Mechanisms of memory: an intermediate level of analysis and organization
Rosemary A Cowell and David E Huber

Research in the last five years has made great strides toward mechanistic explanations of how the brain enables memory. This progress builds upon decades of research from two complementary strands: a Levels of Analysis approach and a Levels of Organization approach. We review how research in cognitive psychology and cognitive neuroscience under these two approaches has recently converged on mechanistic, brain-based theories, couched at the optimal level for explaining cognitive phenomena — the intermediate level. Furthermore, novel empirical and data analysis techniques are now providing ways to test these theories’ predictions, a crucial step in unraveling the mechanisms of memory.

Address
Department of Psychological and Brain Sciences, University of Massachusetts Amherst, United States

Corresponding author: Cowell, Rosemary A (rcowell@umass.edu)

Introduction: levels of analysis versus levels of organization
The goal of life sciences is usually to provide mechanistic explanations of phenomena [1,2]. Bechtel and Abrahamsen [3] define a mechanism as “a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena”. In cognitive neuroscience, such a phenomenon might be “sequential recall effects in human memory”. One challenge in providing such explanations is choosing the appropriate level of analysis and level of biological organization (or scale) at which to construct the theory.

David Marr [4] identified three levels of analysis for understanding information-processing tasks such as vision (Figure 1). The highest, computational level defines the task goals; the middle, representational or algorithmic level, supplies the mechanistic steps by which the task is achieved; and the lowest, implementational level describes the physical realization of the task in biological tissue. The three levels are almost independent: a computational goal may be achieved by multiple algorithms and an algorithm can be implemented in many physical hardwares. According to Marr, an information-processing task should be analyzed by first considering the behavioral goals of the system, because the manner in which a system solves a task will depend more on nature of the problem to be solved than on the hardware that solves it. Thus, a levels of analysis approach is inherently top-down, allowing the researcher to freely invent algorithms and representations to explain behavioral phenomena (in an analogy from physics, the hypothesized representation for the phenomenon of galaxy rotation is dark matter, even though the subatomic particles of dark matter are completely unspecified).

Other researchers have operated under a ‘levels of organization’ approach (Figure 2), in which the problem space for theorizing is defined in terms of levels of biological scale (e.g., synapses, neural assemblies or anatomical systems). In contrast to Marr’s levels of analysis, these levels are tightly coupled; indeed, the goal of theories under this approach is often to define the relation between structures and phenomena at vertically adjacent levels [5,6]. This approach is at least partly reductionist: a phenomenon at one level is explained in terms of structures at the level below (e.g., the behavior of a neural assembly depends on neural coupling via synapses). Although researchers can look both ‘up and down’ to higher and lower levels for inspiration [6,7], they are always constrained by implementational details. Under this approach, a significant challenge is to determine the appropriate ‘intermediate’ sized components: if too small, interactions between component parts cannot account for the phenomenon of interest (e.g., it is difficult to explain the subjective experience of recollection in terms of interacting molecules), and if too large, the component operations become a restatement of the target phenomenon (e.g., a subregion of prefrontal cortex is ‘responsible for retrieval’) [6].
Marr’s Levels of Analysis. There are three levels as which the explanation of a task can reside. The highest level, termed computational, defines the overall goals of the task. The middle level, called representational or algorithmic, supplies the “how” in an explanation of a visual or cognitive phenomenon — the critical, mechanical steps by which the task is carried out. The lowest level, the implementational level, describes the physical realization of the task in the biological tissue of the nervous system. The three levels are only loosely connected — they may influence but do not tightly constrain each other. In particular, understanding the algorithmic/representational level can proceed without specifying the implementational level.

abstract explanations at the computational/algorithmic levels without heed to implementation. Many influential memory theories in this tradition have been formalized with mathematics. One of the earliest theories of learning is Estes’ stimulus sampling theory (SST) [8]. SST hypothesizes a large set of possible stimulus-response ‘elements’, which are statistically sampled in an all-or-none manner. Following on the heels of behaviorism, which disavowed unobservable theoretical constructs, this theory signaled a new approach by postulating both the form of mnemonic representations and the operations that act upon them.

