Feature Review

What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated

Dharshan Kumaran,[1,2,*] Demis Hassabis,[1,3,*] and James L. McClelland[4,*]

We update complementary learning systems (CLS) theory, which holds that intelligent agents must possess two learning systems, instantiated in mammalians in neocortex and hippocampus. The first gradually acquires structured knowledge representations while the second quickly learns the specifics of individual experiences. We broaden the role of replay of hippocampal memories in the theory, noting that replay allows goal-dependent weighting of experience statistics. We also address recent challenges to the theory and extend it by showing that recurrent activation of hippocampal traces can support some forms of generalization and that neocortical learning can be rapid for information that is consistent with known structure. Finally, we note the relevance of the theory to the design of artificial intelligent agents, highlighting connections between neuroscience and machine learning.

Trends

Discovery of structure in ensembles of experiences depends on an interleaved learning process both in biological neural networks in neocortex and in contemporary artificial neural networks.

Recent work shows that once structured knowledge has been acquired in such networks, new consistent information can be integrated rapidly.

Both natural and artificial learning systems benefit from a second system that stores specific experiences, centred on the hippocampus in mammalians.

Replay of experiences from this system supports interleaved learning and can be modulated by reward or novelty, which acts to rebalance the general statistics of the environment towards the goals of the agent.

Recurrent activation of multiple memories within an instance-based system can be used to discover links between experiences, supporting generalization and memory-based reasoning.

Complementary Learning Systems

Twenty years have passed since the introduction of the CLS theory of human learning and memory [1], a theory that, itself, had roots in earlier ideas of Marr and others. According to the theory, effective learning requires two complementary systems: one, located in the neocortex, serves as the basis for the gradual acquisition of structured knowledge about the environment, while the other, centered on the hippocampus, allows rapid learning of the specifics of individual items and experiences. We begin with a review of the core tenets of this theory. We then provide three types of updates. First, we extend the role of replay of memories stored in the hippocampus. This mechanism, initially proposed to support the integration of new information into the neocortex, may support a diverse set of functions [2,3], including goal-related manipulation of experience statistics such that the neocortex is not a slave to the statistics of its environment. Second, we describe recent updates to the theory in response to two key empirical challenges: (i) evidence suggesting that the hippocampus supports some forms of generalization that go beyond those originally envisaged [4–6], and (ii) evidence suggesting that, when new information is consistent with existing knowledge, the time required for its integration into the neocortex may be much shorter than originally suggested [7,8]. In a final section, we highlight links between the core principles of CLS theory and recent themes in machine learning, including neural network architectures that incorporate memory modules that have parallels with the hippocampus. While there remain several issues not yet fully addressed (see Outstanding Questions), the extensions, responses to challenges, and integration with machine learning bring the theory into agreement with many important recent developments and provide a take-off point for future investigation.

[1]Google DeepMind, 5 New Street Square, London EC4A 3TW, UK
[2]Institute of Cognitive Neuroscience, University College London, 17 Queen Square, WC1N 3AR, UK
[3]Gatsby Computational Neuroscience Unit, 17 Queen Square, London WC1N 3AR, UK
[4]Department of Psychology and Center for Mind, Brain, and Computation, Stanford University, 450 Serra Mall, CA 94305, USA

 http://dx.doi.org/10.1016/j.tics.2016.05.004

*Correspondence:
dkumaran@google.com (D. Kumaran),
demishassabis@google.com
(D. Hassabis), mcclelland@stanford.edu
(J.L. McClelland).

Summary of the Theory

CLS theory [1] provided a framework within which to characterize the organization of learning in the brain (Figure 1, Key Figure). Drawing on earlier ideas by David Marr [9], it offered a synthesis of the computational functions and characteristics of the hippocampus and neocortex that not only accounted for a wealth of empirical data (Box 1) but resonated with rational perspectives on the challenges faced by intelligent agents.

Structured Knowledge Representation System in Neocortex

A central tenet of the theory is that the neocortex houses a structured knowledge representation, stored in the connections among the neurons in the neocortex. This tenet arose from the observation that multi-layered neural networks (Figure 2) gradually learn to extract structure when trained by adjusting connection weights to minimize error in the network outputs [10]. Early

Key Figure

Complementary Learning Systems (CLS) and their Interactions.

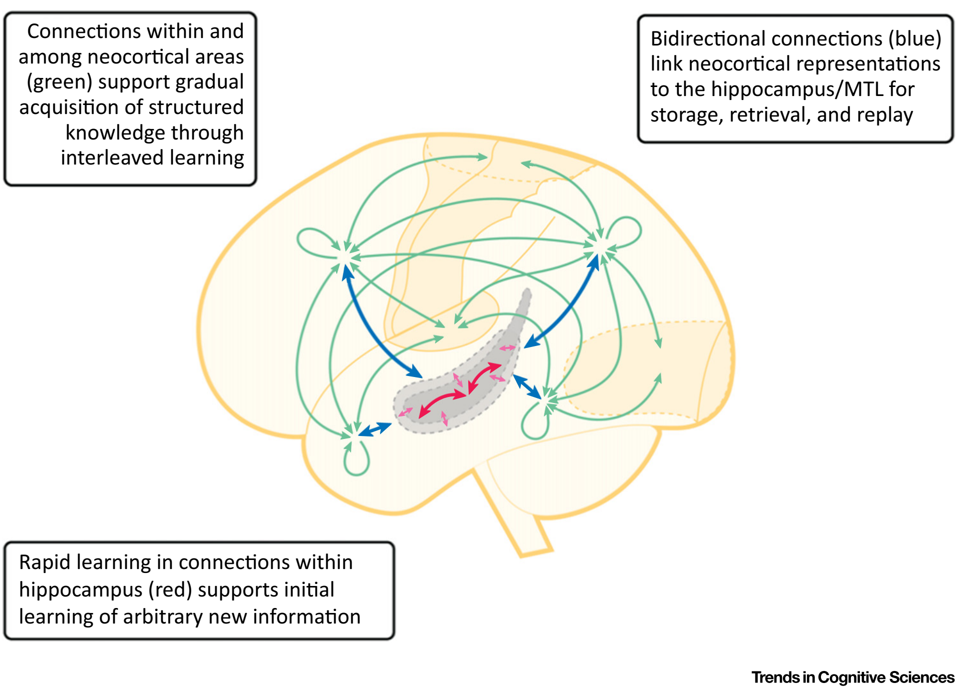


Figure 1. Lateral view of one hemisphere of the brain, where broken lines indicate regions deep inside the brain or on the medial surface. Primary sensory and motor cortices are shown in darker yellow. Medial temporal lobe (MTL) surrounded by broken lines, with hippocampus in dark grey and surrounding MTL cortices in light grey (size and location are approximate). Green arrows represent bidirectional connections within and between integrative neocortical association areas, and between these areas and modality specific areas (the integrative areas and their connections are more dispersed than the figure suggests). Blue arrows denote bidirectional connections between neocortical areas and the MTL. Both blue and green connections are part of the structure-sensitive neocortical learning system in the CLS theory. Red arrows within the MTL denote connections within the hippocampus, and lighter-red arrows indicate connections between the hippocampus and surrounding MTL cortices: these connections exhibit rapid synaptic plasticity (red greater than light-red arrows) crucial for the rapid binding of the elements of an event into an integrated hippocampal representation. Systems-level consolidation involves hippocampal activity during replay spreading to neocortical association areas via pathways indicated with blue arrows, thereby supporting learning within intra-neocortical connections (green arrows). Systems-level consolidation is considered complete when memory retrieval – reactivation of the relevant set of neocortical representations – can occur without the hippocampus.

Box 1. Empirical Evidence Supporting Core Principles of CLS Theory

The Role of the Hippocampus in Memory

Bilateral damage to the hippocampus profoundly affects memory for new information, leaving language, reading, general knowledge, and acquired cognitive skills intact [29,34], consistent with the idea that many types of new learning are initially hippocampus-dependent. Memory for recent pre-morbid information is profoundly affected by hippocampal damage, with older memories being less dependent on the hippocampus and therefore less sensitive to hippocampal lesions [1,34,51,128], supporting gradual integration of learned information into cortical knowledge structures. However, some evidence suggests that memory for specific details of an event can remain MTL-dependent [52,129] as long as the details are retained (e.g., [130]).

Hippocampus Supports Core Computations and Representations of a Fast-Learning Episodic Memory System

Episodic memory is widely accepted to depend on the hippocampus, mediated by a capacity to bind together (i.e., 'auto-associate') diverse inputs from different brain areas that represent the constituents of an event. Indeed, information about the spatial (e.g., place) and non-spatial (e.g., what happened) aspects of an event are thought to be processed primarily by parallel streams before converging in the hippocampus at the level of the DG/CA3 subregions [37]. Two complementary computations – pattern separation and pattern completion – are viewed to be central to the function of the hippocampus for storing details of specific experiences. Evidence suggests that the dentate gyrus (DG) subregion of the hippocampus performs pattern separation, orthogonalizing incoming inputs before **auto-associative storage** in the CA3 region [131–137]. Further, the CA3 subregion is crucial for pattern completion – allowing the output of an entire stored pattern (e.g., corresponding to an entire episodic memory) from a partial input consistent with its function as an attractor network [138,139] (Boxes 2–4).

Hippocampal Replay

A wealth of evidence demonstrates that replay of recent experiences occurs during offline periods (e.g., during sleep, rest) [2,3]. Further, the hippocampus and neocortex interact during replay as predicted by CLS theory [65], putatively to support interleaved learning. A causal role for replay in systems-level consolidation is supported by the finding that optogenetic blockage of CA3 output in transgenic mouse after learning in a contextual fear paradigm specifically reduces sharp-wave ripple (SWR) complexes in CA1 and impairs consolidation [69].

The Hippocampus And Neocortex Support Qualitatively Different Forms of Representation

A recent experiment [140] found initial evidence in favor: the behavior of rats in the Morris water maze early on appeared to reflect individual episodic traces (i.e., an instance-based non-parametric representation), but at a later time-point (28 days after learning) was consistent with the use of a parametric representation putatively housed in the neocortex.

examples were provided by networks that learned to read words aloud [11–13] from repeated, interleaved exposure to the spellings and corresponding sounds of English words. These networks supported the gradual acquisition of a structured knowledge representation in the connection weights among the units in the network, shaped by the statistics of the environment in a fashion that was efficient and generalized to novel examples [1,14,15], while also supporting performance on atypical items occurring frequently in the domain. Such a representation can be described as **parametric** (see Glossary) rather than as item-based (or **non-parametric**) in that the connection weights can be viewed as a set of parameters optimized for the entire domain (e.g., the spellings and sounds of the full set of words in the language) instead of supporting memory for the items *per se*. According to the theory, such networks underlie acquired cognitive abilities of all types in domains as diverse as perception, language, semantic knowledge representation, and skilled action. This idea can be seen as an extension of Marr's original proposal [9], which held that cortical neurons each learned the statistics associated with a particular category.

The CLS theory proposed that learning in such a parametric system will necessarily be slow, for two main reasons: first, each experience represents a single sample from the environment. Given this, a small learning rate allows a more-accurate estimate of the underlying population statistics by effectively aggregating information over a larger number of samples [1]. Second, the optimal adjustment of each connection depends on the values of all of the other connections. Before the ensemble of connections has been structured by experience, the signals specifying how to change connection weights to optimize the representation will be both noisy and weak, slowing initial learning. This issue has proved to be particularly important in deep (i.e., many-layered) neural network architectures that have enjoyed recent successes in machine learning [16] as well as in

Glossary

Attractor network: networks with recurrent connectivity that have stable states which persist in the absence of external inputs, and afford noise tolerance. Discrete/point attractor networks can be used to store multiple memories as individual stable states. Continuous attractor networks have a continuous manifold of stable points which allow them to represent continuous variables (e.g., position in space).

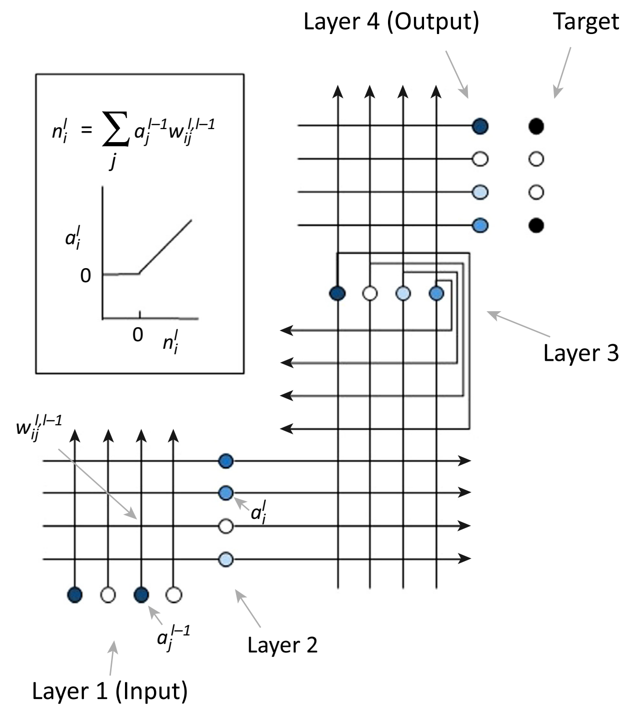
Auto-associative storage: the storage within an attractor network of an input pattern constituting an experience, such that elements of the input pattern are linked together through plasticity within the recurrent connections of the network. The operation of recurrent connections supports functions such as pattern completion, whereby the entire input pattern (e.g., memory of a birthday party) can be retrieved from a partial cue (e.g., the face of a friend).

