CellPress

Hierarchical process memory: memory as an integral component of information processing

Uri Hasson[1], Janice Chen[1], and Christopher J. Honey[2]

[1] Department of Psychology and the Neuroscience Institute, Princeton University, NJ 08544-1010, USA
[2] Department of Psychology, University of Toronto, Toronto ON, M5S 3G3, Canada

Models of working memory (WM) commonly focus on how information is encoded into and retrieved from storage at specific moments. However, in the majority of real-life processes, past information is used continuously to process incoming information across multiple timescales. Considering single-unit, electrocorticography, and functional imaging data, we argue that (i) virtually all cortical circuits can accumulate information over time, and (ii) the timescales of accumulation vary hierarchically, from early sensory areas with short processing timescales (10s to 100s of milliseconds) to higher-order areas with long processing timescales (many seconds to minutes). In this hierarchical systems perspective, memory is not restricted to a few localized stores, but is intrinsic to information processing that unfolds throughout the brain on multiple timescales.

> 'The present contains nothing more than the past, and what is found in the effect was already in the cause.' (Henri L. Bergson)

Memory as a component of all neural processes

In real life, multiple timescales of prior information continuously influence the processing of information in the present. Consider, for example, how prior information shapes language comprehension: each phoneme achieves its meaning in the context of a word, each word in the context of a sentence, and each sentence in the context of a discourse. Thus, past information gathered over milliseconds-, seconds-, and minutes-long timescales all contribute to comprehension. More generally, memories of recent events continuously support the processing of incoming information.

Working memory as a specialized memory store

When information from the recent past is needed for task performance, it is conventionally described as being stored in WM [1,2]. Theories of WM traditionally focus on memory stores: how information enters and leaves them, their capacity, and the robustness of stored information to interference and decay. The separation between information storage and information processing is rooted in analogies with digital computer architectures, where the systems that perform information processing (e.g., CPUs) are separated from the memory systems that store information (e.g., RAM, caches, and hard-disks). Thus, in computer-inspired models of memory, new information is temporarily stored in limited-capacity WM buffers, or old information is made available for present processing when it is loaded into the buffers from long-term memory (LTM) storage (Figure 1A). In such models, the systems of memory storage (WM and LTM) are functionally distinct (and in some cases physically separated, Figure 1B) from the systems that support online information-processing tasks, such as visual and auditory object recognition, biological motion perception, perceptual decision making, and the organization of movement [3].

The innovation provided by the multi-store model (Figure 1) was in specifying how a general purpose WM resource was instantiated via a control system (central executive) operating on a set of functionally specialized buffers (the phonological loop, visuospatial sketchpad, and episodic buffers); this clarified how WM might relate to task performance in wide-ranging task domains. However, as researchers came to consider the number of WM subsystems that would be required to support memory for different kinds of information over multiple timescales, these subdivisions of WM began to appear inadequate [4]. Moreover, the neural circuits identified as WM buffers (e.g., the phonological loop) appear in many cases to be the same as the neural circuits that perform the relevant processing (e.g., of phonological and linguistic information) [5].

Newer perspectives on WM no longer require a physical separation between memory storage and ongoing information processing, but they maintain a functional separation between stores and processing [4,6]. For example, in contemporary theories of visual WM, the visual memory representations are located in the visual processing stream. Nonetheless, the representations in visual WM are functionally separated from new visual input, as top-down fronto-parietal signals are required to shield the contents of WM from interference. Thus, information is still considered to be stored in and retrieved from WM, which has a distinct functional status from the representations of incoming information.

Corresponding author: Hasson, U. (hasson@princeton.edu).