As in SST, Atkinson and Shiffrin [9] appealed to the statistical sampling of latent memory traces, but their theory laid out how a memory moves from sensory memory, to short-term memory, and then into or out of long-term memory (via encoding and retrieval). These different ‘modes’ of memory were defined in terms of content (e.g., visual, verbal, semantic), duration (e.g., less than a second, up to 30 s, a lifetime), and capacity (e.g., a whole visual display, approximately 7 chunks, unlimited). Subsequent work with the Search of Associative Memory (SAM) model specified the mathematical algorithms of statistical sampling and recovery in memory recall [10], and global matching in recognition [11]. Concurrent with the algorithmic developments in the SAM model, other memory researchers formalized memory representations by assuming that every event is specified as a vector of features, with vectors concatenated in a giant repository [12] or convolved with the existing memory structure to form a composite distributed representation [e.g., 13].

Further important progress was made in defining the representations and algorithms of temporal context. Early work with the SAM model [14] proposed that context randomly fluctuates over time, explaining interference and forgetting effects. To explain recall dynamics (e.g., how participants use each recalled item to cue subsequent recall attempts), Howard and Kahana [15] proposed the Temporal Context Model (TCM), in which context fluctuations are not random, but strongly influenced by the current item. This was a radical departure from Atkinson and Shiffrin’s modal model, explaining short-term memory phenomena without reference to a separate short-term memory system. The TCM model was subsequently given feature representations and learning rules compatible with neural network theory, producing the Context Maintenance and Retrieval (CMR) model [16]. Rather than assuming arbitrary features, CMR uses real-world representations extracted using Latent Semantic Analysis (LSA) [17] — a statistical corpus-based analysis of the co-occurrences of words within paragraphs. Attesting to the applicability of this genre of latent representations, such representations are the driving force behind Google and other text-based search engines.

Thus, cognitive models of memory under a levels of analysis approach have arrived at increasingly realistic representations and algorithms. However, they have largely avoided the question of compatibility with biological brains.

Searching for the structures of memory
In contrast to the abstract models of cognitive psychology, cognitive neuroscience research in the same period attempted to identify the brain structures of memory. This more data-driven, levels of organization approach was kick-started by the case of patient H.M. [18,19]. H.M.’s deficits following brain surgery suggested that the medial temporal lobe (MTL) performs a highly selective function — the formation of new, long-term, declarative memories. Moreover, the preservation of H.M.’s other skills implied that the MTL makes no contribution to other functions (e.g., perceptual, linguistic, motoric). Subsequent work demonstrated that declarative memory is further divisible into subtypes such as episodic and semantic [20,21], inspiring the ‘multiple memory systems’ framework in which
Different forms of memory are independent and rely upon distinct brain regions [20,22–24].

Although the multiple memory systems account provides a decomposition of memory by carving it into subtypes based on phenomenology, and mapping each subtype onto intermediate-level structures in the brain, it skirts the question of how each subtype arises. That is, it fails to describe the mechanisms that produce the observed memory phenomena [6,25,26,27]. This exemplifies Bechtel’s claim [6] that the search for cognitive mechanisms has been hindered by a failure to correctly identify the intermediate level: multiple memory systems accounts identify the structural parts but not the intermediate-level operations that act upon them.

Later, memory systems accounts were challenged (or complemented) by components of processing accounts, which decompose memory according to the type of information processing required, such as top-down (conceptual) versus bottom-up (perceptual) processing [28–31]. Bechtel [6] argued that components of processing accounts better dissect the mechanisms of memory, because rather than providing only a descriptive, phenomenal decomposition, they emphasize constituent operations, which are needed to define a working mechanism.

However, even components of processing accounts have provided scant details regarding the specific operations that underpin memory [6,27]. Furthermore, the utility of a component process account is limited by how appropriately it decomposes a phenomenon [27]. For example, one theory decomposes retrieval into recollection and familiarity processes, and maps them onto hippocampal versus neocortical systems [32–35]. We have argued [27,36] that recollection and familiarity are inadequate as components of a mechanistic theory because to identify these processes as memory operations is to make what Gilbert Ryle identified as a ‘category mistake’ [37]: they are subtypes of memory retrieval, not subcomponents. Recollection and familiarity are the high-level phenomena to be explained, rather than intermediate-level operations of memory, so applying them to component structures falls short of a mechanistic explanation. In our view, while 20th century cognitive neuroscience enjoyed success in identifying the brain structures of memory, it failed to identify the operations of memory.