Exemplar models: exemplar models in cognitive science, related to instance-based models in machine learning, operate by computing the similarity of a new input pattern (i.e., presented as external sensory input) to stored experiences. This results in the output of the model, for example a predicted category label for the new input pattern, at which point the process terminates.

Non-parametric: we use this term to refer to algorithms where each experience or datapoint has its own set of coordinates, where capacity can be increased as required – and the number of parameters may grow with the amount of data. K-nearest neighbor constitutes one common example of such a non-parametric instance-based method.

Parametric: we use this term to refer to algorithms that do not store each datapoint, but instead directly learn a function that (for example) predicts the output value for a given input. The number of parameters is typically fixed.

Paired associative inference (PAI) task: a paradigm in which items are organized into (e.g., a hundred) sets of triplets (e.g., ABC) or larger sets (e.g., sextets: ABCDEF). Participants view item pairs (e.g., AB, BC) during the study phase and are tested on their ability to appreciate the indirect relationships between items that



Trends in Cognitive Sciences

Figure 2. A Neocortex-Like Artificial Neural Network. In the complementary learning systems (CLS) theory, neocortical processing is seen as occurring through the propagation of activation among neurons via weighted connections, as simulated using artificial networks of neuron-like units (small circles). Each unit has an input line and an output line (with arrowhead). There is a separate real-valued weight where each output line crosses an input line. The weights are the knowledge that governs processing in the network. During processing (inset), each unit computes a net input (n) from the activations of its inputs and the weights (plus a bias term, omitted here), producing an activation (a) that is a non-linear function of n (one such function shown). The units in a layer may project back onto their own inputs (illustrated for layer 3), simulating recurrent intra-cortical computations and higher layers may project back to lower layers (Figure 1). In the situation shown, the input (lower left) is a pattern in which units are either active ($a = 1$, black) or inactive ($a = 0$, white), and examples of possible activations produced in units of other layers are shown (darker for greater activation). Learning occurs through adjusting the weights to reduce the difference between the output of the network and a target output (upper right) [10,16]. In the case shown, the output activations are similar to the target, but there is some error to drive learning. There are no targets for internal or hidden layers (i.e., layers 2 and 3). These patterns depend on the connection weights, which in turn are shaped by the error-driven learning process.

modeling the neural computations supporting visual processing of objects in primates [17,18]. The considerable advantages of depth in allowing the learning of increasingly complex and abstract mappings [16] are balanced here by the strong interdependencies among connection weights in deep networks [19,20] such that the weights are learned gradually through extensive, repeated, and interleaved exposure to an ensemble of training examples that embody the domain statistics.

Although there are real advantages of a system using structured parametric representations, on its own such a system would suffer from two drastic limitations [1]. First, it is important to be able to base behavior on the content of an individual experience. For example, after experiencing a life-threatening situation – for example, an encounter with a lion at a watering-hole – it would clearly be beneficial to learn to avoid that particular location without the need for further encounters with the lion. The second problem is that the rapid adjustment of connection weights in a multilayer network to accommodate new information can severely disrupt the representation of existing knowledge in it – a phenomenon termed catastrophic interference [1,21–23] that is related to the stability–plasticity dilemma [24]. If the new information about the dangerous lion is forced into a multi-layer network by making large connection weight adjustments just to accommodate this item, this can interfere with knowledge of other less-threatening animals one may already be familiar with.

were never presented together (e.g., A and C).

Paired associative recall task: a paradigm where item pairs are experienced during study (e.g., word pairs such as 'dog–table' in a human experiment, or flavor–location pairs in a rodent experiment), and at test the individual must recall the other item (e.g., specific location) from a cue (the specific flavor, e.g., banana).

Recurrent similarity computation: recurrent similarity computation allows the procedure performed by exemplar models to iterate: that is, the retrieved products from the first step of similarity computation are combined with the external sensory input, and a subsequent round of similarity computation is performed. This process continues until a stable state (i.e., basin of attraction in a neural network) is reached. This allows the model to capture higher-order similarities present in a set of related experiences, where pairwise similarities alone are not informative.

Sharp-wave ripple (SWR): spontaneous neural activity occurring within the hippocampus during periods of rest and slow wave sleep, evident as negative potentials (i.e., sharp waves). Transient high-frequency (~150 Hz) oscillations (i.e., ripples) occur within these sharp waves, which can reflect the replay (i.e., reactivation) of activity patterns that occurred during actual experience, sped up by an order of magnitude.

Sparsity: the proportion of neurons in a given brain region that are active in response to a given stimulus ('population sparseness'). Sparse coding, where a small (e.g., 1%) proportion of neurons is active, is contrasted with densely distributed coding where a relatively large proportion of neurons are active (e.g., 20%).

Instance-Based Representation in the Hippocampal System

Fortunately, a second, complementary learning system can address both problems, affording the rapid and relatively individuated storage of information about individual items or experiences (such as the encounter with the lion). Following Marr's and subsequent proposals [25–27], the CLS theory proposed that the hippocampus and related structures in the medial temporal lobe (MTL) support the initial storage of item-specific information, including the features of the watering hole as well as those of the lion (Figure 1). This proposal has been captured in models of the role of the hippocampus in recognition memory for specific items and in sensitivity to context and co-occurrence of items within the same event or experience [28–36].

In CLS theory, the dentate gyrus (DG) and CA3 subregions of the hippocampus are the heart of the fast learning system (Boxes 2–4). The DG is crucial in selecting a distinct neural activity pattern in CA3 for each experience, even when different experiences are quite similar [25–27,37,38], a process known as pattern separation. Increases in the strengths of connections onto and among the participating neurons in DG and CA3 stabilize the activity pattern for an experience and support reactivation of the pattern from a partial cue: because of the strengthened connections, reactivation of part of the pattern that was activated during storage (features of the watering hole in which the lion was encountered) can then reactivate the rest of the pattern (i.e., the encounter with the lion), a process called 'pattern completion'. Return connections from hippocampus to neocortex then support adaptive behavior (e.g., avoidance of that location).

Note, however, that a hippocampal system acting alone would also be insufficient due to capacity limitations [26] and its limited ability to generalize. Related to the latter point, the use of pattern-separated hippocampal codes for related experiences – in contrast to the relatively dense similarity-based coding scheme thought to operate in the parametric neocortical system [17,39–46] – may be adaptive for some purposes but comes with a cost: it disregards shared structure between experiences, thereby limiting both efficiency and generalization.

The theory is supported by findings that neocortical activity patterns generally show less **sparsity** and exhibit greater similarity-based overlap compared to the hippocampus [17,40,41,43–47] (Boxes 4 and 5). It should be noted, however, that the degree of sparsity and similarity-based overlap varies across subregions of the hippocampus and neocortex (Boxes 4 and 5). While some of the relevant findings have been seen as supporting other theories [48], such differences are fully consistent with CLS and have long been exploited in CLS-based accounts of the roles of specific hippocampal subregions (Boxes 2–4). Similarly, learning rates vary across hippocampal areas in the theory (Box 2) and, likewise, there may be variation in learning rates across neocortical areas (Box 5).

Joint Contribution to Task Performance

In the CLS theory, the hippocampal and neocortical systems contribute jointly to performance in many tasks and many different types of memories. This point applies to tasks that are often thought of as tapping 'episodic memory' (memory for the elements of one specific experience), 'semantic memory' (knowledge of facts, e.g., about the properties of objects) or 'implicit memory' (performance enhancement as a consequence of prior experience that is not dependent on explicit recollection of the prior experience). In the CLS theory, tasks and types of memory are seen as falling on a continuum, with varying degrees of dependence on the two learning systems depending on task, item, and other variables. For example, consider the task of learning a list of paired associates. The list may contain a mixture of pairs with strong, weak, or no discernable prior association (e.g., dog–cat, heavy–suitcase, city–tiger). Recall of the second word of a pair when cued with the first is worse in hippocampal patients than controls, but both groups show better performance on items with stronger prior association [49]. The findings have

Box 2. Functional Roles of Subregions of the Medial Temporal Lobes

Work within the CLS framework [27,116,141] relies on the anatomical and physiological properties of MTL subregions and the computational insights of others [9,25,26] to characterize the computations performed within these structures.

Entorhinal Cortex (ERC) Input to the Hippocampal System

During an experience, inputs from neocortex produces a pattern of activation in the ERC that may be thought of as a compressed description of the patterns in the contributing cortical areas (Figure I: illustrative active neurons in the ERC are shown in blue). ERC neurons give rise to projections to three subregions of the hippocampus proper, the dentate gyrus (DG), CA1, and CA3 [28,84]. *Pattern selection and pattern separation*: novel ERC patterns are thought to activate a small set of previously uncommitted DG neurons (shown in red – these neurons may be relatively young neurons, created by neurogenesis). These neurons, in turn, select a random subset of neurons in CA3 via large 'detonator synapses' (shown as red dots on the projection from DG to CA3) to serve as the representation of the memory in CA3, ensuring that the new CA3 pattern is as distinct as possible from the CA3 patterns for other memories, including those for experiences similar to the new experience (Boxes 3 and 4). *Pattern completion:* recurrent connections from the active CA3 neurons onto other active CA3 neurons are strengthened during the experience, such that if a subset of the same neurons later becomes active, the rest of the pattern will be reactivated. Direct connections from ERC to CA3 are also strengthened, allowing the ERC input to directly activate the pattern in CA3 during retrieval without requiring DG involvement (Box 3). *Pattern reinstatement in ERC and neocortex* [116,141]: The connections from ERC to CA1 and back are thought to change relatively slowly to allow stable correspondence between patterns in CA1 and ERC. Strengthening of connections from the active CA3 neurons to the active CA1 neurons during memory encoding allows this CA1 pattern to be reactivated when the corresponding CA3 pattern is reactivated; the stable connections from CA1 to ERC then allow the appropriate pattern there to be reactivated, and stable connections between ERC and neocortical areas propagate the reactivated ERC pattern to the neocortex. Importantly, the bidirectional projections between CA1 and ERC, and between ERC and neocortex, support the formation and decoding of invertible CA1 representations of ERC and neocortical patterns, and allow recurrent computations. These connections should not change rapidly given the extended role of the hippocampus in memory – otherwise reinstatement in the neocortex of memories stored in the hippocampus would be difficult [61].

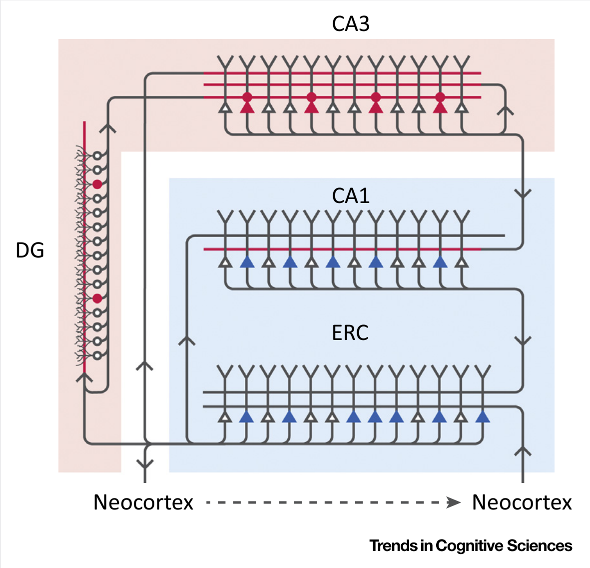


Figure I. Hippocampal Subregions, Connectivity, and Representation. Schematic depictions of neurons (with circular or triangular cell bodies) are shown, along with schematic depictions of projections from neurons in an area to neurons in the same or other areas (grey or colored lines – red coloring indicates projections with highly-plastic synapses, while grey coloring illustrates relatively less-plastic or stable projections). CA1 output to ERC then propagates out to neocortex; ERC and even resulting neocortical activity can be fed back into the hippocampus (broken line) as proposed in the REMERGE model (see below).

been captured in a model [50] (K. Kwok, PhD Thesis, Carnegie-Mellon University, 2003) in which hippocampal and neocortical networks jointly contribute to retrieval. Background associative knowledge is mediated by the cortex and the hippocampus mediates acquisition of associations linking each item pair to the learning context.

