PNAS SPECIAL FEATURE | PSYCHOLOGICAL AND COGNITIVE SCIENCES
NEUROSCIENCE

A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation

Dhairyya Singh[a,1], Kenneth A. Norman[b,c], and Anna C. Schapiro[a,1]

Edited by Katharine Simon, University of California, Irvine, CA; received January 31, 2022; accepted August 11, 2022 by Editorial Board Member Henry L. Roediger III

How do we build up our knowledge of the world over time? Many theories of memory formation and consolidation have posited that the hippocampus stores new information, then "teaches" this information to the neocortex over time, especially during sleep. But it is unclear, mechanistically, how this actually works—How are these systems able to interact during periods with virtually no environmental input to accomplish useful learning and shifts in representation? We provide a framework for thinking about this question, with neural network model simulations serving as demonstrations. The model is composed of hippocampus and neocortical areas, which replay memories and interact with one another completely autonomously during simulated sleep. Oscillations are leveraged to support error-driven learning that leads to useful changes in memory representation and behavior. The model has a non–rapid eye movement (NREM) sleep stage, where dynamics between the hippocampus and neocortex are tightly coupled, with the hippocampus helping neocortex to reinstate high-fidelity versions of new attractors, and a REM sleep stage, where neocortex is able to more freely explore existing attractors. We find that alternating between NREM and REM sleep stages, which alternately focuses the model's replay on recent and remote information, facilitates graceful continual learning. We thus provide an account of how the hippocampus and neocortex can interact without any external input during sleep to drive useful new cortical learning and to protect old knowledge as new information is integrated.

neural network model | oscillations | sleep stages | continual learning

Significance

Sleep is known to be an important time for consolidating our memories, with memory systems in the brain replaying recent memories and stabilizing them for long-term storage. We present a computational model that simulates these interactions between hippocampal and neocortical memory systems, providing an account of how they can autonomously produce useful offline learning. We explain how information stored initially in the hippocampus can help to build neocortical representations during sleep, and how the alternation of different stages of sleep across the night can facilitate graceful integration of new information with existing knowledge.

Building our knowledge of the world over time requires the ability to quickly encode new information as we encounter it, store that information in a form that will serve us well in the long term, and carefully integrate the new information into our existing knowledge structures. These are difficult tasks, replete with pitfalls and trade-offs, but the brain seems to accomplish them gracefully. The Complementary Learning Systems (CLS) framework proposed that the brain achieves these feats through a division of labor across two interacting systems: The hippocampus encodes new information using a sparse, pattern-separated code, supporting rapid acquisition of arbitrary information without interference with existing neocortical knowledge (1). The hippocampus then replays this recently acquired information offline, gradually "teaching" this information to the neocortex. The neocortex uses overlapping, distributed representations adept at representing the structure across these memories, resulting in the construction and updating of semantic knowledge over time. But how can these brain regions interact autonomously, with no input from the environment, to produce useful learning and reshaping of representations? How does the brain move from one memory to another during offline replay, and which of these offline states does it learn from, and how?

CLS considered the possibility that sleep may be a useful time for this teaching to occur, and other perspectives have also focused in on this idea, given the strong coupled dynamics and parallel replay that occurs in these areas during sleep (2–4). Extant theories of consolidation have focused particularly on stages 2 and 3 of non–rapid eye movement (NREM) sleep, when nested oscillations associated with memory replay—hippocampal sharp wave ripples, thalamocortical spindles, and neocortical slow oscillations—reflect especially strong hippocampal–cortical interaction (2–11). These dynamics appear to be causally involved in memory consolidation: Manipulations that enhance hippocampal–cortical synchrony during sleep benefit memory (12–16).

Not all theories agree on whether offline hippocampal–cortical interactions serve to increase the relative reliance on neocortex for episodic memories (17, 18), but most theories agree that the hippocampus helps to build and shape semantic representations in neocortex (19, 20), and these theories often assign a central role to active processing during sleep (ref. 21; cf. ref. 22). Here, we adopt the following core ideas, shared across several perspectives: During sleep, the hippocampus actively helps to build neocortical semantic

Author affiliations: [a]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104; [b]Department of Psychology, Princeton University, Princeton, NJ 08540; and [c]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540

Author contributions: D.S., K.A.N., and A.C.S. designed research; D.S. and A.C.S. performed research; D.S. and A.C.S. analyzed data; and D.S., K.A.N., and A.C.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. K.S. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: dsin@sas.upenn.edu or aschapir@sas.upenn.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2123432119/-/DCSupplemental.

Published October 24, 2022.

representations of information it has recently encoded, and this process transforms the nature of the memories. We assume that learning does occur directly in the cortex during initial awake experience, but that it is not typically strong enough to support behavior without offline hippocampal influence (23). We think this process of the hippocampus helping to shape and strengthen neocortical memory representations can usefully be referred to as "systems consolidation" (24), although the way it is conceptualized here deviates in several respects from Standard Systems Consolidation Theory (17).

Despite all of this evidence and theorizing that the hippocampus "teaches" the neocortex new information in order to build up semantic representations during sleep, it is unclear, mechanistically, how this occurs in autonomous offline interactions. It does not seem likely that simple strengthening of existing connections through Hebbian learning will be sufficient—a more sophisticated restructuring of representations seems to be at work. The original CLS model simulations (1) did not address these questions, since there was no true implemented hippocampus (cf. ref. 25), and, as in many models of replay, the order and nature of inputs to cortex were engineered by hand. Here, we aim to fill this gap with a model that shows how an implemented cortex and hippocampus can interact, replay, and learn autonomously.

We will address two additional related limitations of the original CLS framework, namely, that its solution to continual learning was too slow, and that it relied on the implausible assumption of a stationary environment (26). Hippocampal replay of an episode—for instance, a first encounter with a penguin—was hypothesized to be interleaved over days, weeks, and months with a stationary (unchanging) distribution of bird input from the environment, allowing careful integration of the lone penguin encounter with the general structure of birds. This strategy is slow because it relies on these continued reminders over time about the distribution of information from the environment. But sleep-dependent memory consolidation, including integration of new information with existing knowledge, can impact behavior quite quickly, over one night or several nights of sleep (6, 7, 27–30). The CLS framework fails in nonstationary environments because the environment no longer provides those required reminders of old information. Humans do not have this problem: We can, for example, speak one language much of our lives and then move to a new place where we encounter only a new language, without forgetting the first language. Norman et al. (26) proposed that alternating NREM and REM sleep stages across the course of a night may be the key to solving these problems. The hippocampus and neocortex are less coupled during REM (2), potentially providing an opportunity for the neocortex to visit and explore its existing representational space. This could serve to provide the reminders about remote knowledge needed to avoid new information overwriting the old, without having to wait for the environment to provide those reminders.

We will present proof-of-concept simulations that demonstrate 1) how the hippocampus can begin to autonomously shape neocortical representations during NREM sleep and 2) how alternating NREM/REM sleep stages allows for rapid integration of that new information into existing knowledge. The hippocampus is implemented as our C-HORSE model (Complementary Hippocampal Operations for Representing Statistics and Episodes), which is able to quickly learn new categories and statistics in the environment, in addition to individual episodes (31–33), allowing for a more complete treatment of initial hippocampal learning than has been considered in prior hippocampal replay models.

Once structured information is encoded in the hippocampus, how is it then consolidated? We extend the model with a neocortical area that serves as the target of consolidation and with a sleep environment that allows learning and dynamics between these areas to unfold autonomously offline (Fig. 1). Our learning scheme leverages oscillations during sleep to support self-supervised error-driven learning (26). Error-driven learning typically involves computing the discrepancy between the model's prediction and the actual state of the environment, and then adjusting weights to minimize this error; this ability to contrast "better" and "worse" states and to learn from these discrepancies allows for more sophisticated representation shaping than Hebbian learning, which updates weights between neurons based on simple coactivity, rather than contrasting two states (26, 34). This raises the question of how error-driven learning might be accomplished during sleep, when the brain cannot use the actual state of the external environment as a target for learning. Here, we show how stable patterns of internal activity during sleep can serve as effective targets for offline error-driven learning. A short-term synaptic depression mechanism supports autonomous transitions from one memory to the next, avoiding the need to hand-engineer the model's inputs and their ordering.