1364-6613/
 http://dx.doi.org/10.1016/j.tics.2015.04.006

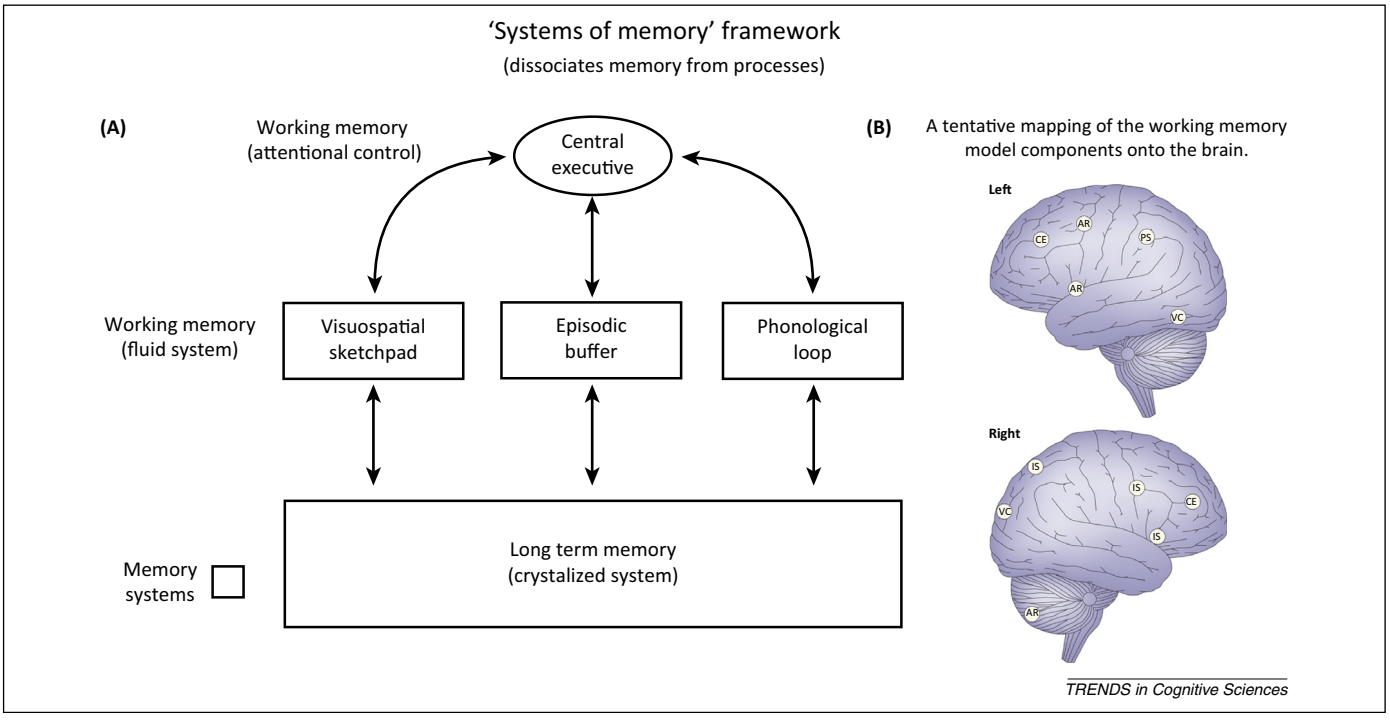

Figure 1. **(A)** In a 'systems of memory' framework, the information storage (boxes) is functionally separated from the information-processing units (arrows). **(B)** A tentative mapping of the working memory (WM) model components onto the brain. Figure adapted from [1], reprinted with permission from Nature Publishing Group.

Although WM models effectively capture behavioral and neural data related to goal-directed control of prior information (e.g., maintaining a visual array over a delay period), we propose that they can also conceptually obscure our understanding of ubiquitous real-life processes in which memory has to be integrated with ongoing processing. We argue that memory and online processing are entangled in many everyday situations, such as reading a book or conversing with a friend. Thus, we suggest replacing the question: 'how is information stored and then retrieved for later processing?' with the question: 'how does prior information continuously shape processing in the present moment?'

Questioning the separation between memory and ongoing processing

The separation of the contents of memory from ongoing information processing has long been questioned. A prominent example is the network memory model [7], which embraced the idea that there are no dedicated memory systems and that memory is an integral part of all neural networks, arguing for a shift in research focus from 'systems of memory' to the 'memory of systems'. The network memory model also incorporated a hierarchical organization of the systemic memory, so that higher-order areas could combine and abstract memories accumulated from lower areas in the hierarchy. In a similar spirit, the theory of active memory questioned the separation between short-term memory and LTM [8]. In this framework, memory is considered as a single entity, either active or inactive, with active memories envisaged as a subset of especially labile memories that are currently being used by the brain to process incoming information [9].

More recently, the separation between memory units and processing units has been questioned by neuroscientists who propose local interactions between memory and perception within visual areas [10–12] and beyond [13], as well as by some LTM researchers who argue for the involvement of the medial temporal lobe (MTL) memory system in perceptual processes [14–17].

Neurobiologically, the idea of functional separation of memory into processes and stores is not well substantiated. In neural circuits, there does not appear to be a separation between neurons that process information and neurons that store information. For example, in the sea slug *Aplysia californica* there are no systems specialized for memory that are separated from perceptual and motor systems [18]; in contrast, short-term and long-term changes in the synaptic efficacy of sensory and motor neurons support learning and memory. More generally, in mammalian circuits it is known that patterns of prior information reshape synapses over minutes and years [19–21], and can alter levels of activation, potentiation, and excitability over milliseconds and minutes [22–24]. Thus, at the biological level, prior information continually influences information processing in the present, and memory is intrinsic to virtually all neural processes.

Psychologists have also critiqued the segregation of memory and ongoing processing. An early step in this direction came in linking memory performance to the hierarchical depth of prior processing of a stimulus [25,26]. Following the rise of connectionist models [27], even more radical theories arose, proposing that the memory needed for task performance (e.g., in language comprehension) is an intrinsic property of distributed circuits that continuously process the language input [28]. In such models, individual differences in language comprehension are ascribed not to differences in capacity, but to differences in linguistic expertise (i.e., the organization of information processing within the language processing

circuits). Another prominent model that emphasized the importance of expertise for explaining mnemonic performance was the Long Term Working Memory model [29], which aimed to account for variations in memory as a function of domain expertise and stimulus semantics.

Finally, computational modeling work has shown in a variety of ways how memory and information processing can be combined in the same circuits [30]. Abstract connectionist models provided some of the earliest and most influential examples [31], while more recent dynamical systems models of information processing have also posited a diversity of integration timescales, enabling each level of a hierarchical system to integrate information over different timescales [32–34].

Process memory framework

Synthesizing these prior ideas with recent empirical work from neurophysiology and neuroimaging, we now outline a framework for how memory serves online information processing. In this framework, we emphasize that traces of past information should not be segregated from ongoing neural processes. To dissociate our notion of memory from the traditional notion of encapsulated memory stores, we will use the term 'process memory' throughout the paper. We use the term process memory, in a broad sense, to mean active traces of past information that are used by a neural circuit to process incoming information in the present moment. Furthermore, we argue for a hierarchical organization of process memory, in which the timescale of memory-dependent processing gradually increases from early sensory areas to high-order areas. The new framework is broadly consistent with the family of distributed memory models described above [7,8,10].