Box 3. Pattern Separation and Completion in Different Subregions of the Hippocampus

Pattern separation and completion [25–27] are defined in terms of transformations that affect the overlap or similarity among patterns of neural activity [28,142]. Pattern separation makes similar patterns more distinct through conjunctive coding [9,25], in which each output neuron responds only to a specific combination of active input neurons. Figures IA and IB illustrate how this can occur. Pattern separation is thought to be implemented in DG (see Box 4), using higher-order conjunctions that reduce overlap even more than illustrated in the figure.

Pattern completion is a process that takes a fragment of a pattern and fills in the remaining features (as in recalling a lion upon seeing the scene where the lion previously appeared) or that takes a pattern similar to a familiar pattern and makes it even more similar to it. Computational simulations [27] have shown how the CA3 region might combine features of pattern separation and completion, such that moderate and high overlap results in pattern completion toward the stored memory, but less overlap results in the creation of a new memory [37,133,143] (Figure IC). In this account, when environmental input produces a pattern in ERC similar to a previous pattern, the CA3 outputs a pattern closer to the one it previously used for this ERC pattern [124,144]. However, when the environment produces an input on the ERC that has low overlap with patterns stored previously, the DG recruits a new, statistically independent cell population in CA3 (i.e., pattern separation [27]). Emerging evidence suggests that the amount of overlap required for pattern completion (as well as other characteristics of hippocampal processing) may differ across the proximal-distal [145,146] and dorso–ventral axes [98,147–150] of the hippocampus, and may be shaped by neuromodulatory factors (e.g., Acetylcholine) [85,151]. Also, incomplete patterns require less overlap with a stored pattern than distorted ones for completion to occur, so that partial cues will tend to produce completion, as when one sees the watering hole and remembers seeing a lion there previously [27].

Several studies point to differences between the CA3 and CA1 regions in how their neural activity patterns respond to changes to the environment [37]: broadly, the CA1 region tends to mirror the degree of overlap in the inputs from the ERC while CA3 shows more discontinuous responses reflecting either pattern separation or completion [134,152].

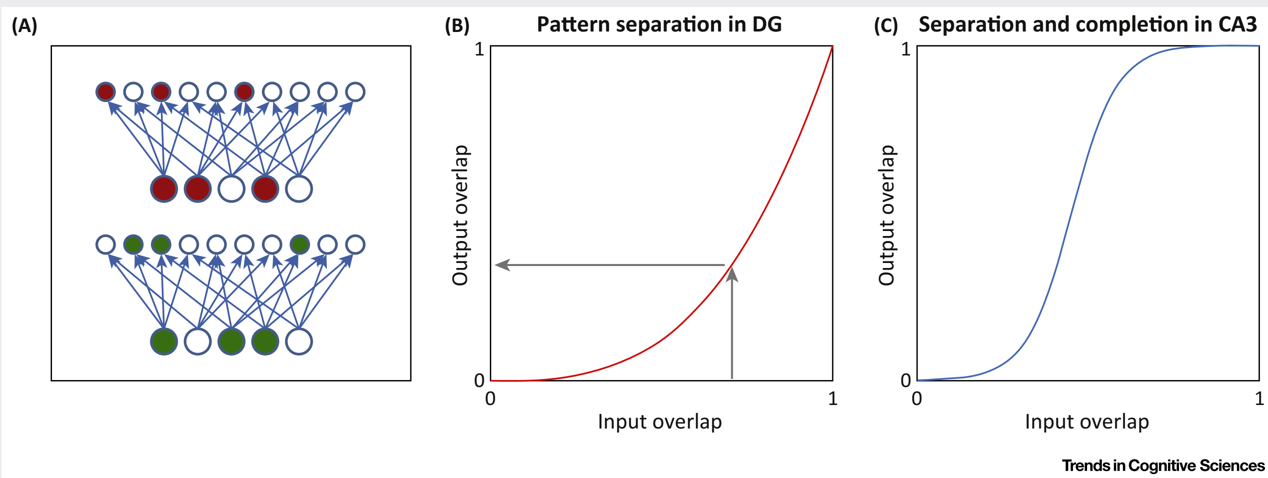


Figure I. Conjunctive Coding, Pattern Separation, and Pattern Completion. (A) A set of 10 conjunctive units with connections from a layer of 5 input units is shown twice with different input patterns. Here each conjunctive unit detects activity in a distinct pair of input units (arrows). The output for each pattern is sparser than the input (i.e., 30% vs 60%, respectively), and the two outputs overlap less than the two corresponding inputs (i.e., 33% vs 67%, respectively; overlap is the number of active units shared by two patterns divided by the number of units active in each). DG may use higher-order conjunctions, magnifying these effects. (B) An illustration of the general form of a pattern separation function, showing the relationship between input and output overlap. Arrows indicate the overlap of the inputs and outputs shown in the left panel. (C) The separation-and-completion profile associated with CA3, where low levels of input overlap are reduced further, while higher levels are increased [27,37].

Replay of Hippocampal Memories and Interleaved Learning

We now consider two important aspects of the CLS theory that are central foci of this review: the replay of hippocampal memories and interleaved learning. According to the theory, the hippocampal representation formed in learning an event affords a way of allowing gradual integration of knowledge of the event into neocortical knowledge structures. This can occur if the hippocampal representation can reactivate or replay the contents of the new experience back to the neocortex, interleaved with replay and/or ongoing exposure to other experiences [1]. In this way the new experience becomes part of the database of experiences that govern the values of the connections in the neocortical learning system [51–53]. Which other memories are selected for interleaving with the new experience remains an open question. Most simply, the hippocampus might replay recent novel experiences interleaved with all other recent experiences still stored in

Box 4. Sparse Conjunctive Coding and Pattern Separation in the Dentate Gyrus

Neuronal codes range from the extreme of localist codes – where neurons respond highly selectively to single entities ('grandmother cells') to dense distributed codes where items are coded through the activity of many (e.g., 50%) neurons in an area [153,154]. While localist codes minimize interference and are easily decodable, they are inefficient in terms of representational capacity. By contrast, dense distributed codes are capacity-efficient; however, they are costly in terms of metabolic cost and relatively difficult to decode. These are endpoints on a continuum quantified by a measure called sparsity, where 'population' sparsity indexes the proportion of neurons that fire in response to a given stimulus/location, and 'lifetime' sparsity indexes the proportion of stimuli to which a single neuron responds [26,153,155]. For example, a population sparsity of 1% means that only 1% of the neurons in a population are active in representing a given input. Two randomly selected sparse patterns tend to have low overlap (for two randomly selected patterns of equal sparsity over the same set of neurons, the average proportion of neurons in either pattern that is active in the other is equal to the sparsity), but neurons still participate in several different memories, making them more efficient than localist codes. Despite variability in estimates of the sparsity of a given brain region [27,153,156,157], the DG is widely believed to sustain among the sparsest neural code in the brain (~0.5–1% population sparseness) [25–27]. The CA3 region, to which the DG projects, is thought to be less sparse (~2.5% [47]). Many studies find less-sparse patterns in CA1 than CA3 [134,152].

The unique functional and anatomical properties of the DG suggest the origins of its sparse, pattern-separated code. The perforant path from the ERC (containing ~200 000 neurons in the rodent) projects to a layer of ~1 million of DG granule cells. Combined with the high levels of inhibition in the DG, this supports the formation of highly sparse, conjunctive representations, such that each neuron in DG responds only when several input neurons are simultaneously active, reducing overlap between similar input patterns [25–27,136]. Evidence also suggests that new DG neurons arise from stem cells throughout adult life; these new neurons may be preferentially recruited in the formation of memories [136], further reducing overlap with previously stored memories. The CA3 pattern for a memory is then selected by the active DG neurons, each of which has a 'detonator' synapse to ~15 randomly selected CA3 neurons. This process helps minimize the overlap of CA3 patterns for different memories, increasing storage capacity and minimizing interference between them, even if the two memories represent similar events that have highly overlapping patterns in neocortex and ERC. Empirical evidence provides support for this, with one study [137] showing that the representation supported by DG was highly sensitive to small changes in the environment, despite evidence that incoming inputs from the ERC were little affected (also see [133,145]). Furthermore, DG lesions impair an animals' ability to learn to respond differently in two very similar environments while leaving the ability to learn to respond differently in two environments that are not similar [136].

the hippocampus. A variant of this scheme would be for new experiences to be interleaved with related experiences activated by the new experience, through the dynamics of a recurrent mechanism (as in the REMERGE model [5], described below). An alternative possibility would be that interleaved learning does not actually involve the faithful replay of previous experiences: instead, hippocampal replay of recent experiences might be interleaved with activation of cortical activity patterns consistent with the structured knowledge implicit in the neocortical network (e.g., [23,54–56]).

Thus, the dual-system architecture proposed by CLS theory effectively harnesses the complementary properties of each of the two component systems, allowing new information to be rapidly stored in the hippocampus and then slowly integrated into neocortical representations. This process, sometimes labeled 'systems level consolidation' [51], arises, within the theory, from gradual cortical learning driven by replay of the new information, interleaved with other activity to minimize disruption of existing knowledge during the integration of the new information.

Empirical Evidence of Replay. Because of its centrality in the theory, we highlight key empirical evidence that replay events really do occur. The data come primarily from rodents, recorded during periods of inactivity (including sleep), in which hippocampal neurons exhibit large irregular activity (LIA) patterns that are distinct from the activity patterns observed during active states [2,3]. During LIA states, synchronous discharges thought to be initiated in hippocampal area CA3 produce **sharp-wave ripples** (SWRs), which are propagated to neocortex. SWRs reflect the reactivation of recent experiences, expressed as the sequential firing of so-called place cells, cells that fire when the animal is at a specific location [2,3,57–59]. These replay events appear to be time-compressed by a factor of about 20, bringing neuronal spikes that were well-separated in time during an actual experience into a time-window that enhances synaptic plasticity both

Box 5. Similarity-Based Coding in High-Level Visual Cortex

High-level visual regions of the neocortex are thought to support distributed representations that are inferred to be less sparse than those of the DG and the CA3/CA1 regions of the hippocampus (Box 4). Population sparseness in the ERC is estimated at 7–10% [158], with high-level sensory cortices exhibiting similar or higher levels of sparseness (e.g., variable estimates [44–46]). Although lifetime sparseness does not directly translate to population sparseness, recent evidence suggests that V4 and inferotemporal cortex (ITc) have a sparseness of ~10% on this measure [159]. It is worth noting that learning rates may vary according to neuronal selectivity and lifetime sparseness, resulting in differences in learning rates across neocortical areas and hippocampal subregions. Neurons in early visual regions that encode frequently-occurring features (i.e., edges) may have a relatively slow learning rate while neurons in higher visual regions and beyond (e.g., ITc and perirhinal cortex) may have a higher learning rate to support the encoding of less-frequently occurring, more-conjunctive features (e.g., individual objects) [12,160,161].

Evidence from electrophysiological recording studies in high-level visual cortical regions such as the ITc in primates provides support for the operation of a similarity-based coding scheme – whereby related categories (e.g., dogs and cats) are represented by overlapping neuronal codes [17,40–43] (Figure I). Representational similarity analysis (RSA) of the ITc population response during passive viewing of pictures reveals coding of fine-grained categorical structure (e.g., of a set of animate and inanimate objects) – that is well fit by deep convolutional neural networks which have algorithmic parallels with feedforward processing in the ventral visual stream [17,40]. While analogous similarity-based coding was observed using fMRI in the human homolog of ITc [41], there was no evidence for greater within-category (cf. between-category) representational similarity in any subregion of the hippocampus in a recent fMRI study [162] which found evidence consistent with the importance of pattern separation in episodic memory. Instead, similarity-based coding in this study was observed in the perirhinal and parahippocampal cortex – MTL regions that project to the ERC and that are typically considered to be intermediate zones (i.e., between the hippocampal and neocortical systems) in CLS theory.