The model's units employ a rate code, which means that oscillations do not have a true frequency that would correspond directly to particular frequencies of activity in the sleeping brain. But we imagine that our NREM oscillations correspond in function to sleep spindles, when individual replay events and enhanced cortical plasticity have been shown to occur (35–37), and our REM oscillations correspond to theta oscillations,

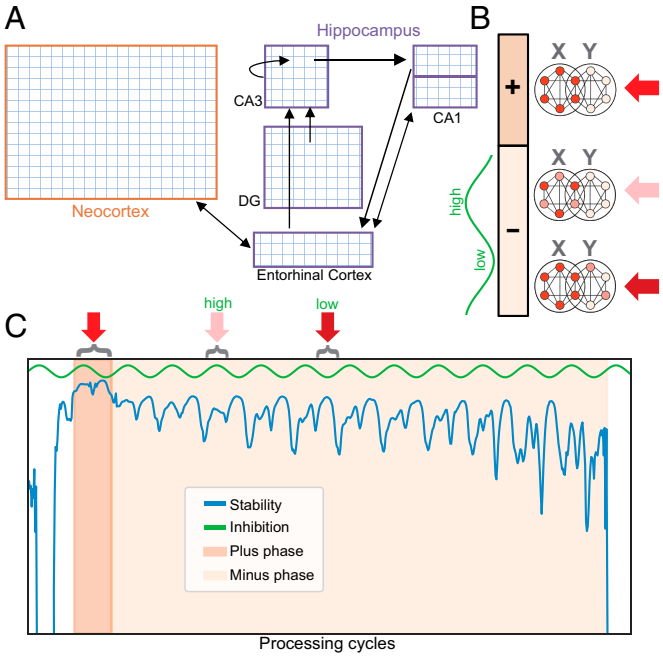

Fig. 1. Model architecture and sleep algorithm. (*A*) The architecture included C-HORSE (our model of the hippocampus; purple outlines), and a neocortical layer as the target of consolidation (orange outline). (*B*) When replaying attractor X, all units participating in the attractor have strong, stable activity during the plus phase. During the minus phase, oscillatory phases of higher inhibition lead to lower activity, with the weakest units in attractor X dropping out, and oscillatory phases of lower inhibition lead to higher activity, causing activity spreading to a nearby attractor, Y. (*C*) Stability trace for a real learning event, with background colors indicating the plus and minus phases. When the model first falls into an attractor, its activity is highly stable and triggers the plus phase. Short-term synaptic depression gradually destabilizes the attractor. The stability drop causes the plus phase to end, and the minus phase begins. As the attractor further destabilizes, the minus phase ends.

which have been associated with enhanced plasticity in that stage (26, 38, 39). We simulate the difference between NREM and REM sleep as a difference in the degree of communication between the hippocampus and neocortex (2). There are, of course, many other important differences between NREM and REM sleep (3), but, for these simulations, we aimed to isolate the contribution of this factor. Together, the simulations demonstrate how sleep can build our neocortical knowledge over time, with the hippocampus initially helping to construct neocortical representations during NREM sleep, and the hippocampus and cortex working together across NREM/REM stage alternation to integrate this new information with existing knowledge.

Model Simulations

In the model, sleep is initiated with a single injection of noise to all units, after which all external input is silenced. The model then autonomously moves from attractor to attractor corresponding to the memories of items learned during wake [exhibiting latching dynamics (26, 40)]. A "memory" is thus defined as the pattern of stable activity across units in the model corresponding to an item experienced during training. We use stability in the model's dynamics as a trigger to initiate a learning trial. When the model initially falls into each attractor state, activity tends to be highly stable from one processing cycle to the next, and this high stability initiates a "plus" phase (Fig. 1 *B* and *C*). Short-term synaptic depression destabilizes the attractor by temporarily weakening synapses in proportion to the coactivity of their connected units. As stability drops, the model transitions into a "minus" phase. Oscillations in inhibition levels are persistently present in the sleeping model, with all the layers of the model on a synchronized oscillatory schedule. However, the effects of the oscillating inhibition on activation become more prominent in the minus phase, as synaptic depression begins to distort an attractor. These oscillations in inhibition levels reveal aspects of the attractor that can be improved: When inhibition is high, the weakest units in the attractor drop out, revealing the parts of the memory that would benefit from strengthening, and, when inhibition is low, nearby potentially interfering memories or spurious associations become active, revealing competitors whose weakening would help stabilize the attractor (Fig. 1 *B* and *C*).

As the attractor activity further destabilizes, the minus phase ends, and the model traverses to the next attractor. Weights are updated at the end of every minus phase via Contrastive Hebbian Learning (CHL), which locally adjusts minus phase unit coactivities (product of sender and receiver activity over time) toward their plus phase coactivities (41). This serves to both strengthen weak parts of an attractor (the contrast between the plus phase and the higher-inhibition moments in the minus phase) and weaken competitors (the contrast between the plus phase and the lower-inhibition moments of the minus phase). Once competitors are sufficiently weakened to avoid high interference, they no longer appear in the low-inhibition phase and are not suppressed any further. The learning algorithm can thus balance both overlap and separation in a manner that reflects the structure of the input. Learning in a hidden layer with distributed representations will tend to emphasize overlap, but the low-inhibition phase of the oscillation serves as a check on creating too much of this overlap. This algorithm is closely related to the Oscillating Learning Rule (OLR) (26), which used oscillations to similarly distort attractors and improve memory, but the OLR requires the network to track where it is in an oscillation and change the sign of the learning rule accordingly. The current learning scheme does not require tracking this information (but does require a measure of stability). Our algorithm's short-term synaptic depression-induced traversal across memory attractors is also related to previously proposed "latching" dynamics in cortical networks (40).

Simulation 1: Building Neocortical Representations of Novel Information. In this simulation, we explored how the hippocampus helps to rapidly shape neocortical representations of novel information. As a demonstration, we simulated a paradigm in which we have found sleep effects that influence behavior quickly (across a night of sleep or even a nap) and that have hallmarks of the construction of new neocortical knowledge (29, 42). We chose this study because it was representative of a larger class of studies demonstrating a benefit of sleep on memory for novel structured information (42); we did not aim to target the details of the findings of this particular study, although that would be a useful enterprise for future work. In the experiment, participants were asked to learn the features of 15 satellites belonging to three different classes (Fig. 2*A* and ref. 29). Each satellite exemplar had features shared with other members of the class, as well as features unique to the exemplar. As in the human participant sleep group, the model was first trained to a learning criterion of 66% and then allowed to sleep, with tests before and after sleep to assess changes in performance (Figs. 2*C* and 3*A*). In this simulation, sleep corresponds to the dynamics of NREM, with tight coupling between hippocampal and neocortical areas.

The model was trained during wake using feature inference, which is analogous to how humans were trained in the experiment. Each satellite's seven features (five visual, class name, and codename) were presented on separate input/output layers (representing entorhinal cortex) as inputs to the hidden layers of the model. Each training trial consisted of presenting the model with a satellite with one feature missing, at which point the model guessed the identity of the missing feature. The correct answer was then clamped onto the corresponding layer. CHL was used to update the weights based on the difference between the model's prediction and the correct answer (41). As in our experiments with human participants (29, 41), we sought to match model performance on shared and unique features during wake training by querying unique features much more often than shared features. The learning rate was much higher in the hippocampus than cortex, so—while learning happened

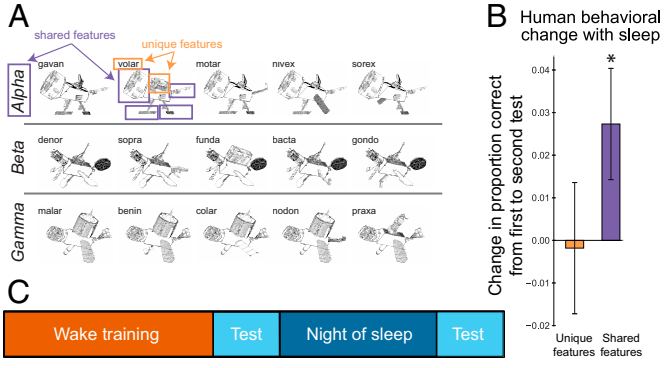

Fig. 2. Schapiro et al. (29) category learning paradigm and results. (*A*) Satellite exemplars studied from three categories. (*B*) Performance change for unique and shared features from presleep to postsleep tests. *$P < 0.05$. (Results are collapsed across the frequency manipulation included in that study.) (*C*) Experimental protocol: Participants studied the satellites in the evening, had an immediate test, and were tested again 12 h later, after a night of sleep.