In the process memory framework, virtually all cortical circuits have the ability to accumulate information over time (Figure 2, red circles). We operationalize the proces-

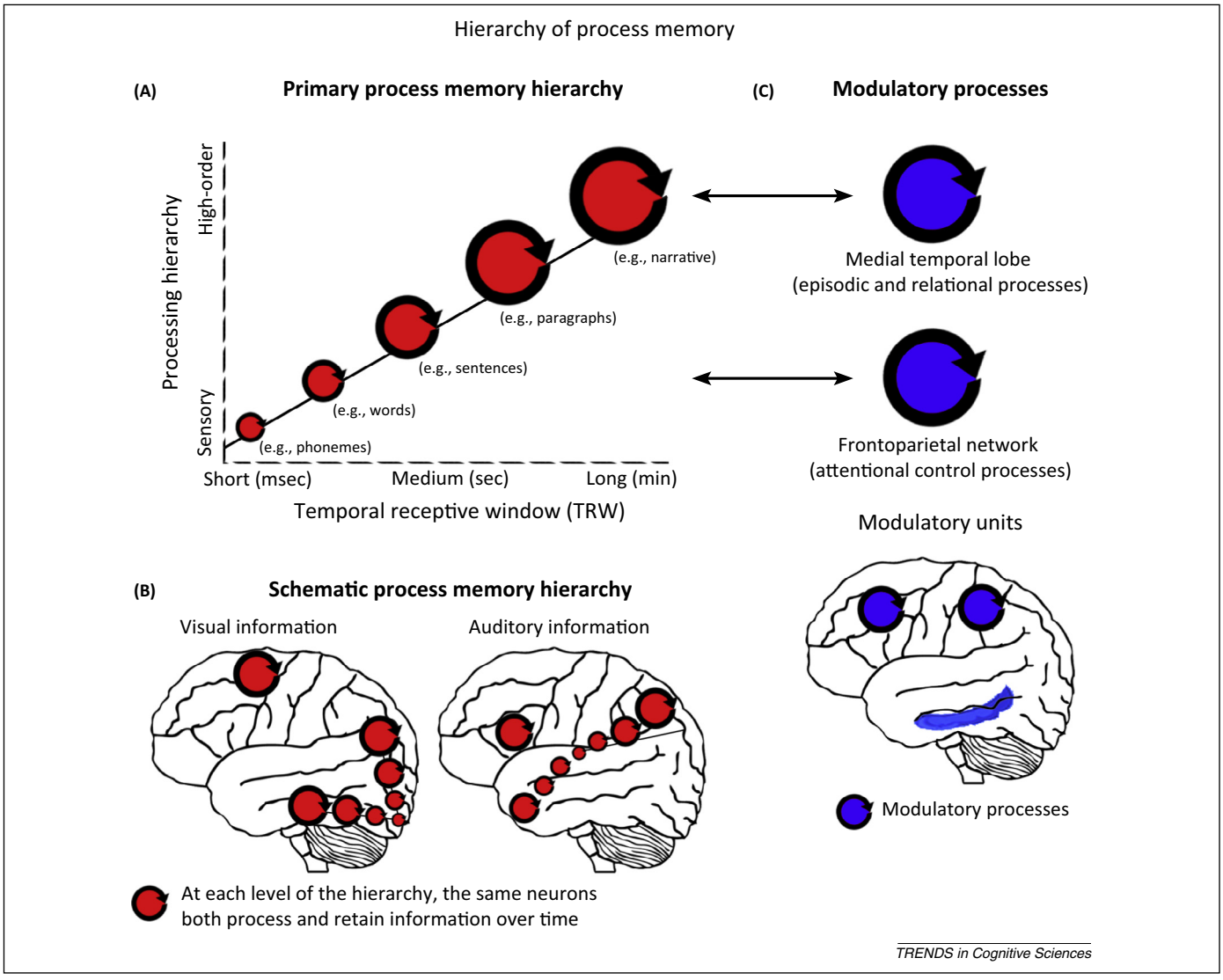

Figure 2. **(A)** A hierarchy of process memory framework. Memory is integral to the operation of each cortical area and there is no separation between the processing units and information storage units. Furthermore, the processing timescale (operationalized by measuring the temporal receptive window [TRW]) in each region increases in a topographically organized manner, from milliseconds in early sensory areas up to minutes in high-order areas. **(B)** A schematic process memory hierarchy for auditory and visual stimulation (for actual data see Figures 3 and 4 and [39]). **(C)** Primary versus modulatory process memory. Two additional processes (blue circles) modulate the primary process memories (red circles) that are located along the hierarchy: attentional control processes (e.g., fronto-parietal network interactions with short-TRW linguistic regions could enable maintenance of a target word across a delay period) and episodic memory processes (e.g., MTL/hippocampal interactions, most likely with long-TRW regions such as retrosplenial cortex, could enable reactivation of an autobiographical episode).

sing timescale of each brain region by measuring its temporal receptive window (TRW): the window of time in which prior information from an ongoing stimulus can affect the processing of newly arriving information. The TRW is defined as a temporal analog of the spatial receptive field. Some areas have a short TRW (e.g., 10s to 100s of milliseconds), enabling them to integrate a few phonemes to detect a word. Other areas have a medium TRW (e.g., several seconds), enabling the integration of sequences of words while parsing a sentence. Still other areas have a long TRW (e.g., 10s to 100s of seconds) that is necessary for the integration of sentences over time while comprehending a narrative.

The TRW increases in an orderly hierarchical fashion from early sensory areas to higher-order perceptual and cognitive areas (Figure 2, size of red circles). Memories of the recent past are not stored in a few dedicated memory stores, but are organized hierarchically across cortical regions that process incoming information. We will now outline evidence for the process memory model, before returning to discuss the relationship between process memory and existing models of WM and LTM.