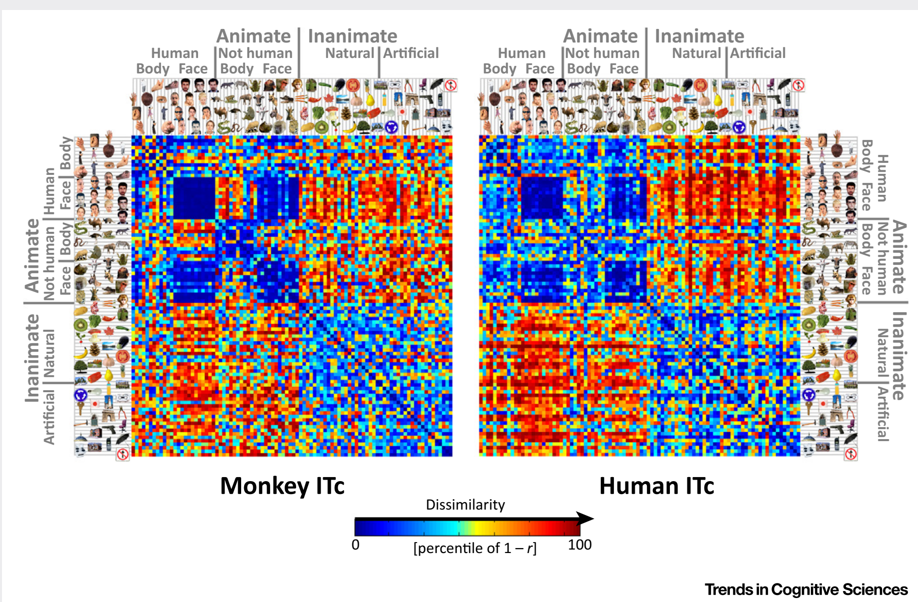


Figure I. Similarity-Based Coding in High-Level Visual Cortex. Representational dissimilarity matrices (RDM) reflect the correlation (i.e., $1 - r$, where r is the Pearson correlation coefficient) between the response of voxel patterns (fMRI in humans [41], right panel) or neuronal populations (electrophysiological recording in monkey [43], left panel) to a set of 92 object images. RDMs are analogous in monkey and human ITc. The RDMs show that the representations of animate objects are similar, as are those of inanimate objects. In addition to this clear animate–inanimate distinction, object coding in ITc exhibits finer categorical structure (e.g., for faces, body parts), visible in these RDMs (also see [41]). Reproduced with permission from [41].

within the hippocampus and between hippocampus and neocortex, and this allows a single event to be replayed many times during a single sleep period [1–3,58,60,61]. Consistent with the proposal that replay events are propagated to neocortex [1–3,60,61], SWRs within hippocampus are synchronized with fluctuations in neocortical activity states [62,63]. Also, hippocampal replay of specific place sequences has been shown to correlate with replay of patterns on grid cells located in the deep layers of the entorhinal cortex that receive the output of the hippocampal circuit [64], as well as more distant neocortical regions [65]. Furthermore, a recent study observed coordinated reactivation of hippocampal and ventral striatal neurons during slow-wave sleep (SWS), with location-specific hippocampal replay preceding activity in reward-sensitive striatal neurons [66]. A causal role for replay is supported by studies showing that the disruption of ripples in the hippocampus produces a significant impairment in systems-level consolidation in rats [67–69].

Additional Roles of Replay. Recent work has highlighted additional roles for replay – both during LIA but also during theta states [3,70,71] – well beyond its initially proposed role in systems-level consolidation. Specifically, recent evidence suggests that hippocampal replay can: (i) be non-local in nature, initiated by place cell activity coding for locations distant from the current position of the animal [3]; (ii) reflect novel shortcut paths by stitching together components of trajectories [72,73]; (iii) support look-ahead online planning during goal-directed behavior [70,74]; (iv) reflect trajectories through parts of environments that have only been seen but never visited [75]; and (v) be biased to reflect trajectories through rewarded locations in the environment [76]. Together, this evidence points to a pervasive role for hippocampal replay in the creation, updating, and deployment of representations of the environment [3]. Notably, these putative functions accord well with perspectives that emphasize the role of the human hippocampus in prospection [77], imagination [78,79], and the potential utility of episodic control of behavior over control based on learned summary statistics in some circumstances (e.g., given relatively little experience in an environment) [80].

Proposed Role for the Hippocampus in Circumventing the Statistics of the Environment. As we have seen, hippocampal activity during LIA does not necessarily reflect a faithful replay of recent experiences. Instead, mounting evidence suggests that replay may be biased towards rewarding events [59,76]. Building on this, we consider the broader hypothesis that the hippocampus may allow the general statistics of the environment to be circumvented by reweighting experiences such that statistically unusual but significant events may be afforded privileged status, leading not only to preferential storage and/or stabilization (as originally envisaged in the theory) but also leading to preferential replay that then shapes neocortical learning. We see this hippocampal reweighting process as being particularly important in enriching the memories of both biological and artificial agents, given memory capacity and other constraints as well as incomplete exploration of environments. These ideas link our perspective to rational accounts that view memory systems as being optimized to the goals of an organism rather than simply mirroring the structure of the environment [81].

A wide range of factors may affect the significance of individual experiences [82,83]: for example, they may be surprising or novel; high in reward value (either positive or negative) or in their informational content (e.g., in reducing uncertainty about the best action to take in a given state). The hippocampus – in receipt of highly processed multimodal sensory information [84] as well as neuromodulatory signals triggered by such factors [83,85] – is well positioned to reweight individual experiences accordingly. Indeed, recent work suggests specific molecular mechanisms that support the stabilization of memories and specific neuromodulatory projections to the hippocampus [83,86–88] that allow the persistence of individual experiences in the hippocampus to be modulated by events that occur both before and afterwards, providing mechanisms by which episodes may be retrospectively reweighted if their significance is enhanced by subsequent events [83,89], thereby influencing the probability of replay.

The importance of the reweighting capability of the hippocampus is illustrated by the following example. Over a multitude of experiences, consider a child gradually acquiring conceptual knowledge about the world, which includes the fact that dogs are typically friendly. Imagine that one day the child experiences an encounter with a frightening, aggressive dog – an event that would be surprising, novel, and charged with emotion. Ideally, this significant experience would not only be rapidly stored within the hippocampus but would also lead to appropriate updating of relevant knowledge structures in the neocortex. While CLS theory initially emphasized the role of the hippocampus in the first stage (i.e., initial storage of such one-shot experiences), here we highlight an additional role for the hippocampus in 'marking' salient but statistically infrequent experiences, thereby ensuring that such events are not swamped by the wealth of typical experiences – but instead are preferentially stabilized and replayed to the neocortex, thereby allowing knowledge structures to incorporate this new information. Although this reweighting would generally be adaptive, it could on occasion have maladaptive consequences. For example, in post-traumatic stress disorder, a unique aversive experience may be transformed into a persistent and dominant representation through a runaway process of repeated reactivation.

Challenges Arising from Recent Empirical Findings

In this section we discuss two significant challenges to the central tenets of CLS theory. Both challenges have recently been addressed through computational modeling work that extends and clarifies the principles of the theory.

The Hippocampus, Inference, and Generalization

Cross-Item Inferences

The first challenge concerns the role of the hippocampus in generalizing from specific experiences to novel situations. As noted above, CLS theory emphasized the crucial role of the hippocampus as a fast learning system relying on sparse activity patterns that minimize overlap even in the representation of very similar experiences. This representation scheme was thought to support memory of specifics, leaving generalization to the complementary neocortical system. Evidence presenting a substantial challenge to this account, however, has come from paradigms where individuals have been shown to rapidly utilize features that create links among a set of related experiences as a basis for a form of inference within or shortly after a single experimental session [4,5,90–95].

The **paired associate inference** (PAI) task [91,96–98] provides an example of a task that involves the hippocampus and captures the essence of requiring cross-item inferences required in other relevant tasks (such as the transitive inference task reviewed in [4]). In the study phase of the PAI task, subjects view pairs of objects (e.g., AB, BC) that are derived from triplets (i.e., ABC) or larger object sets (e.g., sextets: A, B, C, D, E, F; Box 6). In the crucial test trials, subjects are tested on their ability to appreciate the indirect relationships between items that were never presented together (e.g., A and F in the sextet version). Evidence for a role of the hippocampus in supporting inference in such settings [91,92,96–98] naturally raises the question of the neural mechanisms underlying this function, and has been seen as challenging the view that the hippocampus only stores separate representations of specific items or experiences. Indeed, the findings have been taken as supporting 'encoding-based overlap' models [4,95,99,100], in which it is proposed that the hippocampus supports inference by using representations that integrate or combine overlapping pairs of items (e.g., AB and BC in the triplet version of the PAI task).

While the PAI findings weigh against the view that the hippocampus only plays a role in the behaviors based on the contents of a single previous episode, these findings could arise from reliance on separate representations of the relevant AB, AC item pairs. Indeed, the CLS-grounded REMERGE model [5] proposes that representations combining elements of items never experienced together may arise from simultaneous activation of two or more memory traces within the hippocampal system, driven by an interactive activation process occurring within a recurrent circuit, whereby the output of the system can be recirculated back into it as a

Box 6. Generalization Through Recurrence in the Hippocampal System

The REMERGE model (Figure I) [5], which reflects a synthesis of interactive activation and competition (IAC) models [163] and **exemplar models** of memory [108,164,165], constitutes an abstraction and simplification of the multi-stage circuitry of the hippocampal system into two principal layers: feature and conjunctive layers, broadly corresponding to the ERC and hippocampus proper, respectively. The localist coding (e.g., unit AB) in the conjunctive layer reflects an idealization of the sparsely distributed pattern-separated codes in the DG/CA3 subregions of the hippocampus (Boxes 2–4) that support episodic memory (e.g., for trials involving presentation of A and B objects together).

An essential principle of the model – mediated by the bidirectional excitatory connections between feature and conjunctive layers – is the principle of recurrence between the hippocampus proper and neocortical regions such as the ERC (termed 'big-loop' recurrence, to distinguish it from the internal recurrence known to exist within the CA3 region). This allows recirculation of network output as a subsequent input to the system. Intuitively, this functionality is crucial to allowing the model to discover the higher-order structure present within a set of related episodes: an initial probe on the feature layer (e.g., denoting stimuli present on screen during a test trial) prompts the activation of experiences containing these elements on the conjunctive layer, which in turn drives a new pattern of feature layer activity that reflects not only the external input but also the content of retrieved experiences. This in turn leads to the activation of conjunctive units denoting experiences related to the new feature layer pattern, and so on. This can bring about a situation where, for example, the presentation of A and C can result in the activation of AB and BC, which jointly activate B, in turn further activating AB and BC which then suppress other conjuncts involving A and C. This produces a stable state in which AB, BC, and A, B, and C are all activated at the same time – thereby effectively inferring a link between A and C. Longer-range inferences (e.g., B–E) can also be supported by the recurrent mechanism ([5] for details). Formally, the function of the network can be viewed as carrying out recurrent similarity computation. Unlike other exemplar models [108,164,165], in which similarity computation is performed only on external inputs, REMERGE performs such computations on inputs affected by its own outputs.

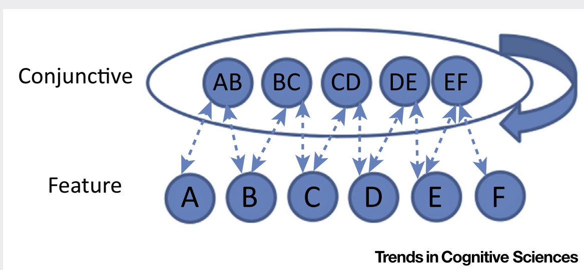


Figure I. A Schematic of the Architecture of REMERGE. Recurrent architecture of REMERGE, showing its two-layer architecture, with input/output units for possible constituents of experiences (A–F), conjunctive units representing pairs of constituents that have occurred together (AB, BC, etc.), bidirectional connections (broken arrows) between conjuncts and their constituents, and recurrent inhibition (broad arrow) among conjunctive units. Adapted from [5].

subsequent input (Box 6). This proposal is consistent with anatomical and physiological evidence [101] (Box 2).

REMERGE, therefore, can be considered to capture the insights of the relational theory of memory [29,102] by allowing the linkage of related episodes within a dynamic memory space, while preserving the assumption that the hippocampal system relies primarily on pattern-separated representations seen as essential for episodic memory [1,9,25–27,30,31,103,104]. Further, the recurrency within the hippocampal system makes the prediction that hippocampal activity may sometimes combine information from several separate episodes – a notion that receives empirical support from neuronal recordings in rodents [72,73]. This generalized replay – simultaneous reactivation of multiple related traces during testing or offline periods – may facilitate the creation of new representations from the recombination of multiple related episodes ('stored generalizations') [5] and the discovery of novel relationships (e.g., shortcuts) [72,73]. Empirical evidence also supports a role for the hippocampus in category- and so-called 'statistical' learning [105–107]: the mechanisms in REMERGE and other related models that rely on separate memory traces for individual items allow weak hippocampal traces that support only relatively poor item recognition to mediate near-normal generalization [5,108].

While encoding-based overlap and retrieval-based models make divergent experimental predictions, empirical evidence to date does not definitely distinguish between them ([4,5,98] for discussion). Indeed, it is conceivable that both mechanisms operate under different circumstances – perhaps as a function of the experimental paradigm under consideration, amount of training, and delay between training and testing (e.g., [109]). It is also worth noting that in reality the difference between encoding-based and retrieval-based models is not absolute: as alluded to above, generalized replay may facilitate the formation of new representations that directly capture distant relationships between items (e.g., the linear hierarchy in the transitive inference paradigm [5,90,110]). Such representations then become the contents of episodic memory, subject to storage in the hippocampus.

