B.P. & Wallenstein, G.V. Encoding and retrieval of episodic memories: role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus* **6**, 693–708 (1996).
33. Shen, B. & McNaughton, B.L. Modeling the spontaneous reactivation of experience-specific hippocampal cell assembles during sleep. *Hippocampus* **6**, 685–692 (1996).
34. O'Reilly, R.C. & Rudy, J.W. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychol. Rev.* **108**, 311–345 (2001).
35. Cho, Y.H. & Kesner, R.P. Involvement of entorhinal cortex or parietal cortex in long-term spatial discrimination memory in rats: retrograde amnesia. *Behav. Neurosci.* **110**, 436–442 (1996).
36. Wiig, K.A., Cooper, L.N. & Bear, M.F. Temporally graded retrograde amnesia following separate and combined lesions of the perirhinal cortex and fornix in the rat. *Learn. Mem.* **3**, 313–325 (1996).
37. Anagnostaras, S.G., Maren, S. & Fanselow, M.S. Temporally graded retrograde amnesia of contextual fear after hippocampal damage in rats: within-subjects examination. *J. Neurosci.* **19**, 1106–1114 (1999).
38. Squire, L.R., Clark, R.E. & Knowlton, B.J. Retrograde amnesia. *Hippocampus* **11**, 50–55 (2001).
39. McCloskey, M. & Cohen, N.J. Catastrophic interference in connectionist networks: the sequential learning problem. in *The Psychology of Learning and Motivation* vol. 24 (ed. Bower, G.) 109–165 (Academic Press, New York, 1989).
40. Glisky, E.L., Schacter, D.L. & Tulving, E. Learning and retention of computer-related vocabulary in memory-impaired patients: method of vanishing cues. *J. Clin. Exp. Neuropsychol.* **8**, 292–312 (1986).
41. Tulving, E., Hayman, C.A. & Macdonald, C.A. Long-lasting perceptual priming and semantic learning in amnesia: a case experiment. *J. Exp. Psychol. Learn. Mem. Cogn.* **17**, 595–617 (1991).
42. Kitchener, E.G., Hodges, J.R. & McCarthy, R. Acquisition of post-morbid vocabulary and semantic facts in the absence of episodic memory. *Brain* **121** (Pt. 7),1313–1327 (1998).
43. Bayley, P.J. & Squire, L.R. Medial temporal lobe amnesia: gradual acquisition of factual information by nondeclarative memory. *J. Neurosci.* **22,** 5741–5748 (2002).

44. Holdstock, J.S., Mayes, A.R., Isaac, C.L., Gong, Q. & Roberts, N. Differential involvement of the hippocampus and temporal lobe cortices in rapid and slow learning of new semantic information. *Neuropsychologia* **40**, 748–768 (2002).
45. Hamann, S.B. & Squire, L.R. On the acquisition of new declarative knowledge in amnesia. *Behav. Neurosci.* **109**, 1027–1044 (1995).
46. Verfaellie, M., Reiss, L. & Roth, H.L. Knowledge of New English vocabulary in amnesia: an examination of premorbidly acquired semantic memory. *J. Int. Neuropsychol. Soc.* **1**, 443–453 (1995).
47. Haist, F., Bowden Gore, J. & Mao, H.. Consolidation of human memory over decades revealed by functional magnetic resonance imaging. *Nat. Neurosci.* **4**, 1139–1145 (2001).
48. Bayley, P.J., Hopkins, R.O. & Squire, L.R. Successful recollection of remote autobiographical memories by amnesic patients with medial temporal lobe lesions. *Neuron* **38**, 135–144 (2003).
49. Káli, S. & Dayan, P. Replay, repair and consolidation. in *Advances in Neural Information Processing Systems 15* (eds. Becker, S., Thrun, S. & Obermayer, K.) 19–26 (MIT Press, Cambridge, Massachusetts, USA, 2003).
50. Riedel, G. *et al.* Reversible neural inactivation reveals hippocampal participation in several memory processes. *Nat. Neurosci.* **2,** 898–905 (1999).