A method for mapping processing timescales

To characterize the TRW for each area of the cerebral cortex, we measured the extent to which traces of prior events (recent memory) influenced moment-to-moment neural activity (online processes) during minutes-long real-life stimuli (such as stories and movies, Box 1). First, minutes-long stimuli were broken into smaller temporal units (e.g., paragraphs, sentences, words) and the order of the units was scrambled, thereby varying the temporal structure of the stimulus at multiple timescales while preserving the atomic elements (using the identical movie frames or elementary sound clips in all conditions). Next, neural activity during the intact and scrambled stimuli was examined for evidence of whether online responses changed as a function of the structure of prior events. Areas with short processing timescales (i.e., short TRWs) were expected to respond in the same way at any given moment regardless of the prior context. Areas with long processing timescales (i.e., long TRWs) were expected to modulate their responses to a given event as a function of prior context over many seconds; for example, responses to a particular word would be affected by information from a previous paragraph.

Neural response dynamics during these temporally extended real-life stimuli were assessed using inter-subject correlation [35]. The response reliability of any given brain region was measured by calculating the correlation of that region's timecourse across multiple subjects during exposure to the same stimulus. High correlations between

Box 1. Dissociation between memory and process is not feasible in real-life contexts

In the systems of memory perspective, memory is segregated from the neural systems that process sensory input and is kept in dedicated WM and LTM stores (see Figure 1 in main text). Based on such conceptualization, many memory studies focus on delay periods in which information has to be actively maintained but not processed (e.g., match-to-sample tasks, Figure IA), and in which the integration of past with present information is undesirable (e.g., remembering a target word in a list of distractors). On the behavioral level, dual-task and filled-delay studies have taught us that WM capacity for arbitrary items is limited (usually to 4 +/−1 items [81,82]) and that remembered content during a filled-delay period is fragile, labile, and susceptible to interference [3]. On the physiological level, these tasks have revealed sustained and selective responses during the delay period in prefrontal cortex and lateral inferior parietal cortex [83,84]. These findings are of lasting value for explaining how people maintain and manipulate information according to rules using attentional control [57]; however, they do not necessarily provide evidence for segregation of processing and memory storage systems. Actively holding information in mind during a delay period seems to rely on attentional mechanisms, not on dedicated WM buffers [85,86], and the evidence suggests that the same brain areas that perform primary processing are also involved in the maintenance of information across delays [4,6,87].

Contemporary views of WM [4,6] are compatible with the process memory framework we propose (see Figure 2 in main text), but there is a key difference in emphasis. In contrast to WM studies that focus on delay periods, in which selected information must be segregated from new input, we highlight the active and ongoing accumulation and integration of information that occurs during online processing. To that end, we explored the kind of memory that is crucially required during continuous natural stimulation, where prior information must be integrated with (rather than segregated from) new input. For example, when listening to a spoken sentence (Figure IB), the brain must concurrently detect acoustic features and integrate them with prior sounds to recognize words, while at the same time integrating each word with the preceding elements in the sentence. By examining neural processing of naturalistic temporally extended stimuli, we underline the importance of an under-studied question in memory research: how are memories of past information, gathered across multiple timescales, used by the brain to continuously process incoming information?

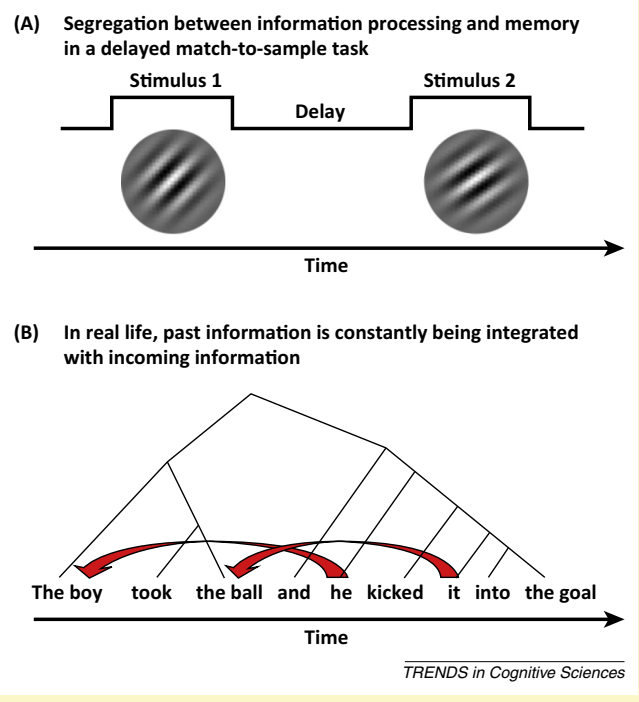

Figure I. **(A)** Example of a delayed match-to-sample task in which the to-be-remembered information is separated in time from the subsequent cue that triggers the use of the information. **(B)** Example of an everyday sentence in which there is a need to integrate each incoming word with the preceding words as the sentence unfolds over time. Red arrows connect a pronoun and its referent; black arrows indicate the need to integrate phrases within and across sentences.

subjects at any given location in the brain indicated the presence of stimulus-driven reliable responses at that location. For more details about the inter-subject correlation method see [35,36].