The distinction between encoding- and retrieval-based models can be related more broadly to the finding of 'concept' cells: hippocampal neurons which come to respond to common features across many events, for example cells for specific odors [111], time-points within an episode [112], attributes of a task [113], and even cells that fire to any picture or the name of a famous person [114]. In Box 7 we review empirical findings concerning concept cells and pattern overlap sometimes observed in parts of hippocampus, and consider how well these findings fit within the perspective that the hippocampus supports pattern separation.

Rapid Schema-Dependent Consolidation

It is useful to distinguish systems-level consolidation from what we refer to as within-system consolidation. The former refers to the gradual integration of knowledge into neocortical circuits, while the latter denotes stabilization of recently formed memories within the hippocampus, perhaps through stabilization of synapses among hippocampal neurons [89]. In the initial formulation of CLS, systems-level consolidation was viewed as temporally extended (e.g., spanning years or even decades in humans [34,51–53]). Although it was noted in [1] that the timeframe could be highly variable (depending, perhaps, on the rate of replay of memory

Box 7. Concept Cells and Nodal Codings?

Reports of concept cells in the hippocampus have been taken as contradicting a tenet of CLS theory, but the existence of such neurons is not necessarily inconsistent with it, given that the theory expects different hippocampal regions to vary in terms of context specificity and also permits variation within hippocampal regions (Box 3). Evidence supporting the CLS prediction of context-specificity in the CA3 and DG comes from a recent intracranial recording study in humans [166]. In this study, neurons in CA3/DG, and also in the subiculum, tended to discriminate between different images of a famous person – with responses correlating with successful performance in a recognition memory task that required discriminating previously experienced targets from similar lures. Neurons in other MTL areas (i.e., entorhinal and parahippocampal cortices) exhibited more invariant 'concept cell like' responses that were not linked to memory performance (the CA1 subregion was sparsely sampled in this study).

It is also interesting to consider the finding of 'splitter' cells in a task where animals must alternate between turning left and right on successive trials in a T maze [167–179]: here, some CA1 and CA3 place cells for locations on the central stem of the T maze are modulated by the trajectory of the rat (e.g., whether it will subsequently turn left or right) whereas others are trajectory-independent. This phenomenon, known as partial remapping [48,170–172], is consistent with the idea that pattern separation is a matter of degree in our theory [27,37]. As such, we should expect partly overlapping representations (i.e., rather than fully independent 'charts' [121]) when environmental changes are sufficiently small (Box 3). We also expect the greatest differentiation in DG, and at an early point in learning. To our knowledge no studies have yet recorded from DG in this paradigm.

In a recent study, representational similarity analysis techniques [173] were applied to ensemble recording data collected while rats performed a context-guided reward discrimination task [113]. As expected, the population codes in CA3 and CA1 were dominated by context and place coding, although other task dimensions – reward value and item – were also represented [113] (also see [174]). Although there was some representational overlap across locations based on value and item, CA3/CA1 codes were consistent with incomplete but still strong pattern separation, especially in the dorsal hippocampus. Overall, these findings appear consistent with the CLS, with the provision that pattern separation is a matter of degree, and may vary by task and region. Why CA3 shows greater specificity than CA1 in some studies but not others requires further exploration.

traces in the hippocampus), recent evidence suggests that this timeframe can be much shorter than anticipated (e.g., as little as a few hours to a couple of days) [7,8]. We focus on empirical data from the influential 'event arena' paradigm which demonstrated striking evidence of this phenomenon [7,8].

In the studies using this paradigm [7], rats were trained to forage for food in an event arena whose location was indicated by the identity of a flavor (e.g., banana) presented to the animal as a cue in a start box (Box 8). Learning of six such flavor–place paired associations (PAs) required multiple sessions distributed over several weeks, and was found to be hippocampus-dependent. Interestingly, although the learning of the original six PAs proceeded at a slow rate, rats were then able to learn two new PAs within the now familiar event arena based on a single exposure to each. Importantly, this one-shot learning was dependent on the presence of prior knowledge, often termed a 'schema': no such rapid learning was observed when rats that had been trained within one event arena were exposed to new PAs within a novel arena. Further, although the hippocampus must be intact for learning of new PAs in the familiar environment, memory for the new PAs remained robust when the hippocampus was surgically removed 2 days later. A follow-up study [8] provided insights into the neural basis of this phenomenon: the expression of genes associated with synaptic plasticity was significantly greater in neocortex very shortly (80 minutes) after rats experienced new PAs in the familiar arena compared to new PAs in the unfamiliar arena. Taken together, these results support the view that rapid systems-level consolidation, mediated by extensive synaptic changes in the neocortex within a short time after initial learning, is possible if the novel information is consistent with previously acquired knowledge.

At face value, the findings from the event arena paradigm [7,8] present a substantial challenge to a core tenet of CLS theory as originally stated: newly acquired memories, the theory proposed, should remain hippocampus-dependent for an extended time to allow for gradual interleaved learning such that integration into the neocortex can take place while avoiding the catastrophic forgetting of previously acquired knowledge. It is worth noting, however, that the simulations presented in the original CLS paper to illustrate the problem of catastrophic interference involved the learning of new information that is inconsistent with prior knowledge. As such, the relationship between the degree to which new information is schema-consistent and the timeframe of systems-level consolidation was not actually explored.

Recent work within the CLS framework [115] addressed this issue using simulations designed to parallel the key features of the event arena experiments [7,8] using the same neural network architecture and content domain that had been used in the original CLS paper as an illustration of the principles of learning in the neocortex. Briefly, the network was first trained to gradually acquire a schema (structured body of knowledge) about the properties of a set of individual animals (e.g., canary is a bird, can fly; salmon is a fish, can swim), paralleling the initial learning phase over several weeks in the event area paradigm ([115] for details). Next, the ability of this trained network to acquire new information was examined. The network was trained on a new item X, whose features were either consistent or inconsistent with prior knowledge (e.g., X is a bird and X can fly, consistent with known birds, or X is a bird but can swim, not fly, inconsistent with the items known to the network), thus mirroring the learning of new PAs under schema-consistent and schema-inconsistent conditions in the event arena studies [7]. Notably, the network exhibited rapid learning of schema-consistent information without disrupting existing knowledge, while schema-inconsistent information was acquired much more slowly and necessitated interleaved training with the already-known examples (e.g., canary) to avoid catastrophic interference. Interestingly, there was also a clear relationship between the profile of weight changes occurring in the network and the consistency of the information being learned. Specifically, even though the same small value of the learning rate parameter was

Box 8. Rapid Integration of New Learning in the Neocortex: When Does it Occur?

In the event arena paradigm [7,8] (Figure I), hippocampal lesions prevent acquisition of new schema-consistent associations. By contrast, hippocampal lesions performed as little as 48 h after learning leave memory intact. One explanation for the crucial but temporary nature of the hippocampal contribution is replay: even a few minutes with the hippocampus intact could allow multiple replays, each one incrementing the strength of intra-neocortical connections. In an investigation of induction of plasticity-related genes in neocortex [8] the hippocampus was intact for 80 minutes after initial exposure to the new associations. These findings raise the broader question of when rapid integration of new learning into the neocortex occurs, and whether it can occur even without a hippocampus.

A substantial body of work from several laboratories now supports the view that a single period of sleep can produce changes in how experiences from a single learning session impact on subsequent responding. As key examples, some studies have reported increased levels of linking inferences [175], and others have reported increased lexical competition and related phenomena [109,176] attributed to a single sleep session. These findings are often interpreted as evidence of rapid systems-level consolidation (e.g., [176]). However, the materials used are not obviously highly consistent with prior knowledge in most cases, and therefore under the CLS framework we would not expect full integration into neocortical networks in such a short time-period. An alternative interpretation (illustrated in [5]) is that replays during sleep increase the strength, robustness, and rate of activation of new hippocampus-dependent traces, and that such strengthening may be sufficient to account for the observed effects. Thus, the findings are consistent with the view that integration of these new memories into neocortical structures proceeds over a considerably longer time-period.

Work with the ‘fast mapping’ paradigm in humans with hippocampal lesions [177] provides another potential source of evidence about rapid neocortical learning of arbitrary new information. In this paradigm, human participants see pairs of pictures of objects – one familiar and one unfamiliar – and are asked a question such as ‘is the numbat's tail pointing up’, inferring that the unfamiliar name ‘numbat’ must refer to the unfamiliar object [177]. Some studies find that patients with extensive hippocampus damage show retention of the new object–name association at a delayed test [178,179], suggesting very rapid neocortical learning even without a hippocampus. However, the finding has proven difficult to replicate [180–182]; future studies should continue to investigate this issue.

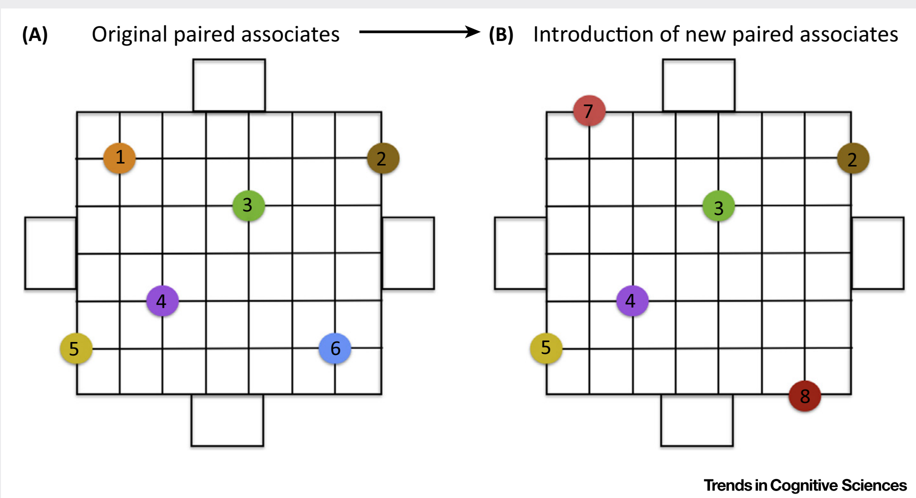


Figure I. Schematic Illustration of the Event Arena Paradigm. (A) Overhead view of 1.6 m × 1.6 m event arena: rats are cued with one of six food flavors (e.g., banana) each associated with a location in the arena (e.g., location 3) and are required to go from any of the four start-boxes to a specific location to retrieve food. (B) Following gradual learning of the original set, two new flavor-place pairs are introduced: (e.g., cinnamon–location7; nutmeg–location8). Rapid schema-dependent one-shot learning of these new PAs is observed (see Box text). Figure based on experimental design described in [7].

used in both simulations, large amplitude weight changes occurred during the learning of schema-consistent, but not schema-inconsistent, information – emulating the schema-dependent pattern of neocortical plasticity-related gene expression reported in [8]. A theoretical analysis of multilayer neural networks makes clear why the model exhibits these effects [20]: the analysis shows that the rate of learning within a multilayered neural network of the type that CLS attributes to the neocortex [20] will always depend on the state of knowledge within the network as well as on the compatibility of new inputs with the structured system this knowledge represents.

The analysis described above thus addresses the challenge to CLS theory posed by the findings from the event arena paradigm [7,8] (Box 8 discusses other issues related to rapid systems-level consolidation). Taken together, this empirical and theoretical research highlights the need for two amendments to the theory as originally stated [115]. First, consider the core tenet of the theory that the incorporation of novel information into neocortical networks must be slow to avoid catastrophic interference: we now know that this statement only applies when new information is inconsistent with existing knowledge in the neocortex. The second important amendment relates to the original dichotomy between the slow-learning neocortical system and a fast-learning system instantiated in the hippocampus. The empirical data, simulations, and theoretical work summarized above demonstrate that the neocortex does not necessarily learn slowly. More accurately, we now characterize the rate of learning in the neocortex as being dependent on prior knowledge rather than being slow *per se*. Because input to the hippocampus depends on the structured knowledge in the cortex, it follows that hippocampal learning will also be dependent on prior knowledge [116]. Future research should explore this issue.

Links Between CLS Theory and Machine-Learning Research

The core principles of CLS theory have broad relevance not only in understanding the organization of memory in biological systems but also in designing agents with artificial intelligence. We discuss here connections between aspects of CLS theory and recent themes in machine-learning research.

Deep Neural Networks and the Slow-Learning Neocortical System

Very deep networks [16], sometimes with more than 10 layers, grew out of earlier computational work [15,16,117] on networks with only a few layers, which were used to model the essential principles of the slow-learning neocortical system within the CLS framework. In general, therefore, deep networks share the characteristics of the slow-learning neocortical system discussed previously: they achieve an optimal parametric characterization of the statistics of the environment by learning gradually through repeated, interleaved exposure to large numbers of training examples.