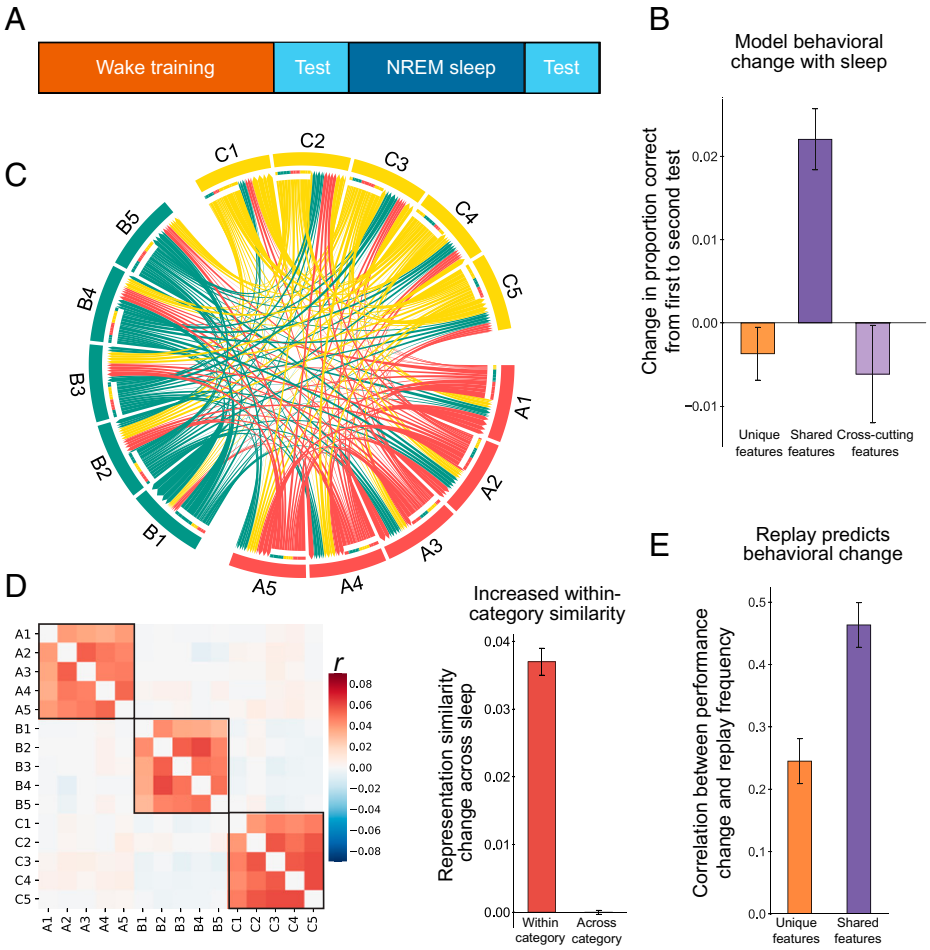

Fig. 3. Simulation 1: Category learning consolidation across one night of sleep. (*A*) Simulation protocol. (*B*) Change in full model performance from before to after sleep for unique features, shared features, and cross-cutting features. (*C*) Chord plot of replay transitions, with A, B, and C indicating categories. Arrow base indicates transition initiation, and head indicates termination. Arrow width indicates proportion of replays for a given satellite. The plot shows that replay was interleaved uniformly across exemplars (with a modest preference for within-category transitions). (*D*) (*Left*) Change in pairwise correlation of neocortical representations across items from before to after sleep. Black boxes indicate within-category changes. (*Right*) Averaged pairwise correlation change of neocortical representations within and across categories. (*E*) Fisher-transformed correlation between exemplar replay frequency during sleep and performance change, averaged across model initializations. All error bars represent ±1 SEM across network initializations.

during wake throughout the model—behavior was initially supported almost entirely by the hippocampus. Cortical representations were, on average, less sparse than hippocampal representations and reflected category structure more strongly. For this paradigm, which requires integrating high-level verbal and visual information, the neocortical module might correspond to an area of the brain representing consolidated semantic knowledge, like the anterior temporal lobe (43). Anterior cingulate cortex/medial prefrontal cortex are also often involved as important neocortical targets of systems consolidation (44, 45).

After the wake learning criterion was achieved, we allowed the model to go into NREM sleep, enabling hippocampal–cortical replay. Based on evidence for enhanced plasticity in sleep (46), we increased the cortical learning rate in the model during NREM sleep. To isolate the impact of replay on neocortical learning, we paused hippocampal learning during sleep. The model traversed from attractor to attractor corresponding to the studied satellite exemplars. Replay occurred concurrently in the hippocampus and neocortical modules. We found that replay was interleaved across exemplars, with a modest preference for within-category transitions (39.4% within-category transitions relative to chance 28.5%; $P < 0.001$ by binomial test [$N = 2{,}400$, $K = 945$]; Fig. 3*C*). We assessed performance changes as a result of NREM sleep and found that performance on shared features improved significantly ($P < 0.001$) while performance on unique features was preserved across sleep (Fig. 3*B*), which corresponds to the pattern observed in the human experiment shown in Fig. 2*B* as well as the behavioral data from our imaging study using this paradigm (42). These results depended on using an error-driven learning algorithm; simple Hebbian learning during sleep resulted in substantial forgetting of unique information and less robust increases in shared feature memory (*SI Appendix*, Fig. S1).

The preferential benefit for shared features could potentially be related to their greater frequency of occurrence, given that they appear many times across exemplars, whereas unique features only appear with one exemplar. To rule out the possibility that sleep is simply benefitting higher-frequency features, we ran a separate set of simulations in which we added an extra feature to each exemplar that appeared across categories but was matched in frequency to the shared features. We found that these "cross-cutting" features did not benefit from sleep replay (Fig. 3*B*), suggesting that neocortical learning was sensitive to the true category structure, rather than simply frequency of feature appearance.

For each simulation, we assessed whether the number of times a particular satellite was replayed was associated with improvement in memory for that satellite. This resulted in a correlation across 15 exemplars for each simulation, which we Fisher

transformed and averaged across simulations. We found that, across simulations, there was a positive correlation between performance changes and replay frequency (Fig. 3*E*), which was especially strong for shared features, indicating that replay during sleep was driving performance changes. This is consistent with the results from our imaging study of this paradigm, where we found that awake replay of the satellites in the hippocampus (which we took to be representative of what would continue to happen during sleep) predicted memory improvement across sleep (42).

We next examined change in neocortical representations across sleep. We conducted a representation similarity analysis by calculating pairwise correlations of patterns of neocortical unit activity for all satellite exemplars presleep and postsleep. We found that, from before to after sleep, there was increased within-category similarity and no change in across category similarity (Fig. 3*D*).

The results from simulation 1 provide a way of thinking about how the hippocampus can help the neocortex build up new semantic representations. With a fast learning rate in the hippocampus and a slow rate in the neocortex during wake learning, the hippocampus will be responsible for most behavior prior to sleep. Offline, the hippocampus can help the neocortex to reinstate stable versions of the new memories. Because the neocortex has more highly distributed representations than the hippocampus, it tends to find the shared structure across exemplars, resulting in improved understanding of the category structure in this paradigm. These dynamics occur completely autonomously, with three key ingredients: 1) Short-term synaptic depression causes the hippocampus and neocortex to move in tandem from one attractor to the next, resulting in interleaved replay of recent experience; 2) as synaptic depression destabilizes an attractor, oscillations increasingly distort the attractor to reveal aspects of the memory in need of change; and 3) the model learns toward highly stable states and away from the subsequent, relatively unstable states, resulting in memory shifts and strengthening.

Simulation 2: Integration of Recent and Remote Knowledge. We next turned to simulating the potential role of sleep in supporting continual learning. We defined two related environments, Env 1 and Env 2. Each environment contained 10 items, with seven units each. Two out of seven units overlapped between the first item in Env 1 and the first item in Env 2, between the second item in Env 1 and the second item in Env 2, and so on. There was no overlap across items within an environment. This is a variant of the classic AB–AC interference paradigm (47–50). To simulate new learning in Env 2 after having fully consolidated Env 1 in neocortex, we trained the neocortical hidden layer fully on Env 1. Next, the complete network was trained on the related Env 2, with hippocampus primarily supporting performance, as in simulation 1, given its faster learning rate. We then allowed the model to go into either an alternating sleep protocol (alternating five times back and forth between epochs of NREM and REM sleep) or into five sequential NREM epochs or five sequential REM epochs (Fig. 4*A*). NREM was implemented as above, with strong coupled dynamics between the hippocampus and neocortex, whereas REM allowed neocortex to explore its attractor space without influence from the hippocampus (for simplicity, we allowed no hippocampal influence).

We expected that Env 2 replay would unfold during NREM sleep in a manner analogous to simulation 1, with the hippocampus helping the neocortex to replay and learn the Env 2 patterns. During REM, with no influence from the hippocampus, we expected that the neocortex would be able to replay the well-consolidated remote Env 1 memories. This could serve to provide reminders about Env 1 to prevent Env 2 from overwriting Env 1. We observed, as expected, that NREM tended to focus on Env 2 items (lower distance from plus phase patterns to Env 2 items; Fig. 4*B*) whereas REM tended to focus on the remote Env 1 items.