**Mechanism at an intermediate level: structures, operations and representations**

As briefly summarized above, the levels of analysis approach elucidated mnemonic representations and operations in the abstract but did not tie them to the brain, whereas the levels of organization approach delineated the brain structures underpinning memory but did not specify its neural operations, nor its representations. The convergence of these approaches was inevitable, and fruitful.
McClelland et al.’s Complementary Learning Systems (CLS) account [38] was an early example of convergence. Pairing neuropsychological facts (e.g., that damage to the hippocampus affects some but not all kinds of memory) with information-processing principles of learning and representation, this work and subsequent extensions [39] supplied a mechanism for many long-term declarative memory phenomena. It proposed candidate operations for encoding and retrieval — for example, pattern separation, pattern completion, ‘sharpening’ of stimulus representations — and tied them to neuroarchitectural properties. Moreover, specifying the operations required explication of the representations: in pattern completion, the activity pattern representing each stimulus must be sparse so that a partial input does not trigger completion of an incorrect representation owing to excessive overlap [40–42]. In fact, it is difficult to say whether the operations of CLS are simply the best means by which to arrive at representations with the desired properties (i.e., representations are the central explanatory factor), or rather the properties of CLS representations are an inevitable consequence of the operations (i.e., operations are the crux of the account). The point is that we need both operations and representations for a mechanistic explanation.

Recognizing the importance of representations led to another promising approach. Echoing earlier accounts that assigned ‘memory systems’ or ‘processes’ to brain regions, this new framework instead assigns ‘representations’. This approach analyses vision and memory in information processing terms, asking what representations are needed for each task (a levels of analysis philosophy), but exploiting the fact that neurons can be characterized in terms of their receptive fields, that is, what they represent (a data-driven premise). For example, the Representational-Hierarchical account assumes that successive regions in the ventral visual stream and MTL capture increasingly complex conjunctions of features, culminating in objects in perirhinal cortex and associations of items with context, time and space in the hippocampus. Critically, the contribution of a brain region to cognition is said to be determined by what information it represents rather than what process it computes [36,43–45, see also Ref. 46]. This view is now well evidenced [47*,48–52]. Although this will surely prove an imperfect account of memory, especially for other levels of biological scale (e.g., within subregions of the hippocampus, functional distinctions may be best characterized by operations), it improves on multiple memory systems and process-based accounts because the representations that it postulates are intermediate-level components.

What have we learned about the intermediate-level mechanisms of memory?

Thus, cognitive psychology and cognitive neuroscience have converged on a detailed sketch of the brain structures, operations and representations of memory. Recent technical advances have given us new methods for probing neural representations and functional networks. Armed with these methods, what advances in mechanistic understanding have been built on that sketch?

Many researchers have risen to Marr’s challenge of determining the representations underlying cognition by exploiting new tools such as multi-variate pattern analysis (MVPA; [53]) and representational similarity analysis [54]. That is, memory theories developed by cognitive psychologists at an abstract level — with representational structures that were freely invented — are now being tested in the brain. For example, fMRI studies have supported and refined the temporal context theory [15], by revealing information-bearing states that reflect temporal context (e.g., [55–58]). In another example, hierarchical schemas in memory, well-established in the abstract [59,60], have been fleshed out in neural terms by electrophysiological studies in rodents employing novel analysis techniques [61]. Another abstract construct proposed to be important for the structure and segmentation of episodic memory is the event boundary [62]. Modern brain imaging has confirmed this proposal and helped unpack its mechanisms, for example by demonstrating that the stability of hippocampal representations across events is related to memory for temporal order [63,64]. Brunec et al. [65] recently suggested that boundary segmentation mechanisms are shared across spatial and episodic domains, integrating two hippocampal functions that have often been studied separately.

A second, more data-driven approach within neuroscience has fine-tuned the intermediate-level structures and operations of memory. For example, recent work has uncovered functional networks rather than focal, contiguous brain regions as the structural components of memory [66–68]. Using new connectivity analyses, previously unknown structural components such as the Default Mode Network (DMN) have been revealed [69] and the DMN has been linked to episodic memory via possible roles in memory consolidation [70] and autobiographical reminiscence [71]. Further, the operations of memory are undergoing significant revision. By applying classifiers to intracranial electrophysiological data collected in humans performing an episodic recall task, Weidemann et al. [72] arrived at a novel proposition: that semantic and episodic information interact primarily through retrieval operations, not through the similarity of representations laid down at encoding.

Other breakthroughs have stemmed from combining theory-driven and data-driven approaches. Zhang et al. [73] used RSA to identify stimulus-specific activity patterns in humans performing a memory task, then tracked the replay of these representations during sleep. Replay triggered by hippocampal ‘ripples’ during non-REM sleep predicted later memory. Thus, guided by theory (that consolidation depends on replay) and exploiting modern
tools for mining complex data (RSA), this study shed light on the representations and operations of memory consolidation. In a similarly top-down/bottom-up approach, Howard Eichenbaum, Mark Howard and colleagues have combined new empirical evidence for ‘time cells’ in the hippocampus [74,75] with existing theories of hippocampal function, to generate new theories that integrate temporal, spatial and mnemonic processing in the hippocampus [76–78].