In recent years, deep networks have achieved state-of-the-art performance in several domains, including image recognition and speech recognition [16], made possible through increased computing power and algorithmic development. Their power resides in their ability to learn successively more abstract representations from raw sensory data (e.g., the image of an object) – for example oriented edges, edge combinations, and object parts – through composing multiple processing layers that perform non-linear transformations. One class of deep networks, termed convolutional neural networks (CNNs), has been particularly successful in achieving state-of-the-art performance in challenging object-recognition tasks (e.g., ImageNet [118]). CNNs are particularly suited to the task of object recognition because their architecture naturally builds in robustness to changes in position through the use of a hierarchy of convolutional filters where units within a feature map at each layer share the same weights, thereby allowing them to detect the same feature at different locations. Interestingly, CNNs have recently also been shown to provide a good model of object recognition in primates at both behavioral and neural levels (e.g., V4, inferotemporal cortex) [17,18,40].

Neural Networks and Replay

For the purposes of machine learning, deep networks are often trained in interleaved fashion because the examples from the entire dataset are available throughout. This is not generally the case, however, in a developmental or online learning context, when intelligent agents need to

Box 9. Experience Replay in Deep Q-Networks

Instead of employing a standard online learning method in which each unit of play experience (consisting of a state, action, next state and resulting reward) is used immediately to adjust connection weights and then discarded, an experience replay buffer similar to the hippocampus is used. This allows learning based on randomly chosen subsets of recent experiences stored in the replay buffer ([119] for details) to be interleaved with ongoing game-play. The approach is in line with findings cited above [66] that hippocampal replay reactivates reward related neurons in striatum, in accord with the hypothesis that hippocampus-dependent RL facilitates learning during off-line periods.

Experience replay in the DQN architecture was crucial in (i) maximizing data efficiency, allowing each unit of experience to be reused in many updates (e.g., mirroring benefits of repeated time-compressed hippocampal replay) and (ii) smoothing out learning and avoiding unstable response policies that can result from the tendency of the current policy to bias the experienced samples. The approach minimizes learning from consecutive samples, which is undesirable owing to their strongly correlated nature and inconsistent with the implicit assumptions built into neural-network learning algorithms. Instead experience replay allows updates within the deep Q-network to be performed on non-adjacent samples from a set of recent experiences in a fashion that breaks up these correlations while still relying on relevant statistics. The dramatic advantage of a network implementing interleaved learning through experience replay was illustrated by the effects of disabling replay on network performance: this caused a severe drop in performance to at best ~30% of when experience replay was present [119]. Note that the uniform sampling mechanism as implemented treats all transitions in the replay memory as if they were equal. Recent work [183] shows that biasing replay towards significant events – specifically, experiences that are associated with high reward prediction errors – yields further gains. This mechanism, which resonates with the role of the hippocampus in reweighting experiences as discussed above, allows information to be harvested from rare experiences that may be particularly informative.

learn and make decisions while gathering experiences and/or where the data distribution is changing perhaps as the agent's abilities change. Here, recent machine-learning research has drawn inspiration from CLS theory about the role of hippocampal replay. Implementation of an 'experience replay' mechanism was crucial to developing the first neural network (Deep Q-Network or DQN) capable of achieving human-level performance across a wide variety of Atari 2600 games by successfully harnessing the power of deep neural networks and reinforcement learning (RL) [119] (Box 9).

Continual Learning and the Hippocampus

Continual learning – the name machine-learning researchers use for the ability to learn successive tasks in sequential fashion (e.g., tasks A, B, C) without catastrophic forgetting of earlier tasks (e.g., task A) – remains a fundamental challenge in machine-learning research, and addressing it can be considered a prerequisite to developing artificial agents we would consider truly intelligent. A principal motivation for incorporating a fast-learning hippocampal system as a complement to the slow neocortical system in CLS theory was to support continual learning in the neocortex: hippocampal replay was proposed to mediate interleaved training of the neocortex (i.e., intermixed examples from tasks A, B, C) despite the sequential nature of the real-world experiences. We draw attention here to a second relatively underexplored reason why the hippocampus may facilitate continual learning, particularly over relatively short timescales (i.e., before systems-level consolidation).

As discussed previously, the hippocampus is thought to represent experiences in pattern-separated fashion, whereby in the idealized case even highly similar events are allocated neuronal codes that are non-overlapping or orthogonal (e.g., [26]). Notably, the advantages of this coding scheme for episodic memory – reduction of interference between similar but distinct events – may also have significant benefits for continual learning. Specifically, this mechanism allows the rapid creation of distinct non-interfering representations for multiple tasks to which an agent has been exposed in sequential fashion. The utility of this function, and the ubiquity of continual learning, is well established in the domain of spatial navigation, where the notion of a task can be related to that of an environmental context: rodents are able to learn and sustain robust representations of many different environments (e.g., >10 environments in [120]), with each environment being represented by a pattern-separated representational space

Box 10. Neural Networks with External Memory and the Hippocampus

The neural Turing machine (NTM) [125] consists of two basic components: an external memory and a neural network controller that is distinguished by its ability to interact with the external memory (Figure I). An external memory allows specific inputs (such as items to be remembered) or the results of intermediate computations to be written to it, and then to be read out in a content- or location-based addressable fashion [184].

The controller interacts with the external memory through write and read heads that focus on particular parts of the memory matrix through attentional addressing mechanisms. Content-based addressing focuses attention on memory slots based on their similarity to the current values (i.e., 'key') emitted by the controller. The graded, similarity-based nature of these addressing mechanisms allows the architecture to be trained using the continuous learning signals that drive learning in other deep neural networks [10]. The controller may be a feedforward network, but is more typically a recurrent network exploiting specialized long-short-term memory (LSTM) modules [185] that can learn to retain information over very extended numbers of time-steps. In contrast to standard neural networks, the architecture of the NTM allows a separation of computation from memory, as in conventional computers [125]. This allows the NTM to learn to perform algorithms independently of the variables concerned (also see [186]).

While parallels have been drawn between the external memory of the NTM and working memory [125], the characteristics of its external memory can easily be related to long-term memory systems as well. Indeed, content-based addressable external memories of this kind share functionalities with attractor networks [145], an architecture often used to model the computational functions performed by the CA3 subregion of the hippocampus (e.g., storage and retrieval of episodic memories) [187]. There are further points of connection between the operation of the NTM and the hippocampus: information is not stored and retained indiscriminately; instead it is selected based on an estimate of potential future relevance (see section 'Proposed Role for the Hippocampus in Circumventing the Statistics of the Environment').

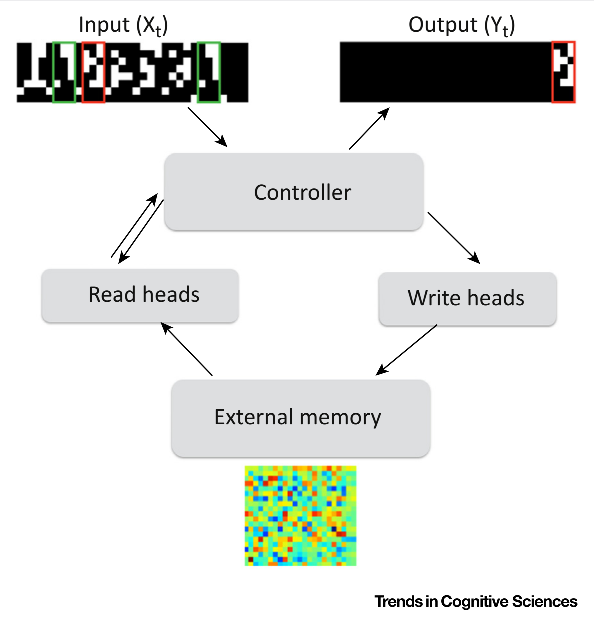


Figure I. NTM and the Paired Associative Recall Task. The input to the controller is a sequence of column vectors. The network receives one column per time-step, and the figure shows the columns presented over 29 consecutive time-steps indexed by t. The input here consists of a sequence of items, where each item is three binary random vectors presented in adjacent time-steps. Two items are highlighted, one in a green box and one in a red box. A delimiter symbol (in row 4) appears in the time-step preceding each item. After three items have been presented, a different delimiter symbol (row 5) occurs followed by a query (single item in green box). The network responds correctly with the appropriate target (red box). Schematic representation of external memory matrix shown. Adapted with permission from [125].

(putatively implemented as a continuous attractor, called a 'chart' [121]) within the CA3 subregion of the hippocampus, and within which specific locations are further individuated [122–124].

Neural Networks with External Memory and the Hippocampus

Recent work has suggested that deep networks may be considerably enhanced by the addition of an external memory. For example, an external memory is used in the neural Turing machine

(NTM) [125], and this memory has content-addressable properties akin to those of the **attractor networks** used to model pattern completion in the hippocampus (Box 10). Such an external memory has been shown to support functionalities such as the learning of new algorithms, for example performing **paired associative recall** (Box 10 [125]) or question-and-answering (Q&A [126,127]) – a class of machine-learning paradigms where textual outputs are required based on queries (e.g., Q: where is Bill?; A: the bathroom) requiring inference over a knowledge database (e.g., a set of sentences).

It is also worth noting that the neuropsychological testing of story recall can be considered to be a version of the Q&A task used in machine learning (e.g., [126]). When the amount of story content to be retained exceeds a few sentences, this task is crucially dependent on the memory storage properties of the hippocampus. Indeed, the specific working of the REMERGE model of the hippocampus – **recurrent similarity computation**, such that the output of the episodic system is recirculated as a new input – has parallels in a recent machine-learning algorithm developed for the purpose of Q&A, termed a 'memory network' [127]. Specifically, a learned, dense feature-vector representation of an input query (e.g., 'where is the milk?') is used to retrieve the sentence with the most similar feature vector in the database (e.g., 'Joe left the milk'): a combined feature representation of the initial query and retrieved sentence is then used to identify similar sentences earlier in the story ('Joe traveled to the office'); this process iterates until a response is emitted by the network ('the office'). The joint dependence of this system on input/output feature representations that are developed gradually through training with a large corpus of text and on individual stored sentences nicely parallels the complementary roles of neocortical and hippocampal representations in CLS theory and REMERGE.

Concluding Remarks

We have argued that the core features of the memory architecture proposed by CLS theory continue to provide a useful framework for understanding the organization of learning systems in the brain. We have, however, refined and extended the theory in several ways. First, we now encompass a broader and more-significant role for the hippocampus in generalization than previously thought. Second, we have amended the statement that neocortical learning is constrained to be slow *per se* – instead, we now clarify that the rate of neocortical learning is dependent on prior knowledge and can be relatively fast under some conditions. Together, these revisions to the theory imply a softening of the originally strict dichotomy between the characteristics of neocortical (slow learning, parametric, and therefore generalizing) and hippocampal (fast-learning, item-based) systems. In addition, we have extended the proposed functions for the fast-learning hippocampal system, suggesting that this system can circumvent the general statistics of the environment by reweighting experiences that are of significance. Finally, we have highlighted the broad applicability of the principles of CLS theory to developing agents with artificial intelligence, an area which we hope will continue to rise in interest and become a significant direction for future research (see Outstanding Questions).

Outstanding Questions

Under what conditions does the proposed hippocampal reweighting of experiences result in a biased neocortical model of environmental structure?

Are hippocampal representations updated to incorporate changes in neocortical representations (the 'index maintenance' problem), and if so how?

What is the fate of hippocampal memory traces after systems-level consolidation is complete?

What are the precise conditions under which rapid systems-level consolidation can occur?

Are hippocampal memory traces susceptible to reconsolidation in a way that mirrors amygdala-dependent memories (e.g., in fear-conditioning paradigms)?

What neocortical mechanisms complement hippocampal replay in facilitating continual learning?

What algorithmic functionalities and implementational schemes are desirable for an external memory module, both for human learners and for artificial agents?

Acknowledgments

We are very grateful to Adam Cain for help with creating the figures and Greg Wayne and Nikolaus Kriegeskorte for comments on an earlier version of the paper.