We found that alternating NREM and REM allowed the neocortex to improve Env 2 performance without a catastrophic decrease in Env 1 performance (Fig. 4*C*). Consecutive epochs of NREM sleep resulted in greater improvement in Env 2 performance, but also led to a much more substantial deterioration in Env 1 performance. As expected, consecutive epochs of REM sleep resulted in no improvement in Env 2 performance, as well as no Env 1 damage. We also investigated whether the mere presence of REM sleep is enough for preserving Env 1 performance or whether the alternating schedule is critical. To test this, we ran a blocked schedule of five NREM epochs followed by five REM epochs. We predicted that the blocked schedule would not allow for the same level of Env 1 preservation. As predicted, we found that the blocked schedule led to substantially greater Env 1 damage in comparison to the alternating schedule (*SI Appendix*, Fig. S2), indicating that

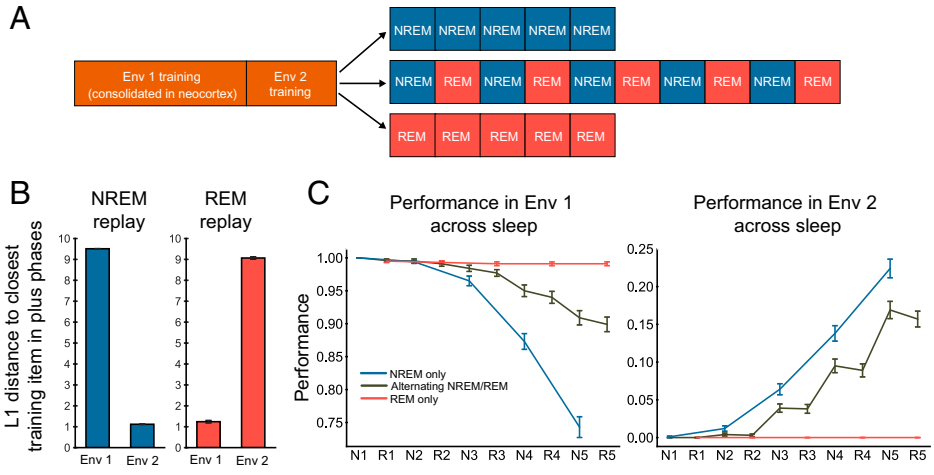

Fig. 4. Simulation 2: Continual learning via alternating sleep stages. (*A*) Simulation protocol. (*B*) L1 distance of plus phase attractors to Env 1 and Env 2 item patterns in REM and NREM. Smaller distance indicates more replay of those patterns. (*C*) Mean performance across sleep for Env 1 and Env 2 items for NREM-only vs. alternating NREM/REM (N/R) vs. REM-only conditions.

the alternating schedule is necessary for Env 1 protection—the alternating schedule allows REM to intermittently reinstate clean Env 1 items and repair any NREM-induced damage to these items before the damage accumulates beyond repair.

We also sought to characterize the role oscillations play in these simulations. We ran our simulations with the same task and either turned off oscillations completely during sleep (*SI Appendix*, Fig. S3) or selectively turned off oscillations in either NREM or REM epochs in the alternating condition (*SI Appendix*, Fig. S4). We found that, without oscillations, NREM sleep is substantially worse at facilitating cortical learning of Env 2 items, and that REM is impaired in its ability to protect Env 1 performance, indicating that oscillations play an important role in our framework.

Alternating NREM and REM may thus support graceful continual learning, with NREM helping to build neocortical representations of recent information while REM acts to protect the old information from this new potential interference (51). This alternation should be especially important to the degree that new and old information overlap and therefore threaten interference; when new information is unrelated to prior knowledge (as in simulation 1), reminders of old knowledge are likely not needed (52).

Discussion

Our simulations demonstrate how the hippocampus and neocortex can interact autonomously offline to build new neocortical knowledge and to gracefully integrate new information with existing cortical knowledge. Simulation 1 demonstrates how the hippocampus can begin to shape neocortical knowledge during a first bout of NREM sleep. We think of this as the very beginning of systems consolidation, not the complete process. Still, one night of sleep (or even a nap) is sufficient to appreciably change behavior in many paradigms, including the task simulated here (29, 42). Sleep begins by seeding the network with random activity. The network then falls into an attractor that happens to resemble that activity, and moves from attractor to attractor due to a short-term synaptic depression mechanism which weakens synapses as a function of their recent activity. We find that this mechanism results in interleaved replay of attractors, which sets the stage for useful structure learning (1). When the network first falls into an attractor, activity is highly stable, and the network uses these stable states as targets to learn toward. Oscillations then distort these attractors, revealing contrasting states to learn away from—both weak parts of an attractor in need of strengthening and competing attractors in need of weakening. During NREM, the hippocampus helps the neocortex reinstate strong versions of attractors (the neocortex's own attractors for new information would not be strong enough to support offline learning on their own), and this autonomous error-driven learning allows useful weight updates from these states. Most prior models of replay, including the original CLS model, do not have this fully autonomous dynamic, with some notable recent exceptions (53–56). Typically, the model architect exerts some control over the input to the model on a given replay trial, and, indeed, defines what a "trial" is. However, a large part of the computational puzzle of understanding learning during sleep is understanding how the system can generate its own learning trials, and our account adds to other recent efforts providing a demonstration of how this can work.

These mechanisms thus act as hypotheses for how sleep carries out autonomous replay: Once the system has visited an attractor, there should be an endogenous mechanism of leaving this attractor in order to transition to the next. We hypothesize that this is accomplished by a form of adaptation at the synaptic level. Unit-level adaptation or homeostatic mechanisms may also be possible (54), although they should result in less ability to transition between closely related memories (which are likely to share units). To define offline learning trials, we hypothesize that the brain contrasts periods of stable activity with periods of perturbed activity, consistent with several accounts of biologically plausible error-driven learning (57–59). We further hypothesize that oscillations (spindles in NREM and theta in REM) enhance neocortical learning during sleep by distorting attractors during the unstable (minus phase) periods in ways that are informative about flaws in the network's representation. This implies that dampening or eliminating these oscillations would reduce the efficacy of learning during sleep, consistent with findings from both NREM and REM sleep (38, 60).

In simulation 1, as in our human experiment results, we found that these autonomous learning dynamics resulted in enhanced memory for features of exemplars shared across a category, while preserving memory for exemplar-unique features. This was not due to shared features appearing more frequently in the stimulus set, as frequency-matched features that appeared across categories did not show this improvement. Sleep resulted in increased similarity for neocortical representations of items in the same category, consistent with the literature suggesting that sleep helps to extract the structure or central tendencies of a domain (45, 61–63). To the extent that detailed information is present in the hippocampus, it should be able to help neocortex learn this detailed information, and there is extensive evidence that sleep, indeed, benefits retention of detailed information (28, 64, 65). However, the overlapping nature of representations in the neocortex (especially relative to the hippocampus) means that the neocortex will tend to emphasize shared structure over the course of systems consolidation.

In simulation 2, we explored the role of alternating NREM and REM sleep stages in enabling successful continual learning. We simulated learning of a new environment, reliant on the hippocampus, after the neocortex had already consolidated an old, related environment. We found that the hippocampus helps build neocortical representations of the new environment during NREM sleep (as in simulation 1), and the neocortex then reinstates the old environment during REM sleep, allowing for integration of the new information into neocortex without overwriting the old. Our proposal is strongly consistent with studies suggesting REM is important for integrating new information with existing knowledge (66) and that this process is set up by prior memory reactivation during slow wave sleep (67, 68). It is also consonant, more generally, with the idea that both NREM and REM sleep are critical for memory integration, with the ordering of the stages being consequential (69, 70).

This simulation demonstrates how a night of sleep can serve to protect against interference even in nonstationary environments. The original CLS framework did not provide an account of this learning scenario, as it required the environment to continue to provide reminders of old knowledge, the role taken over by REM here. The sleep model of Norman et al. (26) showed how REM can provide the necessary reminder and repair function (71), but that model required continued exposure from the environment over time to learn new knowledge, the role taken over by NREM here. Deep neural networks performing both experience replay and generative replay have had impressive success in tackling the problem of continual learning in complex nonstationary environments (72–76). Our approach is related in benefitting from offline interleaved replay, but it

fleshes out the autonomous interactions between the hippocampus and cortex that may support this learning, and explains how this can happen quickly upon transition to a new environment through the alternation of NREM and REM sleep.