A hierarchical topography of process memory

Mapping temporal receptive windows using fMRI, electrocorticography (ECoG), and single-unit recording has revealed a large-scale topographic organization of processing timescales along the auditory and visual processing streams [37–39]. Figure 3A presents the gradual transition along the superior temporal gyrus, from short TRWs in early auditory cortex to long TRWs in the temporoparietal junction and angular gyrus, as measured using fMRI while subjects listened to a story scrambled over multiple timescales. Early auditory areas (A1+) responded reliably (i.e., high inter-subject correlation) at all scrambling levels, from the intact full story (FS), to scrambled paragraphs (P), scrambled sentences (S), scrambled words (W), and backward speech (B). These sensory regions were denoted as having short process memory (short TRWs). Further up the processing hierarchy, more and more of the stimulus history was found to affect processing in the present moment. In areas with especially long process memory (long TRWs), such as the temporoparietal junction (TPJ), angular gyrus (AG), and medial prefrontal cortex (mPFC), the cortical activity at each moment depended on information that had previously arrived over 10s of seconds. In these higher-order areas the responses were reliable only at the full story (FS) and paragraphs (P) levels [38]. The process memory hierarchy was not confined to the processing of temporally extended linguistic input, as an analogous topographic organization was found in the visual system when subjects viewed silent movies [39]. Finally, a similar topographic gradient of TRWs was observed in the visual and auditory processing streams for an audio-visual movie using ECoG (Figure 3B), which replicated the fMRI findings with a direct neurophysiological measurement [37].

Slow neural dynamics are more pronounced in areas with long TRWs

In the primate brain, the TRW of an area (i.e., its processing timescale) covaries with the timescale of its intrinsic neural dynamics. In other words, intrinsically faster neural dynamics are observed in areas with shorter TRWs, whereas intrinsically slower neural dynamics are observed in areas with longer TRWs. Recently, the spike-count autocorrelation during short resting periods was measured in seven cortical areas in the macaque monkey, revealing a hierarchical ordering in which sensory and prefrontal areas exhibited shorter and longer timescales, respectively [40] (Figure 4; see also [101]). Similarly, using ECoG [37] it was observed that neuronal population activity in higher-order regions exhibited a greater proportion of low-frequency fluctuations (and increased temporal autocorrelation), while in early sensory areas there was a greater proportion of high-frequency fluctuations (and decreased temporal autocorrelation). In both studies, the gradient of timescales of neural dynamics was observed in the absence of any stimulus, suggesting that it may be an intrinsic property of neural circuits. Similar dynamical organizations have been reported using fMRI [41].

Together, these results suggest that areas with faster neural dynamics accumulate information over shorter timescales, whereas areas with slower neural dynamics accumulate information over longer timescales. Thus, the abundant slow fluctuations of neural dynamics [42], which are commonly ignored or treated as entirely artifactual, can actually be connected to the processing of real-life information, which is structured on timescales of milliseconds, seconds, and minutes [37,41,43]. Recent modeling has connected the hierarchical organization of dynamical timescales to changes in excitatory–inhibitory balance (via changes in spine density) [44]; other models have proposed a role for large-scale anatomical organizations [45], and the topography of neuromodulators also appears likely to play a role.

Scaling of TRW size as a function of information rate

The results presented so far suggest that TRWs increase gradually from sensory areas with short TRWs up to higher-order areas with long TRWs. The next study asked whether the size of the TRW should be defined in fixed temporal units (e.g., milliseconds, seconds, and minutes) or informational units (e.g., phonemes, words, and sentences). Fortunately, temporal units and informational units are easily dissociated in real-life speech. The fastest speakers of American English will articulate a sentence about twice as fast as the slowest speakers [46]. If temporal integration windows are defined based on informational units, then neural response timecourses should be rescaled in time when the incoming information is rescaled in time. In accordance with such a prediction, a temporal scaling of neural responses was observed throughout auditory, linguistic, and extra-linguistic brain areas in response to a linear scaling (speeding or slowing) of the incoming speech rate [47]. These data suggest that the process memory integration window should be measured on relative rather than absolute timescales, with the brain scaling its neural dynamics in response to compression or dilation of the input.

The temporal rescaling of neural responses in accordance with rescaled stimuli began to break down when stimuli were presented at double speed (50% duration stimulus); this is also when intelligibility began to be impaired [47]. Furthermore, it is important to emphasize that, even when the timescales of processing were rescaled, the TRW hierarchy was preserved. Thus, just as spatial receptive fields can spatially rescale as a function of task, context, and attention demands [48–51], process memory integration windows can temporally rescale according to the rate at which information is arriving [52,53].

Linking process memory with other types of memory

We have argued here for the parsimonious idea that process memory is an integral feature of many cortical processing systems. Moreover, we proposed that process memory increases in a hierarchical manner across the cerebral cortex. How does the process memory framework relate to the classic distinctions [54] between subtypes of