References

1. McClelland, J.L. *et al.* (1995) 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
2. O'Neill, J. *et al.* (2010) Play it again: reactivation of waking experience and memory. *Trends Neurosci.* 33, 220–229
3. Wikenheiser, A.M. and Redish, A.D. (2015) Decoding the cognitive map: ensemble hippocampal sequences and decision making. *Curr. Opin. Neurobiol.* 32, 8–15
4. Zeithamova, D. *et al.* (2012) The hippocampus and inferential reasoning: building memories to navigate future decisions. *Front. Hum. Neurosci.* 6, 1–14

5. Kumaran, D. and McClelland, J.L. (2012) Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychol. Rev.* 119, 573–616
6. Eichenbaum, H. (2004) Hippocampus: cognitive processes and neural representations that underlie declarative memory. *Neuron* 44, 109–120
7. Tse, D. *et al.* (2007) Schemas and memory consolidation. *Science* 316, 76–82
8. Tse, D. *et al.* (2011) Schema-dependent gene activation and memory encoding in neocortex. *Science* 333, 891–895
9. Marr, D. (1971) Simple memory: a theory for archicortex. *Philos. Trans. R. Soc. L. B Biol. Sci.* 262, 23–81
10. Rumelhart, D.E. *et al.* (1986) Learning representations by back-propagating errors. *Nature* 323, 533–536
11. Sejnowski, T.J. and Rosenberg, C.R. (1987) Parallel networks that learn to pronounce English text. *Complex Syst.* 1, 145–168
12. Guyonneau, R. *et al.* (2004) Temporal codes and sparse representations: a key to understanding rapid processing in the visual system. *J. Physiol. Paris* 98, 487–497
13. Plaut, D.C. *et al.* (1996) Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115
14. Rogers, T.T. and McClelland, J.L. (2004) *Semantic Cognition: A Parallel Distributed Processing Approach,* MIT Press
15. Rumelhart, D.E. (1990) Brain style computation: learning and generalization. In *An Introduction to Electronic and Neural Networks* (Zornetzer, S.F. *et al.*, eds), pp. 405–420, Academic Press
16. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
17. Yamins, D.L. *et al.* (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624
18. Yamins, D.L. and DiCarlo, J.J. (2016) Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365
19. Saxe, A.M. *et al.* (2015) Learning hierarchical categories in deep neural networks. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pp. 1271–1276, Cognitive Science Society
20. Saxe, A.M. *et al.* (2014) *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks.*
21. McCloskey, M. and Cohen, N.J. (1989) Catastrophic forgetting in connectionist networks: the problem of sequential learning. In *The Psychology of Learning and Motivation (Vol. 20)* (Bower, G. H., ed.), pp. 109–165, Academic Press
22. Ratcliff, R. (1990) Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychol. Rev.* 97, 285–308
23. French, R.M. (1999) Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* 3, 128–135
24. Carpenter, G.A. and Grossberg, S. (1987) A massively parallel architecture for a self-organizing neural pattern recognition architecture. *Comput. Vision, Graph. Image Process.* 37, 54–115
25. McNaughton, B.L. and Morris, R.G. (1987) Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends Neurosci.* 10, 408–415
26. Treves, A. and Rolls, E.T. (1992) Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus* 2, 189–199
27. O'Reilly, R.C. and McClelland, J.L. (1994) Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus* 4, 661–682
28. Knierim, J.J. *et al.* (2006) Hippocampal place cells: parallel input streams, subregional processing, and implications for episodic memory. *Hippocampus* 16, 755–764
29. Cohen, N.J. and Eichenbaum, H.B. (1994) *Memory, Amnesia and the Hippocampal System,* MIT Press
30. O'Reilly, R.C. and Rudy, J.W. (2001) Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychol. Rev.* 108, 311–345
31. Norman, K.A. and O'Reilly, R.C. (2003) Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychol. Rev.* 110, 611–646
32. Mayes, A. *et al.* (2007) Associative memory and the medial temporal lobes. *Trends Cogn. Sci.* 11, 126–135
33. Davachi, L. (2006) Item, context and relational episodic encoding in humans. *Curr. Opin. Neurobiol.* 16, 693–700
34. Squire, L.R. *et al.* (2004) The medial temporal lobe. *Annu. Rev. Neurosci.* 27, 279–306
35. Schiller, D. *et al.* (2015) Memory and space: towards an understanding of the cognitive map. *J. Neurosci.* 35, 13904–13911
36. O'Reilly, R.C. *et al.* (2014) Complementary learning systems. *Cogn. Sci.* 38, 1229–1248
37. Knierim, J.J. and Neunuebel, J.P. (2016) Tracking the flow of hippocampal computation: pattern separation, pattern completion, and attractor dynamics. *Neurobiol. Learn. Mem.* 129, 38–49
38. Johnston, S.T. *et al.* (2016) Paradox of pattern separation and adult neurogenesis: a dual role for new neurons balancing memory resolution and robustness. *Neurobiol. Learn. Mem.* 129, 60–68
39. Bengio, Y. *et al.* (2013) Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1798–1828
40. Khaligh-Razavi, S.M. and Kriegeskorte, N. (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10, e1003915
41. Kriegeskorte, N. *et al.* (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141
42. Clarke, A. and Tyler, L.K. (2014) Object-specific semantic coding in human perirhinal cortex. *J. Neurosci.* 34, 4766–4775
43. Kiani, R. *et al.* (2007) Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J. Neurophysiol.* 97, 4296–4309
44. McNaughton, B.L. (2010) Cortical hierarchies, sleep, and the extraction of knowledge from memory. *Artificial Intell.* 174, 205–2014
45. Leibold, C. and Kempter, R. (2008) Sparseness constrains the prolongation of memory lifetime via synaptic metaplasticity. *Cereb. Cortex* 18, 67–77
46. Rolls, E.T. *et al.* (1997) The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. *Exp. Brain Res.* 114, 149–162
47. Barnes, C.A. *et al.* (1990) Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Prog. Brain Res.* 83, 287–300
48. McKenzie, S. *et al.* (2015) Representation of memories in the cortical–hippocampal system: results from the application of population similarity analyses. *Neurobiol. Learn. Mem.* Published online December 31, 2015. http://dx.doi.org/10.1016/j.nlm.2015.12.008
49. Cutting, J. (1978) A cognitive approach to Korsakoff's syndrome. *Cortex* 14, 485–495
50. McClelland, J.L. (2011) Memory as a constructive process: the parallel-distributed processing apporach. In *The Memory Process: Neuroscientific and Humanist Perspectives* (Nalbantian, P. *et al.*, eds), pp. 99–129, MIT Press
51. Frankland, P.W. and Bontempi, B. (2005) The organization of recent and remote memories. *Nat. Rev. Neurosci.* 6, 119–130
52. Winocur, G. *et al.* (2010) Memory formation and long-term retention in humans and animals: convergence towards a transformation account of hippocampal–neocortical interactions. *Neuropsychologia* 48, 2339–2356
53. Squire, L.R. *et al.* (1984) The medial temporal region and memory consolidation: a new hypothesis. In *Memory Consolidation: Psychobiology of Cognition* (Weingartner, H. and Parker, E.S., eds), pp. 185–210, Psychology Press
54. Robins, A. (1996) Consolidation in neural networks and in the sleeping brain. *Conn. Sci.* 8, 259–276
55. Tononi, G. and Cirelli, C. (2014) Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron* 81, 12–34

56. Norman, K.A. *et al.* (2005) Methods for reducing interference in the complementary learning systems model: oscillating inhibition and autonomous memory rehearsal. *Neural Netw.* 18, 1212–1228
57. Skaggs, W.E. and McNaughton, B.L. (1996) Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science* 271, 1870–1873
58. Wilson, M.A. and McNaughton, B.L. (1994) Reactivation of hippocampal ensemble memories during sleep. *Science* 265, 676–679
59. Carr, M.F. *et al.* (2011) Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nat. Neurosci.* 14, 147–153
60. Buzsaki, G. (1989) Two-stage model of memory trace formation: a role for 'noisy' brain states. *Neuroscience* 31, 551–570
61. Kali, S. and Dayan, P. (2004) Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nat. Neurosci.* 7, 286–294
62. Sirota, A. *et al.* (2003) Communication between neocortex and hippocampus during sleep in rodents. *Proc. Natl. Acad. Sci. U.S. A.* 100, 2065–2069
63. Battaglia, F.P. *et al.* (2004) Hippocampal sharp wave bursts coincide with neocortical 'up-state' transitions. *Learn. Mem.* 11, 697–704
64. Ólafsdóttir, H. *et al.* (2016) Coordinated grid and place cell replay during rest. *Nat. Neurosci.* Published online April 18, 2016. http://dx.doi.org/10.1038/nn.4291
65. Ji, D. and Wilson, M.A. (2007) Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci.* 10, 100–107
66. Lansink, C.S. *et al.* (2009) Hippocampus leads ventral striatum in replay of place–reward information. *PLoS Biol.* 7, e1000173
67. Ego-Stengel, V. and Wilson, M.A. (2010) Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus* 20, 1–10
68. Girardeau, G. *et al.* (2009) Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.* 12, 1222–1223
69. Nakashiba, T. *et al.* (2009) Hippocampal CA3 output is crucial for ripple-associated reactivation and consolidation of memory. *Neuron* 62, 781–787
70. Johnson, A. and Redish, A.D. (2007) Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* 27, 12176–12189
71. Wikenheiser, A.M. and Redish, A.D. (2015) Hippocampal theta sequences reflect current goals. *Nat. Neurosci.* 18, 289–294
72. Wu, X. and Foster, D.J. (2014) Hippocampal replay captures the unique topological structure of a novel environment. *J. Neurosci.* 34, 6459–6469
73. Gupta, A.S. *et al.* (2010) Hippocampal replay is not a simple function of experience. *Neuron* 65, 695–705
74. Pfeiffer, B.E. and Foster, D.J. (2013) Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* 497, 74–79
75. Ólafsdóttir, H.F. *et al.* (2015) Hippocampal place cells construct reward related sequences through unexplored space. *Elife* 4, e06063
76. Bendor, D. and Wilson, M.A. (2012) Biasing the content of hippocampal replay during sleep. *Nat. Neurosci.* 15, 1439–1444
77. Schacter, D.L. and Addis, D.R. (2007) The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philos. Trans. R. Soc. B Biol. Sci.* 362, 773–786
78. Hassabis, D. and Maguire, E.A. (2007) Deconstructing episodic memory with construction. *Trends Cogn. Sci.* 11, 299–306
79. Hassabis, D. *et al.* (2007) Patients with hippocampal amnesia cannot imagine new experiences. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1726–1731
80. Lengyel, M. and Dayan, P. (2007) Hippocampal contributions to control: the third way. *Neural Inf. Process. Syst.*
81. Anderson, J.R. and Milson, R. (1989) Human memory: an adaptive perspective. *Psychol. Rev.* 96, 703
82. Lisman, J.E. and Grace, A.A. (2005) The hippocampal–VTA loop: controlling the entry of information into long-term memory. *Neuron* 46, 703–713
83. Lisman, J. *et al.* (2011) A neoHebbian framework for episodic memory; role of dopamine-dependent late LTP. *Trends Neurosci.* 34, 536–547
84. van Strien, N.M. *et al.* (2009) The anatomy of memory: an interactive overview of the parahippocampal-hippocampal network. *Nat. Rev. Neurosci.* 10, 272–282
85. Hasselmo, M.E. (1999) Neuromodulation: acetylcholine and memory consolidation. *Trends Cogn. Sci.* 3, 351–359
86. McNamara, C.G. *et al.* (2014) Dopaminergic neurons promote hippocampal reactivation and spatial memory persistence. *Nat. Neurosci.* 17, 1658–1660
87. Sara, S.J. (2009) The locus coeruleus and noradrenergic modulation of cognition. *Nat. Rev. Neurosci.* 10, 211–223
88. McGaugh, J.L. (2004) The amybdala modulates the consolidation of memories of emotionally arousing experiences. *Annu. Rev. Neurosci.* 27, 1–28
89. Redondo, R.L. and Morris, R.G. (2011) Making memories last: the synaptic tagging and capture hypothesis. *Nat. Rev. Neurosci.* 12, 17–30
90. Kumaran, D. (2012) What representations and computations underpin the contribution of the hippocampus to generalization and inference? *Front. Hum. Neurosci.* 6, 157
91. Bunsey, M. and Eichenbaum, H. (1996) Conservation of hippocampal memory function in rats and humans. *Nature* 379, 255–257
92. Zeithamova, D. and Preston, A.R. (2010) Flexible memories: differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. *J. Neurosci.* 30, 14676–14684
93. Preston, A.R. *et al.* (2004) Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus* 14, 148–152
94. Dusek, J.A. and Eichenbaum, H. (1997) The hippocampus and memory for orderly stimulus relations. *Proc. Natl. Acad. Sci. U.S. A.* 94, 7109–7114
95. Shohamy, D. and Wagner, A.D. (2008) Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron* 60, 378–389
96. Zeithamova, D. *et al.* (2012) Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron* 75, 168–179
97. Milivojevic, B. *et al.* (2015) Insight reconfigures hippocampal-prefrontal memories. *Curr. Biol.* 25, 821–830
98. Schlichting, M.L. *et al.* (2015) Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat. Commun.* 6, 8151
99. Eichenbaum, H. *et al.* (1999) The hippocampus, memory, and place cells: is it spatial memory or a memory space? *Neuron* 23, 209–226
100. Howard, M.W. *et al.* (2005) The temporal context model in spatial navigation and relational learning: toward a common explanation of medial temporal lobe function across domains. *Psychol. Rev.* 112, 75–116
101. Kloosterman, F. *et al.* (2004) Two reentrant pathways in the hippocampal–entorhinal system. *Hippocampus* 14, 1026–1039
102. Eichenbaum, H. and Cohen, N.J. (2014) Can we reconcile the declarative memory and spatial navigation views on hippocampal function? *Neuron* 83, 764–770
103. Burgess, N. (2006) Computational models of the spatial and mnemonic functions of the hippocampus. In *The Hippocampus* (Andersen, P. *et al.*, eds), pp. 715–750, Oxford University Press
104. Willshaw, D.J. *et al.* (2015) Memory, modelling and Marr: a commentary on Marr (1971) 'Simple memory: a theory of archicortex'. *Philos. Trans. R. Soc. B Biol. Sci.* 370, 20140383
105. Schapiro, A.C. *et al.* (2014) The necessity of the medial temporal lobe for statistical learning. *J. Cogn. Neurosci.* 26, 1736–1747
106. Knowlton, B.J. and Squire, L.R. (1993) The learning of categories: parallel brain systems for item memory and category knowledge. *Science* 262, 1747–1749
107. Shohamy, D. and Turk-Browne, N.B. (2013) Mechanisms for widespread hippocampal involvement in cognition. *J. Exp. Psychol. Gen.* 142, 1159–1170