Overall, we view the process of sleep-dependent memory consolidation not as a simple strengthening of individual memories or weakening of noise (77) but, instead, as a restructuring that acts to update our internal models of the world to better reflect the environment over time. According to our framework, new information learned over the course of one waking period will be quickly encoded by the hippocampus. We predict that the hippocampus and neocortex will then concurrently replay this information in interleaved order during NREM sleep, with the hippocampus helping the neocortex reinstate higher-fidelity versions of recent experiences than the neocortex could support without hippocampal influence. This hippocampally mediated interleaved replay should support new learning in neocortex that serves to especially emphasize the structure of the new domain, increasing representational overlap between related entities. This prediction has not yet been directly tested, but it is consistent with a finding of increased representational overlap over time for related memories in the medial prefrontal cortex (45). In keeping with recent findings in humans and rodents (78, 79), we also predict that the hippocampus will support memory consolidation in situations where it was not required for initial learning, so long as the hippocampus was encoding the details of the experience in parallel with other regions.

Our account also makes predictions about sleepstage contributions to memory consolidation. First, NREM sleep should more frequently visit attractors corresponding to recent experience, and REM sleep should more frequently visit attractors corresponding to remote, consolidated knowledge, although this is not absolute. This prediction is consistent with the finding that NREM dreams incorporate more recent episodic information in comparison to REM dreams, which incorporate more existing semantic information (80). There have been very few empirical demonstrations of replay during REM (81, 82), and our account suggests that this may be partly because REM is not as focused on recent experience, which is the subject of most experiments. Our account also predicts that integration of new information without disruption to existing knowledge requires NREM–REM alternations. If REM is suppressed, or (more specifically) if replay during REM is suppressed, the account predicts that old information that is related to new information will eventually be overwritten (Fig. 4*C*).

Our model's simulations are intended as demonstrations as opposed to a full account of offline consolidation, but there are exciting possible future directions for this work, as well as complementary modeling frameworks that already help provide a more complete understanding. For example, our current model does not simulate changes in memory across periods of wake. Often, sleep studies find a marked reduction in performance with wake, including in our satellite category learning study (29). Reductions in performance can be simulated through interference caused by waking experience and/or decay in weight strengths over time. However, the story is not so simple, because awake replay can be beneficial for memory (83–86). It may be that awake replay works against other causes of memory degradation, and/or that awake replay does not result in the kind of lasting improvements to performance that sleep replay does; perhaps awake replay leads to short-term, local improvements, but the specialized coordinated hippocampal–cortical replay dynamics that exist during sleep are needed for persistent systems-level change (87).

We have used the word "replay" in the manner it is used in the modeling literature—offline reactivation of any information previously experienced in the environment. Sometimes the word is reserved, in the empirical literature, for the sequential reactivation of states experienced in a particular order. Most of the rodent literature has studied this sequential reactivation, and it will be important for future versions of our model to simulate these kinds of sequential paradigms, taking inspiration from other modeling frameworks that have focused on simulating sequential reactivation (54, 88, 89).

Another limitation of our framework is that, although oscillations are critical to our results, our use of a rate code (where unit activation values correspond to a rate of firing across time) makes the modeling of oscillations more abstract, and limits our ability to explain or make predictions about detailed oscillatory dynamics across the brain. However, a biologically detailed thalamocortical neural network model using a Hebbian learning rule has been developed to engage with these dynamics and provides important insights into how they may contribute to memory consolidation (54–56). We view this model and others that make more direct contact with lower level neural mechanisms (90) as complementary to our approach. Having these models at different levels of abstraction is useful in creating the full bridge from individual neurons to dynamics across systems to higher-level behavioral phenomena—for example, our use of a rate code allows for easier simulation of more complex tasks, and may help our framework scale to larger networks. Whereas prior, more biologically detailed models have focused on the idea of simple direct transfer of information from hippocampus to cortex, our framework provides an account of qualitative changes in the nature of memory over the course of systems consolidation.

Another important missing piece in the current modeling framework, which has been tackled in other frameworks (52, 53, 88, 91, 92), is an explanation of how memories are prioritized for replay, for example, along dimensions of emotion, reward, or future relevance (61, 93). We and others have also found evidence for prioritization of weaker memories for offline processing (29, 42, 49, 71, 94). Prioritization of weaker information is not something that falls naturally out of the current framework; an additional tagging mechanism is needed to mark memories with higher uncertainty or error during or after learning for later replay.

Our current simulations focused on hippocampal influence on neocortical representations during sleep, with hippocampal learning turned off to isolate the effects of cortical learning, but future simulations should also consider sleep-dependent learning locally within the hippocampus (39), including differential learning in the monosynaptic pathway (MSP) and trisynaptic pathway (TSP). Some of the behavioral change seen in the sleep consolidation literature could arise from this local hippocampal learning (22).

The hippocampus module in our framework is a version of our C-HORSE model (31–33), which rapidly learns both new statistical and episodic information, in its MSP and TSP, respectively. The original CLS framework proposed that the hippocampus handles the rapid synaptic changes involved in encoding new episodes, in order to prevent the interference that would occur from attempting to make these synaptic changes directly in neocortex. The neocortex then slowly extracts the statistics across these episodes to build up semantic information over time (weeks, months, or years). But we can learn new semantic and statistical information much more quickly than this (within minutes or hours), and the hippocampus is likely responsible for

much of this learning (31, 33, 79). Our view is thus that the hippocampus acts as a temporary buffer for both statistics and episodes—for any novel information requiring large synaptic changes that might cause interference if implemented directly in neocortex.

The hippocampus in the original CLS account (and also in experience replay models) functions somewhat like the TSP in our hippocampus model, replaying individual episodes and exemplars with high fidelity. The MSP, however, is more sensitive to the structure across experiences (31, 33). This suggests the possibility of generalized replay, which has, indeed, been observed (95, 96), and which may serve to catalyze the consolidation of structured information (97).

We have modeled the hippocampus and one generic neocortical module, but, of course, the brain is much more heterogeneous and hierarchical than this. We could consider the TSP and the neocortex module as two points on a spectrum, with the TSP learning very quickly using orthogonalized representations, and the neocortex learning slowly using overlapping representations. The MSP has intermediate learning speed and overlap in its representations, and it is possible that hierarchically organized neocortical regions form a gradient (98), with slower learning as a region gets closer to the sensory periphery—from hippocampus, to MTL cortex, to high-level sensory regions, to low-level sensory regions. Each region could help train the adjacent slower region through concurrent offline replay. The hippocampus—both the MSP and TSP—may be especially important for tracking information across the time period of 1 day (or across one waking period, for animals sleeping in shorter bouts). Then, with the first bout of sleep, the hippocampus starts to help adjacent neocortical areas build representations of the new information from the prior waking period.

While some learning does occur in the neocortex during initial waking exposure, both in our model and in the brain, it may not typically stabilize or be strong enough to support behavior prior to hippocampal influence during sleep (23, 99, 100). When information is strongly consistent with prior knowledge, there may be the possibility for small tweaks that allow direct integration into neocortical areas without the need for extensive offline restructuring (101, 102). Our account, as well as the original CLS account, predicts that, with enough exposure, any information, even completely novel information, can eventually be learned in the neocortex without hippocampal influence [perhaps explaining the high functioning of developmental amnesiacs (103)]. However, novel information can be learned much faster if it goes through the process of quick encoding in the hippocampus followed by neocortical integration offline.

In summary, the model provides a framework for understanding how the hippocampus can help shape new representations in neocortex, and how alternating NREM/REM sleep cycles across the night continues to allow additions to this knowledge as we encounter new information over time. We hope it will inspire empirical tests and provide a foundation for exploring the mechanisms underlying a range of sleep-dependent memory findings.

Methods

Wake Training. We implemented our simulations in the Emergent neural network framework (104). We departed from the Emergent default learning in using a fully error-driven learning scheme with CHL (41). CHL computes errors locally but behaves similarly to the backpropagation algorithm (59, 105). In each trial, CHL contrasts two states: a "plus" state, in which targets are provided, and a "minus" state, where the model makes a guess without a target. Given a learning rate parameter ε and settled sender (s) and receiver (r) activations at each synapse in the last cycle of the plus (+) and minus (−) phases, each weight update (Δw) is calculated as

$$\Delta w = \varepsilon[(s^+r^+) - (s^-r^-)].$$

In the first simulation, the model was trained on 15 satellite exemplars from three categories (Fig. 2*A*). All results are averages across 100 model initializations. All weights were randomly sampled from uniform distributions with every new initialization (means and ranges are listed in *SI Appendix*).

In simulation 2, the training protocol involved, first, overtraining the model with only the neocortical hidden layer active on Env 1 items to simulate consolidated remote knowledge. We trained the model for 30 epochs after reaching zero error. We then switched to Env 2 training, with both the hippocampus and neocortex engaged in learning; the learning process continued until our error measure reached zero. Weight changes were calculated via CHL (Eq1) based on the coactivation differences between the plus and minus phases in both simulations. All results are averages across 100 model initializations.