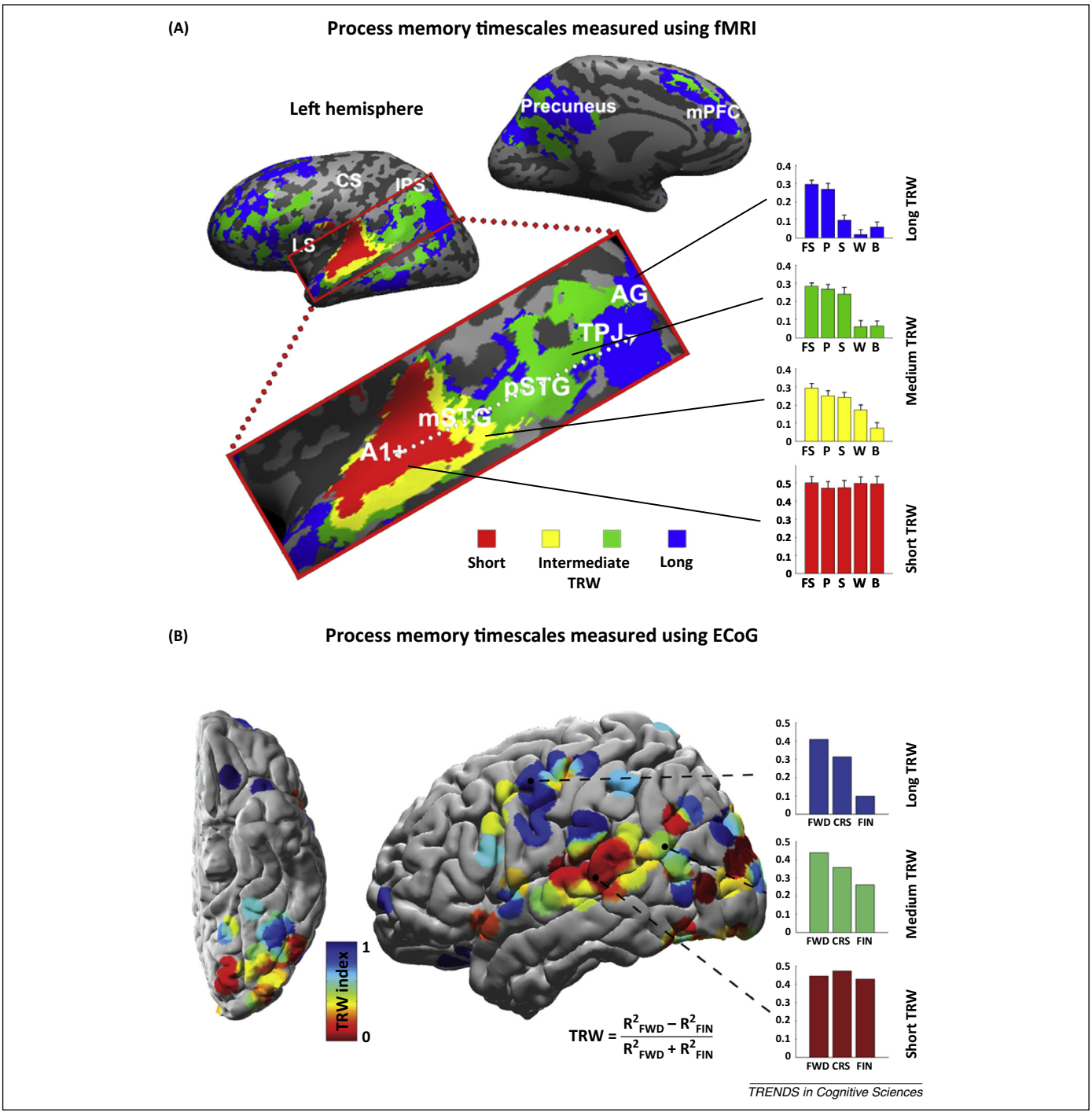

Figure 3. Hierarchical topography of temporal receptive windows (TRWs). **(A)** fMRI map of the gradual transition from short to long TRWs along the temporal–parietal axis mapped using audio narratives. The color of each voxel indicates the shortest timescale of coherence in the stimulus that produced a reliable inter-subject response (red: story played backward; yellow: story with word-order scrambled; green: story with sentence-order scrambled; blue: story with paragraph-order scrambled). (A, inset) BOLD (blood oxygen level-dependent fMRI) timecourses in early auditory areas (A1+) were reliable across subjects exposed to the same stimulus; this was true at all scrambling levels, from the intact full story (FS), to scrambled paragraphs (P), scrambled sentences (S), scrambled words (W), and backward speech (B). Further up the processing hierarchy, more and more of the stimulus history affected responses in the present moment. At the top of the hierarchy, areas such as the temporal parietal junction (TPJ) responded reliably only at the full story and paragraph levels. Figure adapted from [38]. **(B)** Electrocorticography (ECoG) map of the gradual transition from short to long TRWs mapped using an audiovisual movie. The TRW indices computed within five individual subjects are displayed on a standard surface. Shorter TRWs were predominantly found near primary sensory areas, while longer TRWs were found further away from sensory areas. The TRW index was defined as the difference in response reliability between the intact and scrambled stimuli, as a proportion of the sum of their respective reliabilities. (B, inset) Early auditory areas (A1+) responded reliably across all scrambling levels, from the intact full movie (FWD), to the coarse scrambled movie (CRS), and fine scrambled movie (FIN). Further up the processing hierarchy, more and more of the stimulus history affected responses in the present moment. At the top of the hierarchy, areas such as the lateral prefrontal cortex responded with much greater reliability at the intact movie and coarse scrambled movie levels. Figure adapted from [37].

memories, both declarative (semantic and episodic) and non-declarative (procedural, priming, and conditioning)?

The frameworks are compatible, but answer different questions. Traditional memory categories are defined based on the types of stimuli remembered (e.g., visual or auditory), the type of learning (e.g., one-shot or repeated exposures), and the types of behavior that the memory supports (e.g., recollection or recognition). The process