108. Nosofsky, R.M. (1984) Choice, similarity, and the context theory of classification. *J. Exp. Psychol. Learn. Mem. Cogn.* 10, 104–114
109. Tamminen, J. *et al.* (2015) From specific examples to general knowledge in language learning. *Cogn. Psychol.* 79, 1–39
110. Walker, M.P. and Stickgold, R. (2010) Overnight alchemy: sleep-dependent memory evolution. *Nat. Rev. Neurosci.* 11, 218
111. Wood, E.R. *et al.* (1999) The global record of memory in hippocampal neuronal activity. *Nature* 397, 613–616
112. Eichenbaum, H. (2014) Time cells in the hippocampus: a new dimension for mapping memories. *Nat. Rev. Neurosci.* 15, 732–744
113. McKenzie, S. *et al.* (2014) Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron* 83, 202–215
114. Quiroga, R.Q. *et al.* (2005) Invariant visual representation by single neurons in the human brain. *Nature* 435, 1102–1107
115. McClelland, J.L. (2013) Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *J. Exp. Psychol. Gen.* 142, 1190–1210
116. McClelland, J.L. and Goddard, N.H. (1996) Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus* 6, 654–665
117. Hinton, G.E. *et al.* (1986) Distributed representations. In *Explorations in the Microstructure of Cognition. Vol. 1: Foundations* (Rumelhart, D.E. *et al.*, eds), pp. 77–109, MIT Press
118. Krizhevsky, A. *et al.* (2012) Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 25, 1106–1114
119. Mnih, V. *et al.* (2015) Human-level control through deep reinforcement learning. *Nature* 518, 529–533
120. Alme, C.B. *et al.* (2014) Place cells in the hippocampus: eleven maps for eleven rooms. *Proc. Natl. Acad. Sci. U.S.A.* 111, 18428–18435
121. Samsonovich, A. and McNaughton, B.L. (1997) Path integration and cognitive mapping in a continuous attractor neural network model. *J. Neurosci.* 17, 5900–5920
122. Buzsaki, G. and Moser, E.I. (2013) Memory, navigation and theta rhythm in the hippocampal–entorhinal system. *Nat. Neurosci.* 16, 130–138
123. Renno-Costa, C. *et al.* (2014) A signature of attractor dynamics in the CA3 region of the hippocampus. *PLoS Comput. Biol.* 10, e1003641
124. Wills, T.J. *et al.* (2005) Attractor dynamics in the hippocampal representation of the local environment. *Science* 308, 873–876
125. Graves, A. *et al.* (2014) *Neural Turning machines.*
126. Sukhbaatar, S. *et al.* (2015) End-to-end memory networks. *NIPS* 2431–2439
127. J. Weston, *et al*. *Memory Networks*. Published online October 15, 2014 http://arxiv.org/abs/1410.3916
128. Scoville, W.B. and Milner, B. (1957) Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry* 20, 11–12
129. Nadel, L. and Moscovitch, M. (1997) Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* 7, 217–227
130. Moscovitch, M. *et al.* (2005) Functional neuroanatomy of remote episodic, semantic and spatial memory: a unified account based on multiple trace theory. *J. Anat.* 207, 35–66
131. Yassa, M.A. and Stark, C.E. (2011) Pattern separation in the hippocampus. *Trends Neurosci.* 34, 515–525
132. Liu, X. *et al.* (2012) Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature* 484, 381–385
133. Leutgeb, J.K. *et al.* (2007) Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science* 315, 961–966
134. Leutgeb, S. *et al.* (2004) Distinct ensemble codes in hippocampal areas CA3 and CA1. *Science* 305, 1295–1298
135. Bonnici, H.M. *et al.* (2011) Decoding representations of scenes in the medial temporal lobes. *Hippocampus* 22, 1143–1153
136. McHugh, T.J. *et al.* (2007) Dentate gyrus NMDA receptors mediate rapid pattern separation in the hippocampal network. *Science* 317, 94–99
137. Neunuebel, J.P. and Knierim, J.J. (2014) CA3 retrieves coherent representations from degraded input: direct evidence for CA3 pattern completion and dentate gyrus pattern separation. *Neuron* 81, 416–427
138. Nakazawa, K. *et al.* (2002) Requirement for hippocampal CA3 NMDA receptors in associative memory recall. *Science* 297, 211–218
139. Jezek, K. *et al.* (2011) Theta-paced flickering between place-cell maps in the hippocampus. *Nature* 478, 246–249
140. Richards, B.A. *et al.* (2014) Patterns across multiple memories are identified over time. *Nat. Neurosci.* 17, 981–986
141. Ketz, N. *et al.* (2013) Theta coordinated error-driven learning in the hippocampus. *PLoS Comput. Biol.* 9, e1003067
142. Kumaran, D. and Maguire, E.A. (2009) Novelty signals: a window into hippocampal information processing. *Trends Cogn. Sci.* 13, 47–54
143. Moser, E.I. and Moser, M.B. (2003) One-shot memory in hippocampal CA3 networks. *Neuron* 38, 147–148
144. Chaudhuri, R. and Fiete, I. (2016) Computational principles of memory. *Nat. Neurosci.* 19, 394–403
145. Lee, H. *et al.* (2015) Neural population evidence of functional heterogeneity along the CA3 transverse axis: pattern completion versus pattern separation. *Neuron* 87, 1093–1105
146. Lu, L. *et al.* (2015) Topography of place maps along the CA3-to-CA2 axis of the hippocampus. *Neuron* 87, 1078–1092
147. Collin, S.H. *et al.* (2015) Memory hierarchies map onto the hippocampal long axis in humans. *Nat. Neurosci.* 18, 1562–1564
148. Poppenk, J. *et al.* (2013) Long-axis specialization of the human hippocampus. *Trends Cogn. Sci.* 17, 230–240
149. Strange, B.A. *et al.* (2014) Functional organization of the hippocampal longitudinal axis. *Nat. Rev. Neurosci.* 15, 655–669
150. Ranganath, C. and Ritchey, M. (2012) Two cortical systems for memory-guided behaviour. *Nat. Rev. Neurosci.* 13, 713–726
151. Hasselmo, M.E. and Schnell, E. (1994) Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: computational modeling and brain slice physiology. *J. Neurosci.* 14, 3898–3914
152. Vazdarjanova, A. and Guzowski, J.F. (2004) Differences in hippocampal neuronal population responses to modifications of an environmental context: evidence for distinct, yet complementary, functions of CA3 and CA1 ensembles. *J. Neurosci.* 24, 6489–6496
153. Olshausen, B.A. and Field, D.J. (2004) Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14, 481–487
154. Quiroga, R.Q. *et al.* (2008) Sparse but not 'grandmother-cell' coding in the medial temporal lobe. *Trends Cogn. Sci.* 12, 87–91
155. Ahmed, O.J. and Mehta, M.R. (2009) The hippocampal rate code: anatomy, physiology and theory. *Trends Neurosci.* 32, 329–338
156. Tolhurst, D.J. *et al.* (2009) The sparseness of neuronal responses in ferret primary visual cortex. *J. Neurosci.* 29, 2355–2370
157. Vinje, W.E. and Gallant, J.L. (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276
158. Quirk, G.J. *et al.* (1992) The positional firing properties of medial entorhinal neurons: description and comparison with hippocampal place cells. *J. Neurosci.* 12, 1945–1963
159. Rust, N.C. and DiCarlo, J.J. (2012) Balanced increases in selectivity and tolerance produce constant sparseness along the ventral visual stream. *J. Neurosci.* 32, 10170–10182
160. Barlow, H.B. (1972) Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* 1, 371–394
161. Grossberg, S. (1987) Competitive learning: from interactive activation to adaptive resonance. *Cogn. Sci.* 11, 23–63
162. LaRocque, K.F. *et al.* (2013) Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. *J. Neurosci.* 33, 5466–5474
163. McClelland, J.L. and Rumelhart, D.E. (1981) An interactive activation model of context effects in letter perception. Part 1. An account of the basic findings. *Psychol. Rev.* 88, 375–407

164. Medin, D.L. and Schaffer, M.M. (1978) Context theory of classification. *Psychol. Rev.* 85, 207–238
165. Hintzman, D.L. (1986) 'Schema abstraction' in a multiple-trace memory model. *Psychol. Rev.* 93, 411–428
166. Suthana, N.A. *et al.* (2015) Specific responses of human hippocampal neurons are associated with better memory. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10503–10508
167. Wood, E.R. *et al.* (2000) Hippocampal neurons encode information about different types of memory episodes occurring in the same location. *Neuron* 27, 623–633
168. Ferbinteanu, J. and Shapiro, M.L. (2003) Prospective and retrospective memory coding in the hippocampus. *Neuron* 40, 1227–1239
169. Bower, M.R. *et al.* (2005) Sequential-context-dependent hippocampal activity is not necessary to learn sequences with repeated elements. *J. Neurosci.* 25, 1313–1323
170. MacDonald, C.J. *et al.* (2013) Distinct hippocampal time cell sequences represent odor memories in immobilized rats. *J. Neurosci.* 33, 14607–14616
171. Markus, E.J. *et al.* (1995) Interactions between location and task affect the spatial and directional firing of hippocampal neurons. *J. Neurosci.* 15, 7079–7094
172. Skaggs, W.E. and McNaughton, B.L. (1998) Spatial firing properties of hippocampal CA1 populations in an environment containing two visually identical regions. *J. Neurosci.* 18, 8455–8466
173. Kriegeskorte, N. *et al.* (2008) Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2, 4
174. Komorowski, R.W. *et al.* (2009) Robust conjunctive item-place coding by hippocampal neurons parallels learning what happens where. *J. Neurosci.* 29, 9918–9929
175. Ellenbogen, J.M. *et al.* (2007) Human relational memory requires time and sleep. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7723–7728
176. Dumay, N. and Gaskell, M.G. (2007) Sleep-associated changes in the mental representation of spoken words. *Psychol. Sci.* 18, 35–39
177. Coutanche, M.N. and Thompson-Schill, S.L. (2014) Fast mapping rapidly integrates information into existing memory networks. *J Exp Psychol Gen.* 143, 2296–2303
178. Sharon, T. *et al.* (2011) Rapid neocortical acquisition of long-term arbitrary associations independent of the hippocampus. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1146–1151
179. Merhav, M. *et al.* (2014) Neocortical catastrophic interference in healthy and amnesic adults: a paradoxical matter of time. *Hippocampus* 24, 1653–1662
180. Smith, C.N. *et al.* (2014) Comparison of explicit and incidental learning strategies in memory-impaired patients. *Proc. Natl. Acad. Sci. U.S.A.* 111, 475–479
181. Warren, D.E. and Duff, M.C. (2014) Not so fast: hippocampal amnesia slows word learning despite successful fast mapping. *Hippocampus* 24, 920–933
182. Greve, A. *et al.* (2014) No evidence that 'fast-mapping' benefits novel learning in healthy older adults. *Neuropsychologia* 60, 52–59
183. Schaul, T. *et al.* (2016) Prioritized experience replay. In *International Conference on Learning Representations*.
184. Gallistel, C.R. (1990) *The Organization of Learning,* MIT Press
185. Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. *Neural Comput.* 9, 1735–1780
186. Santoro, A. *et al.* (2016) Meta-Learning with memory augmented neural networks. In *International Conference in Machine Learning*.
187. Treves, A. and Rolls, E.T. (1994) Computational analysis of the role of the hippocampus in memory. *Hippocampus* 4, 374–391