Wake Testing. For both simulations, testing involved presenting the input patterns and allowing the model to generate an output using its learned weights. In simulation 1, performance was computed for each of the 78 shared and 27 unique features across the satellite exemplars on feature-held-out testing trials in each testing epoch. The model's performance on each trial was judged as correct if it generated activity on the correct units greater than a threshold of 0.5 and activity on incorrect units less than 0.5. Shared and unique proportion-correct scores were then separately calculated by averaging binary correct/incorrect responses in each feature category for each model initialization. These results were then averaged across random initializations.

In simulation 2, testing involved presenting an item on the input layer and then evaluating whether it generated activity on the correct units in the output layer greater than a threshold of 0.5 and activity on incorrect units less than 0.5. Each testing epoch consisted of testing all 10 Env 1 items and all 10 Env 2 items, and performance on each environment's items was calculated separately by averaging binary correct/incorrect responses across the ten items.

Sleep Activity Dynamics. Sleep epochs were initiated with a random noise injection, after which all external inputs were silenced. Recirculating activity dynamics allowed the model to autonomously reinstate learned item attractors. Each synapse in the model was subject to short-term synaptic depression on its weight as a function of coactivation-induced calcium accumulation. Given Ca_{inc}, Ca_{dec} time constant parameters, s sender activation, r receiver activation, and wt synaptic weight, the calcium update on each processing cycle was computed as

$$\Delta Ca = Ca_{inc}(1 - Ca)\Big((r \cdot s \cdot wt) - Ca_{dec}\Big)Ca.$$

Given a calcium-based depression threshold parameter Ca_{thr}, and gain parameter SD_{gain}, synaptic depression was computed on the weights as follows:

If $Ca > Ca_{thr}$,

$$Wt = Wt \cdot \left(1 - \left(\left(SD_{gain}\,(1 - Ca_{thr})^{-1}\right)(Ca - Ca_{thr})\right)\right)^2.$$

Inhibitory oscillations were implemented via a parameterized sinusoidal wave. Given the default layer FFFB (Feedforward Feedback) inhibitory conductance Gi parameter Gi_{def}, amplitude A, period P, a midline shift S and processing cycle c, the FFFB Gi parameter Gi_c for a given cycle for each layer was set as

$$Gi_c = Gi_{def} * A \cdot sin\Big(2\pi \cdot (P \cdot c)^{-1}\Big) + S.$$

Sleep Learning. Sleep learning events were defined by the stability of model activity, calculated as the average temporal autocorrelation of layer activity in the model. Given n model layers' activity L indexed by i, processing cycle c and the Pearson's correlation function $Corr$, stability was computed as

$$Stability = \sum_{i=0}^{n} Corr(L_i(c), (L_i(c-1))n^{-1}.$$

Plus phases were marked as contiguous cycles where stability was greater than a strict plus threshold (0.999965 and 0.9999 for simulations 1 and 2, respectively). Minus phases were marked by the periods following plus phases where stability was greater than the minus threshold (0.997465 and 0.9899 for simulations 1 and 2, respectively). These thresholds were set in order to obtain

desirable plus and minus phase attractor dynamics in the two simulations. In a scaled-up network exposed to many domains, it may be more feasible to keep these parameters stable across tasks, a possibility that could be explored in future work. Weights were updated at the end of every minus phase. Weight changes were computed using a modified version of the CHL learning rule used in wake training (Eq1). Changes were based on the difference in average minus and plus phase coactivations at each synapse,

$$\Delta w = \varepsilon[\overline{(s^+r^+)} - \overline{(s^-r^-)}].$$

This average allows coactivity dynamics throughout the minus phase (both high and low phases of inhibition) to contribute to the contrast with the plus phase.

Sleep Stages and Simulation Protocols. NREM in both simulations involved all C-HORSE layers and the neocortical hidden layer being active, allowing unimpeded communication between the hippocampus and neocortex. We set up the simulations with the possibility for bidirectional influence, given evidence that the neocortex influences hippocampal replay (106), but, given the stronger learning of new information in the hippocampus, we expect the effective influence to run mostly from hippocampus to neocortex (107). Indeed, we found very similar results when limiting the flow of activity from hippocampus to neocortex during NREM. For REM sleep, C-HORSE and the neocortex were disconnected (lesioned connection from entorhinal cortex to hippocampus), implementing the idea that neocortex dynamics are less influenced by the hippocampus during REM. As we were focused on neocortical learning, C-HORSE projections in both sleep stages were nonlearning. Only input/output layer <−> neocortical layer projections were updated during sleep.

In simulation 1, after the wake training criterion was achieved, the model executed one 30,000-cycle epoch of NREM sleep. Test epochs before and after sleep established performance change over sleep. In the second simulation, after neocortex-only Env 1 item training and full model Env 2 item training, the model either performed five alternating 10,000-cycle epochs of NREM and REM sleep or five 10,000-cycle epochs of consecutive NREM. Test epochs established performance after every sleep epoch. We did some parameter searching to find the approximately best performing length of the sleep blocks. We found that, if blocks were too short, there was not enough replay for there to be substantial performance changes during sleep. If the blocks were too long in simulation 1 (for example, 100,000 cycles), we found that–while we replicated the shared benefit–the model's performance on unique features deteriorated, suggesting eventual deterioration of replay dynamics. In simulation 2, if the sleep blocks were too long (for example, 50,000 cycles), NREM sleep caused so much damage to cortical Env 1 representations that REM sleep was not able to facilitate their repair.

In the simulation 2 alternating condition, we ran 10 total sleep blocks with 5 alternating blocks of NREM and REM sleep each in order to match the number of alternations in an average night of adult human sleep (108). To explore whether the protection to Env 1 performance conferred by the alternating schedule would continue with additional sleep blocks, we ran simulations in which we allowed the model to sleep for either 15 alternating blocks of NREM and REM each (30 total blocks) or 15 sequential NREM blocks. We found that, while there was a reduction in Env 1 performance over the sleep blocks in both the alternating NREM/REM and NREM-only conditions, there was a continued substantial benefit conferred by the alternating NREM/REM condition to Env 1 performance.

Data, Materials, and Software Availability. Data deposition: Model code has been deposited in GitHub (https://github.com/schapirolab/SinghNormanSchapiro_PNAS22) (109).

ACKNOWLEDGMENTS. We are grateful to James Antony, Elizabeth McDevitt, and Sharon Thompson-Schill for helpful discussions, and to Randall O'Reilly, Seth Herd, and Diheng Zhang for model implementation support. This work was supported by NIH grant R01 MH069456 to K.A.N. and Charles E. Kaufman Foundation grant KA2020-114800 to A.C.S.