In light of these advances, some are advocating a new approach to cognition more broadly [27**,47*,79**,80,81]. Under this approach, research into brain function puts the computational and representational capabilities of a brain region (or network) first, and cognitive function second. Rather than carving cognition into folk-psychologically intuitive functions and mapping these onto brain structures, this approach first asks what a brain structure computes or represents, then considers how it can contribute to cognition [27**,36,79**,82]. A mechanism in one brain region for performing some function can be recapitulated in another region to serve another function (e.g., ‘pattern completion’ may underpin recollection in hippocampus and priming in visual cortex [27**,47]). Conversely, a brain region may serve more than one cognitive function: the perirhinal cortex contributes to recognition memory and visual discrimination [48,83] and the hippocampus is functionally versatile because its associative representations are suited to many cognitive purposes and its computations are flexible [79**,84,85].

Conclusions

Twentieth century memory research delineated key memory phenomena, sketched their representations and algorithms in the abstract, and identified the brain structures that support them. Subsequent convergence of abstract models with neuropsychological accounts has produced mechanistic, brain-based explanations. These gains were achieved by choosing intermediate-level components while adopting a levels of analysis philosophy that asks how a brain region’s representations and operations might serve the goals of a cognitive task. The intermediate level is optimal for explanations in cognitive neuroscience because its components — structures like the hippocampus, representational properties such as ‘associative/relnational’, and operations like pattern completion — lie immediately below the level of phenomena to be explained — for example, recall effects and their relation to regional brain activation or focal lesions. Better still, being couched in function-neutral terms of operations and representations, these accounts are not limited to explaining phenomena that count as ‘memory’. Instead, they build bridges between memory, attention, perception, conceptual knowledge, decision-making, and more. Modern neuroscience tools, along with the new intermediate-level currency of operations and representations, are revising the textbook view of brain function by drawing parallels between phenomena that were previously consigned to separate chapters.

Conflict of interest statement

Nothing declared.

CRediT authorship contribution statement

Rosemary A Cowell: Conceptualization, Writing - original draft, Writing - review & editing, Funding acquisition. David E Huber: Writing - review & editing, Funding acquisition.

Acknowledgements

This work was funded by NSF award BCS-1554871 to R.A.C. and by NIMH award 1RF1MH14277 to R.A.C. and D.E.H.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


27. Cowell RA, Barense MD, Sazil PS: A roadmap for understanding memory: decomposing cognitive processes into operations and representations. eNeuro 2019, 6

This review argues that cognitive processes — introspectively identifiable ‘mental events’ like recollection — are inadequate constructs for characterizing neural-cognitive mechanisms, because they confine lower-level components of the mechanisms that we seek to understand. Recollection involves both an operation (pattern completion) and a neural representation (high-dimensional, associative content). To uncover the mechanisms of recollection, we must decompose it into its operations and representations.


47. Ross DA et al.: Hippocampal engagement during recall • depends on memory content. Cereb Cortex 2018, 28:2685-2698

This study demonstrated that representational content, rather than memonic process, determines whether a brain region contributes to memory retrieval. It tested part-cued (i.e., pattern completion-like) retrieval of complex, associative memories (i.e., recollection) and simple, single-item memories (i.e., non-associative retrieval or ‘object recollection’). The hippocampus, traditionally assumed critical for recollection, was engaged in recall only for complex stimuli, and not for objects; object recall engaged lower-level regions, perirhinal and lateral occipital cortex.


75. Mau W et al.: The same hippocampal CA1 population simultaneously codes temporal information over multiple timescales. Curr Biol 2018, 28:1499-1508.e4
This study reported ensembles of “time cells” in hippocampus that simultaneously carry information about temporal sequence order on timescales of seconds, minutes and days, providing compelling empirical evidence for mechanisms proposed in a computational model of hippocampal function [Ref. [75]].
This guest editorial introduces a special issue discussing new perspectives on hippocampal function, in which the various articles outline a diversity of functional roles for the hippocampus, including perception, working memory, attention and implicit memory. Together, the arguments laid out by these authors usher in a new philosophy on how to approach brain function.
This theoretical synthesis suggests that the hippocampus transforms episodic experiences into conceptual knowledge by deploying many different cognitive operations including attentional biasing, pattern completion-based retrieval, memory integration, pattern separation, and the generation of prediction error.