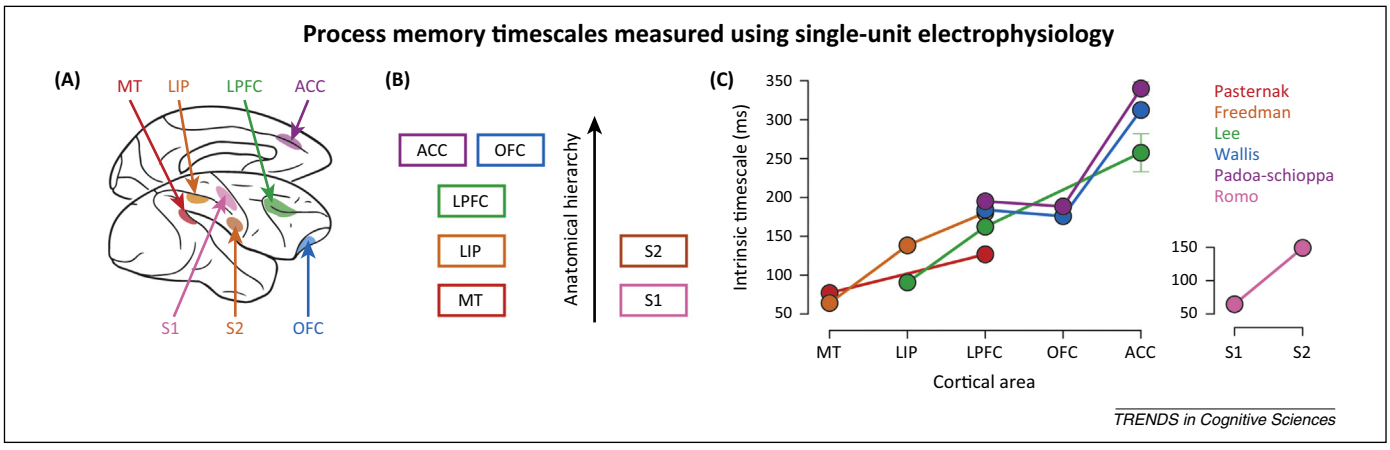

Figure 4. Spike-count autocorrelation reveals a hierarchical ordering of intrinsic timescales. **(A)** Across multiple experiments, single-unit data were recorded from seven cortical areas in the macaque monkey: MT, LIP, LPFC, OFC, ACC, S1 and S2. **(B)** Areas arranged according to their anatomical hierarchy as defined by their long-range projection patterns. **(C)** Intrinsic timescales increase gradually along the visual-prefrontal hierarchy. Error bar indicates standard error of fit parameters. Figure adapted from [40], reprinted with permission from Nature Publishing Group. Abbreviations: ACC, anterior cingulate cortex; LPFC, lateral prefrontal cortex; LIP, lateral intraparietal cortex; MT, Middle temporal visual cortex; OFC, orbitofrontal cortex; S1, S2, Primary and secondary somatosensory cortex (anterior and lateral parietal cortex).

memory framework is concerned with active memory that is intrinsic to the information processing taking place within a circuit. Thus, there may be similar amounts of process memory, supported by similar circuit mechanisms, even across circuits with very different functions. For example, premotor circuits may contain some of the procedurally learned memories for steering a bicycle, whereas visual cortex may contain the memories that later support recognition of a dynamic emotional facial expression. Nevertheless, both organizing a steering movement and recognizing a dynamic facial expression may require integrating information over a second of time. Thus, the types of processes performed in the circuit will define the functional properties of the stored information, but even very different processes may have a common functional signature and some common mechanisms for integrating information. Our model is also compatible with alternative processing-based parcellations of memory, such as that proposed in [55].

In the process-memory framework, the same cortical neurons that process information (e.g., in sensory perception) also store the information. Nonetheless, additional processes are needed to manipulate, control, and consolidate these process memories in specific contexts. In particular, there are two major modulatory processes that act on the primary process memories: attentional control processes supported by fronto-parietal circuits (related to traditional WM), and binding and consolidation processes supported by MTL circuits (related to episodic memory) [102] (Figure 2C).

Process memory and the attentional perspective on WM

In many contemporary perspectives, the term 'working memory' has become almost synonymous with attention and cognitive control [56]. The deep connection between WM and attention arises because, in classic WM paradigms such as digit manipulation and delayed-match-to-sample, subjects are asked to actively preserve some aspects of the incoming information and task goals in mind (Box 1). When the to-be-remembered content is fragile and labile, attention must be used to shield prior information from interference with new information [3,57]. Shielding against processing of new input while actively maintaining information in memory is a special case of the more common situations discussed above, in which memory and processing of incoming information are intertwined (Box 1).

Although attentional control is a fundamental aspect of cognition, and we incorporate it as a key modulator in our model (Figure 2C), its role in ongoing perception and comprehension may be obscured by the label 'working memory'. The work of memory is performed in virtually every neural circuit, and attentional systems modulate this ongoing processing in accordance with rule- or goal-related constraints. Thus, both the designated memory buffers framework and the attentional-control framework of WM fail to account for the tight reliance of online neural processes on memory across multiple levels of the processing hierarchy. One perspective on top-down modulation that is more consistent with our distributed model is provided by hierarchical predictive coding models (Box 2).

Process memory and the MTL

Our framework suggests that processing timescales increase gradually along the cortical hierarchy, from early sensory areas with short (milliseconds-long) integration windows up to higher-order areas with minutes-long integration windows (Figure 2A,B). The responses in areas with the longest processing timescales, at the apex of the hierarchy, seem to be influenced by information accumulated over many minutes. These areas overlap broadly with the default mode network (DMN), and include the angular gyrus, retrosplenial cortex, precuneus, posterior cingulate cortex, and mPFC. Many studies implicate these areas in the encoding and retrieval of episodic memories (memory for situated experiences tied to a particular time and place [58,59]), as well as in a variety of high-level cognitive processes such as decision making [60–62], self-representation [63–65], prospective planning [63], and social reasoning [65].