1. J. L. McClelland, B. L. McNaughton, R. C. O'Reilly, 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).
2. S. Diekelmann, J. Born, The memory function of sleep. *Nat. Rev. Neurosci.* **11**, 114–126 (2010).
3. B. Rasch, J. Born, About sleep's role in memory. *Physiol. Rev.* **93**, 681–766 (2013).
4. J. G. Klinzing, N. Niethard, J. Born, Mechanisms of systems memory consolidation during sleep. *Nat. Neurosci.* **22**, 1598–1610 (2019).
5. M. Geva-Sagiv, Y. Nir, Local sleep oscillations: implications for memory consolidation. *Front. Neurosci.* **13**, 813 (2019).
6. J. Tamminen, J. D. Payne, R. Stickgold, E. J. Wamsley, M. G. Gaskell, Sleep spindle activity is associated with the integration of new memories and existing knowledge. *J. Neurosci.* **30**, 14356–14360 (2010).
7. J. Tamminen, M. A. Lambon Ralph, P. A. Lewis, The role of sleep spindles and slow-wave activity in integrating new information in semantic memory. *J. Neurosci.* **33**, 15376–15381 (2013).
8. F. Xia *et al.*, Parvalbumin-positive interneurons mediate neocortical-hippocampal interactions that are necessary for memory consolidation. *eLife* **6**, e27868 (2017).
9. D. Khodagholy, J. N. Gelinas, G. Buzsáki, Learning-enhanced coupling between ripple oscillations in association cortices and hippocampus. *Science* **358**, 369–372 (2017).
10. N. K. Logothetis *et al.*, Hippocampal-cortical interaction during periods of subcortical silence. *Nature* **491**, 547–553 (2012).
11. D. Ji, M. A. Wilson, Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci.* **10**, 100–107 (2007).
12. H.-V. V. Ngo, T. Martinetz, J. Born, M. Mölle, Auditory closed-loop stimulation of the sleep slow oscillation enhances memory. *Neuron* **78**, 545–553 (2013).
13. N. Maingret, G. Girardeau, R. Todorova, M. Goutierre, M. Zugaro, Hippocampo-cortical coupling mediates memory consolidation during sleep. *Nat. Neurosci.* **19**, 959–964 (2016).
14. J. Ladenbauer *et al.*, Promoting sleep oscillations and their functional coupling by transcranial stimulation enhances memory consolidation in mild cognitive impairment. *J. Neurosci.* **37**, 7111–7124 (2017).
15. M. Geva-Sagiv *et al.*, Hippocampal-prefrontal neuronal synchrony during human sleep mediates memory consolidation (Presented at Society for Neuroscience, 2021) [Poster]. https://www.abstractsonline.com/pp8/#!/10485/presentation/10246. Accessed 1 December 2021.
16. N. Ketz, A. P. Jones, N. B. Bryant, V. P. Clark, P. K. Pilly, Closed-loop slow-wave tACS improves sleep-dependent long-term memory generalization by modulating endogenous oscillations. *J. Neurosci.* **38**, 7314–7326 (2018).
17. L. R. Squire, P. Alvarez, Retrograde amnesia and memory consolidation: A neurobiological perspective. *Curr. Opin. Neurobiol.* **5**, 169–177 (1995).
18. G. Winocur, M. Moscovitch, M. Sekeres, Memory consolidation or transformation: Context manipulation and hippocampal representations of memory. *Nat. Neurosci.* **10**, 555–557 (2007).
19. L. Nadel, M. Moscovitch, Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* **7**, 217–227 (1997).
20. G. Winocur, M. Moscovitch, Memory transformation and systems consolidation. *J. Int. Neuropsychol. Soc.* **17**, 766–780 (2011).
21. M. Moscovitch, A. Gilboa, Systems consolidation, transformation and reorganization: Multiple trace theory, trace transformation theory and their competitors. PsyArXiv [Preprint] (2021). https://doi.org/10.31234/osf.io/yxbrs. Accessed 4 December 2021.
22. A. P. Yonelinas, C. Ranganath, A. D. Ekstrom, B. J. Wiltgen, A contextual binding theory of episodic memory: Systems consolidation reconsidered. *Nat. Rev. Neurosci.* **20**, 364–375 (2019).
23. T. Kitamura *et al.*, Engrams and circuits crucial for systems consolidation of a memory. *Science* **356**, 73–78 (2017).
24. J. W. Antony, A. C. Schapiro, Active and effective replay: Systems consolidation reconsidered again. *Nat. Rev. Neurosci.* **20**, 506–507 (2019).
25. G. Kowadlo, A. Ahmed, D. Rawlinson, One-shot learning for the long term: Consolidation with an artificial hippocampal algorithm. ArXiv [Preprint] (2021). https://doi.org/10.48550/arXiv.2102.07503. Accessed 10 May 2022.
26. K. A. Norman, E. L. Newman, A. J. Perotte, Methods for reducing interference in the Complementary Learning Systems model: Oscillating inhibition and autonomous memory rehearsal. *Neural Netw.* **18**, 1212–1228 (2005).
27. A. Goto *et al.*, Stepwise synaptic plasticity events drive the early phase of memory consolidation. *Science* **374**, 857–863 (2021).
28. M. Friedrich, I. Wilhelm, J. Born, A. D. Friederici, Generalization of word meanings during infant sleep. *Nat. Commun.* **6**, 6004 (2015).
29. A. C. Schapiro *et al.*, Sleep benefits memory for semantic category structure while preserving exemplar-specific information. *Sci. Rep.* **7**, 14869 (2017).
30. M. Huguet, J. D. Payne, S. Y. Kim, S. E. Alger, Overnight sleep benefits both neutral and negative direct associative and relational memory. *Cogn. Affect. Behav. Neurosci.* **19**, 1391–1403 (2019).
31. A. C. Schapiro, N. B. Turk-Browne, M. M. Botvinick, K. A. Norman, 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 (2017).
32. Z. Zhou, D. Singh, M. C. Tandoc, A. C. Schapiro, Distributed representations for human inference. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.07.29.454337. Accessed 1 December 2021.
33. J. Sučević, A. C. Schapiro, A neural network model of hippocampal contributions to category learning. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.01.12.476051. Accessed 18 January 2022.
34. R. C. O'Reilly, Y. Munakata, M. J. Frank, T. E. Hazy, Computational cognitive neuroscience, ed. 4. https://CompCogNeuro.org. Accessed 1 December 2021.
35. N. Niethard, H.-V. V. Ngo, I. Ehrlich, J. Born, Cortical circuit activity underlying sleep slow oscillations and spindles. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E9220–E9229 (2018).
36. J. Durkin *et al.*, Cortically coordinated NREM thalamocortical oscillations play an essential, instructive role in visual system plasticity. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 10485–10490 (2017).
37. A. Peyrache, J. Seibt, A mechanism for learning with sleep spindles. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190230 (2020).
38. R. Boyce, S. D. Glasgow, S. Williams, A. Adamantidis, Causal evidence for the role of REM sleep theta rhythm in contextual memory consolidation. *Science* **352**, 812–816 (2016).
39. G. R. Poe, D. A. Nitz, B. L. McNaughton, C. A. Barnes, Experience-dependent phase-reversal of hippocampal neuron firing during REM sleep. *Brain Res.* **855**, 176–180 (2000).

