CellPress

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.

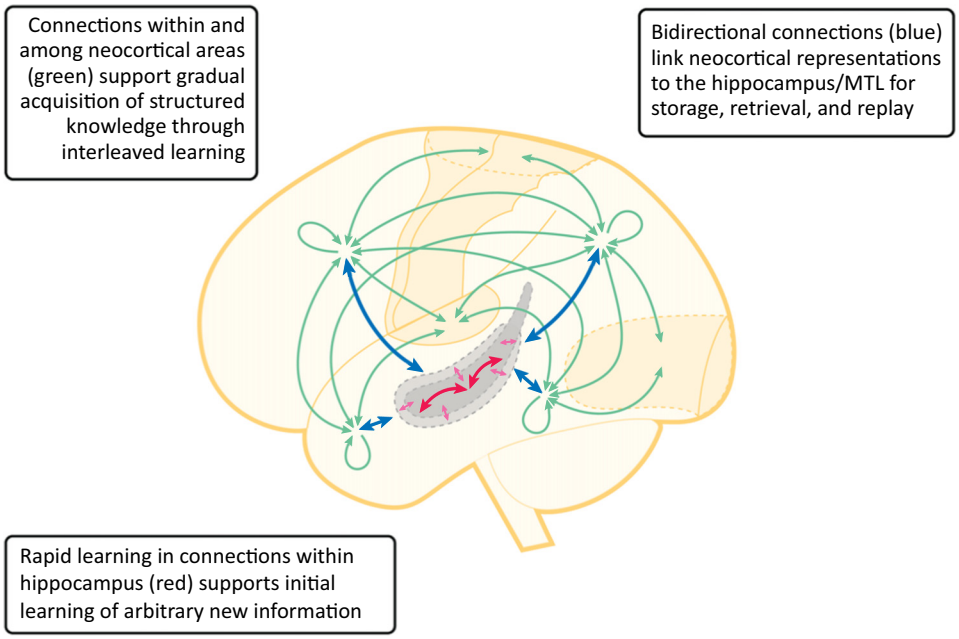

Figure 1. Lateral view of one hemisphere of the brain, where broken lines indicate regions deep inside the brain or on the medial surface. Primary sensory and motor cortices are shown in darker yellow. Medial temporal lobe (MTL) surrounded by broken lines, with hippocampus in dark grey and surrounding 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.