40. A. Treves, Frontal latching networks: A possible neural basis for infinite recursion. *Cogn. Neuropsychol.* **22**, 276–291 (2005).
41. D. H. Ackley, G. E. Hinton, T. J. Sejnowski, A learning algorithm for Boltzmann machines. *Cogn. Sci.* **9**, 147–169 (1985).
42. A. C. Schapiro, E. A. McDevitt, T. T. Rogers, S. C. Mednick, K. A. Norman, Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nat. Commun.* **9**, 3920 (2018).
43. M. A. L. Ralph, E. Jefferies, K. Patterson, T. T. Rogers, The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* **18**, 42–55 (2017).
44. M. T. R. van Kesteren, D. J. Ruiter, G. Fernández, R. N. Henson, How schema and novelty augment memory formation. *Trends Neurosci.* **35**, 211–219 (2012).
45. A. Tompary, L. Davachi, Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron* **96**, 228–241.e5 (2017).
46. L. Sun, H. Zhou, J. Cichon, G. Yang, Experience and sleep-dependent synaptic plasticity: From structure to activity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190234 (2020).
47. M. McCloskey, N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem" in *Psychology of Learning and Motivation*, G. H. Bower, Ed. (Elsevier, 1989), **vol. 24**, pp. 109–165.
48. J. M. Barnes, B. J. Underwood, Fate of first-list associations in transfer theory. *J. Exp. Psychol.* **58**, 97–105 (1959).
49. S. Drosopoulos, C. Schulze, S. Fischer, J. Born, Sleep's function in the spontaneous recovery and consolidation of memories. *J. Exp. Psychol. Gen.* **136**, 169–183 (2007).
50. J. M. Ellenbogen, J. C. Hulbert, R. Stickgold, D. F. Dinges, S. L. Thompson-Schill, Interfering with theories of sleep and memory: Sleep, declarative memory, and associative interference. *Curr. Biol.* **16**, 1290–1294 (2006).
51. B. Baran, J. Wilson, R. M. C. Spencer, REM-dependent repair of competitive memory suppression. *Exp. Brain Res.* **203**, 471–477 (2010).
52. J. L. McClelland, B. L. McNaughton, A. K. Lampinen, Integration of new information in memory: New insights from a complementary learning systems perspective. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190637 (2020).
53. T. L. Hayes *et al.*, Replay in deep learning: Current approaches and missing biological elements. ArXiv [Preprint] (2021). 10.48550/arXiv.2104.04132. Accessed 28 January 2022.
54. O. C. González, Y. Sokolov, G. P. Krishnan, J. E. Delanois, M. Bazhenov, Can sleep protect memories from catastrophic forgetting? *eLife* **9**, e51005 (2020).
55. P. Sanda *et al.*, Bidirectional interaction of hippocampal ripples and cortical slow waves leads to coordinated spiking activity during NREM sleep. *Cereb. Cortex* **31**, 324–340 (2021).
56. Y. Wei, G. P. Krishnan, M. Komarov, M. Bazhenov, Differential roles of sleep spindles and sleep slow oscillations in memory consolidation. *PLOS Comput. Biol.* **14**, e1006322 (2018).
57. M. Welling, G. E. Hinton, "A new learning algorithm for mean field Boltzmann machines" in *Artificial Neural Networks – ICANN 2002*, R. Dorronsoro, Ed. (Lecture Notes in Computer Science, Springer, 2002), **vol. J**, pp. 351–357.
58. B. Scellier, Y. Bengio, Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Front. Comput. Neurosci.* **11**, 24 (2017).
59. R. C. O'Reilly, Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Comput.* **8**, 895–938 (1996).
60. G. Girardeau, K. Benchenane, S. I. Wiener, G. Buzsáki, M. B. Zugaro, Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.* **12**, 1222–1223 (2009).
61. R. Stickgold, M. P. Walker, Sleep-dependent memory triage: Evolving generalization through selective processing. *Nat. Neurosci.* **16**, 139–145 (2013).
62. N. Landmann *et al.*, The reorganisation of memory during sleep. *Sleep Med. Rev.* **18**, 531–541 (2014).
63. P. A. Lewis, S. J. Durrant, Overlapping memory replay during sleep builds cognitive schemata. *Trends Cogn. Sci.* **15**, 343–351 (2011).
64. J. Mirković, M. G. Gaskell, Does sleep improve your grammar? Preferential consolidation of arbitrary components of new linguistic knowledge. *PLoS One* **11**, e0152489 (2016).
65. A. Hanert, F. D. Weber, A. Pedersen, J. Born, T. Bartsch, Sleep in humans stabilizes pattern separation performance. *J. Neurosci.* **37**, 12238–12246 (2017).
66. D. J. Cai, S. A. Mednick, E. M. Harrison, J. C. Kanady, S. C. Mednick, REM, not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy of Sciences* **106**, 10130–10134 (2009).
67. L. J. Batterink, C. E. Westerberg, K. A. Paller, Vocabulary learning benefits from REM after slow-wave sleep. *Neurobiol. Learn. Mem.* **144**, 102–113 (2017).
68. J. Tamminen, M. A. Lambon Ralph, P. A. Lewis, Targeted memory reactivation of newly learned words during sleep triggers REM-mediated integration of new memories and existing knowledge. *Neurobiol. Learn. Mem.* **137**, 77–82 (2017).
69. A. Giuditta *et al.*, The sequential hypothesis of the function of sleep. *Behav. Brain Res.* **69**, 157–166 (1995).
70. M. Strauss *et al.*, Order matters: Sleep spindles contribute to memory consolidation only when followed by rapid-eye-movement sleep. *Sleep* **45**, zsac022 (2022).
71. E. A. McDevitt, K. A. Duggan, S. C. Mednick, REM sleep rescues learning from interference. *Neurobiol. Learn. Mem.* **122**, 51–62 (2015).
72. Z. Li, D. Hoiem, Learning without forgetting. arXiv [Preprint] (2016). https://arxiv.org/abs/1606.09282. Accessed 28 January 2022.
73. H. Shin, J. K. Lee, J. Kim, J. Kim, Continual learning with deep generative replay. arXiv [Preprint] (2017). https://arxiv.org/abs/1705.08690. Accessed 30 January 2022.
74. G. M. van de Ven, H. T. Siegelmann, A. S. Tolias, Brain-inspired replay for continual learning with artificial neural networks. *Nat. Commun.* **11**, 4069 (2020).
75. S.-A. Rebuffi, A. Kolesnikov, G. Sperl, C. H. Lampert, iCaRL: Incremental classifier and representation learning. arXiv [Preprint] (2017). https://arxiv.org/abs/1611.07725. Accessed 30 January 2022.
76. A. Gepperth, C. Karaoguz, A bio-inspired incremental learning architecture for applied perceptual problems. *Cognit. Comput.* **8**, 924–934 (2016).
77. G. Tononi, C. Cirelli, Sleep and the price of plasticity: From synaptic and cellular homeostasis to memory consolidation and integration. *Neuron* **81**, 12–34 (2014).
78. A. Sawangjit *et al.*, The hippocampus is crucial for forming non-hippocampal long-term memory during sleep. *Nature* **564**, 109–113 (2018).
79. A. C. Schapiro *et al.*, The hippocampus is necessary for the consolidation of a task that does not require the hippocampus for initial learning. *Hippocampus* **29**, 1091–1100 (2019).
80. E. J. Wamsley, R. Stickgold, Memory, sleep and dreaming: Experiencing consolidation. *Sleep Med. Clin.* **6**, 97–108 (2011).
81. K. Louie, M. A. Wilson, Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron* **29**, 145–156 (2001).
82. M. Schönauer *et al.*, Decoding material-specific memory reprocessing during sleep in humans. *Nat. Commun.* **8**, 15404 (2017).
83. A. Tambini, L. Davachi, Persistence of hippocampal multivoxel patterns into postencoding rest is related to memory. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19591–19596 (2013).
84. A. Tambini, L. Davachi, Awake reactivation of prior experiences consolidates memories and biases cognition. *Trends Cogn. Sci.* **23**, 876–890 (2019).
85. A. C. Schapiro, T. T. Rogers, N. I. Cordova, N. B. Turk-Browne, M. M. Botvinick, Neural representations of events arise from temporal community structure. *Nat. Neurosci.* **16**, 486–492 (2013).
86. S. P. Jadhav, C. Kemere, P. W. German, L. M. Frank, Awake hippocampal sharp-wave ripples support spatial memory. *Science* **336**, 1454–1458 (2012).
87. L. R. Squire, L. Genzel, J. T. Wixted, R. G. Morris, Memory consolidation. *Cold Spring Harb. Perspect. Biol.* **7**, a021766 (2015).
88. M. G. Mattar, N. D. Daw, Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* **21**, 1609–1617 (2018).
89. X. Liu, D. Kuzum, Hippocampal-cortical memory trace transfer and reactivation through cell-specific stimulus and spontaneous background noise. *Front. Comput. Neurosci.* **13**, 67 (2019).
90. D. F. Tomé, S. Sadeh, C. Clopath, Coordinated hippocampal-thalamic-cortical communication crucial for engram dynamics underneath systems consolidation. *Nat. Commun.* **13**, 840 (2022).
91. W. Sun, M. Advani, N. Spruston, A. Saxe, J. E. Fitzgerald, Organizing memories for generalization in complementary learning systems. bioRxiv [Preprint] (2021). https://www.biorxiv.org/content/10.1101/2021.10.13.463791v1. Accessed 29 January 2022.
92. A. Santoro, P. W. Frankland, B. A. Richards, Memory transformation enhances reinforcement learning in dynamic environments. *J. Neurosci.* **36**, 12228–12242 (2016).
93. E. T. Cowan, A. C. Schapiro, J. E. Dunsmoor, V. P. Murty, Memory consolidation as an adaptive process. *Psychon. Bull. Rev.* **28**, 1796–1810 (2021).
94. D. Denis *et al.*, The roles of item exposure and visualization success in the consolidation of memories across wake and sleep. *Learn. Mem.* **27**, 451–456 (2020).
95. A. S. Gupta, M. A. A. van der Meer, D. S. Touretzky, A. D. Redish, Hippocampal replay is not a simple function of experience. *Neuron* **65**, 695–705 (2010).
96. Y. Liu, R. J. Dolan, Z. Kurth-Nelson, T. E. J. Behrens, Human replay spontaneously reorganizes experience. *Cell* **178**, 640–652.e14 (2019).
97. D. Kumaran, J. L. McClelland, Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychol. Rev.* **119**, 573–616 (2012).
98. A. Roxin, S. Fusi, Efficient partitioning of memory systems and its importance for memory consolidation. *PLOS Comput. Biol.* **9**, e1003146 (2013).
99. L. Himmer, M. Schönauer, D. P. J. Heib, M. Schabus, S. Gais, Rehearsal initiates systems memory consolidation, sleep makes it last. *Sci. Adv.* **5**, eaav1695 (2019).
100. E. Lesburguères *et al.*, Early tagging of cortical networks is required for the formation of enduring associative memory. *Science* **331**, 924–928 (2011).
101. D. Tse *et al.*, Schemas and memory consolidation. *Science* **316**, 76–82 (2007).
102. J. L. McClelland, Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *J. Exp. Psychol. Gen.* **142**, 1190–1210 (2013).
103. F. Vargha-Khadem *et al.*, Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277**, 376–380 (1997).
104. B. Aisa, B. Mingus, R. O'Reilly, The emergent neural modeling system. *Neural Netw.* **21**, 1146–1152 (2008).
105. X. Xie, H. S. Seung, Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural Comput.* **15**, 441–454 (2003).
106. G. Rothschild, E. Eban, L. M. Frank, A cortical-hippocampal-cortical loop of information processing during memory consolidation. *Nat. Neurosci.* **20**, 251–259 (2017).
107. M. E. Hasselmo, Neuromodulation: Acetylcholine and memory consolidation. *Trends Cogn. Sci.* **3**, 351–359 (1999).
108. E. Hartmann, The 90-minute sleep-dream cycle. *Arch. Gen. Psychiatry* **18**, 280–286 (1968).
109. D. Singh, K. A. Norman, A. C. Schapiro, A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation. GitHub. https://github.com/schapirolab/SinghNormanSchapiro_PNAS22. Deposited 30 August 2022.