The DMN is functionally and anatomically connected to the MTL and hippocampus [66,67], but it remains unclear to what extent the hippocampus is needed for the

Box 2. Outstanding questions

- Which biophysical circuit models can integrate memory with online processing? Network recurrence is a powerful mechanism for incorporating memory into online processing, but it remains challenging to model how information accumulated over many seconds of time can modify online processing in a local neural circuit [88–90].
- What is the role of the MTL system in the accumulation and manipulation of information over minutes of time? Areas with long TRWs, at the apex of the hierarchy, seem to be able to accumulate information over minutes. These areas are functionally connected to the MTL system. Will MTL lesions diminish the ability of cortical areas with long TRWs to accumulate information over many minutes?
- Is information transmitted upward along the process memory hierarchy in a continuous or pulsatile manner? For example, does an area with a 'sentence' timescale transmit information continuously to an area with a 'paragraph' timescale, or is the communication primarily at the end of the sentence?
- Does process memory require top-down feedback within the hierarchy? Computing predictions further into the future is made easier with more information about the past; thus process memory and predictive coding may be deeply intertwined. Are predictive signals transmitted in a top-down manner along the hierarchy, from areas with longer timescales to areas with shorter timescales?
- How is process memory affected by the allocation of attention? In contrast to two-state, multi-state, and continuous state models of WM [91–93], the process memory framework does not assign a central role to attention. Because attention is capacity-limited, it is unclear how it can support temporal integration simultaneously across diverse circuits and hierarchical levels, as required in many daily-life situations (Box 1). Nonetheless, attention certainly modulates process memory (see Figure 2C in main text), and the nature of this modulation requires further investigation.
- How can we quantify the influence of prior information? Information theoretic frameworks may enable us to precisely measure the information that the history of a circuit contains about its future, across a range of timescales [94]. Synergy metrics [95–97] can quantify the additional information about the present state that is provided by joint knowledge of past and present input.
- What is the relationship between process memory and information integration during decision making? Neurons in lateral parietal and frontal areas can accumulate time-varying evidence for or against choice alternatives [98–100], demonstrating history-dependent processing in a local circuit. This history-dependence is usually simple because the circuit response depends only on the present input and the present circuit state; future studies should investigate information integration in tasks that require more complex history-dependent processing.

DMN to retain information over many minutes. Hippocampal damage strikingly impacts the ability to retain episodic memories [68], but exactly how long new information can be maintained in cortical areas without hippocampal involvement is not clear. For example, hippocampal amnesics can retain stimulus information for long enough to engage in a conversation [68], summarize a short passage of prose [69,70], and play a complex communicative game [71]. Such observations suggest that a meaningful continuous context (e.g., a conversation or listening to a story) may enable information to persist in cortical areas for a few minutes without relying on the hippocampus (though hippocampal circuits do appear to contribute to ongoing processing when they are intact [72–74]). Furthermore, the important role of the hippocampus in encoding and retrieval of episodic memories over long timescales [75] does not exclude its involvement in binding of relational information over short timescales [72,76–78]. Measurements of neural responses to scrambled narratives in MTL lesion patients will provide important constraints on how these regions modulate process memory.

Links between process memory and active LTM

Stored knowledge is as crucial for online processing of incoming information as is recently acquired information. After all, understanding the meaning of a word, sentence, or idea relies heavily on information gathered throughout the lifetime of the individual. Because online processing, at each level of the hierarchy, relies both on recent memories (e.g., information accumulated during an ongoing conversation) and distant memories (e.g., long-term knowledge), both types of memory must be integrated within each neural circuit. Thus, the argument against dedicated WM stores can naturally be extended to question the notion of dedicated LTM stores that are separate from the circuits that process information [79].

Inspired by prior researchers [8], we suggest that memory be conceived as a single entity that can be either in an active (process memory) state or an inactive (LTM) state. Process memory refers to prior information that is currently used by an area to process incoming information; to influence ongoing processing, the prior information must be in an active state. The active information is composed of the stimulus information accumulated in a given circuit throughout the event, as well as a subset of LTMs activated during the processing of the incoming information. Note that the word 'active' is not necessarily synonymous with sustained elevation in firing rates [80], given that information can be sustained in a neuronal circuit by short-term calcium-mediated synaptic facilitation in the absence of recurrent activity [22]. Inactive LTMs, by contrast, are simply the long-lasting structural features of a circuit (e.g., synaptic patterning) that are not currently affecting the processing of incoming information in that circuit. Inactive LTMs may be brought into an active state via the influence of other active process memories in the circuit, or via modulation from MTL input or fronto-parietal control areas (Box 2).

Concluding remarks

In this article we have argued that the traditional dissociation between memory and ongoing information processing is artificial. Diverse cortical functions, ranging from the smooth pursuit of a swiftly-moving object, to resolving an anaphoric reference, up to the integration of information across multiple paragraphs of text, all require active integration of past information with new information. We have reviewed data from multiple sources indicating that timescales of processing vary in a hierarchical fashion across the cerebral cortex, with shorter (milliseconds to seconds) timescales in sensory regions and a gradient of lengthening timescales (seconds to minutes) in higher-order cortices. Instead

of compartmentalizing memory into increasingly specialized storage systems, we highlight the fact that memory is an integral component of the processing conducted in each neural circuit. Process memory is an especially important factor in real-world cognition and perception, which requires continuous information integration, not only maintenance over delays. We propose the process memory framework as a biologically motivated model of the memory that is intrinsic to ongoing, integrative information processing in naturalistic settings.

Acknowledgments

U.H. and J.C were supported by a National Institute of Mental Health award (R01-MH094480). C.J.H. was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2014-04465). For insightful comments and contributions to the ideas in this manuscript, we thank Mariam Aly, Bradley R. Buchsbaum, Alin I. Coman, Jarrod A. Lewis-Peacock, Kenneth A. Norman, Rosanna K. Olsen, and Jordan Poppenk.

References

1 Baddeley, A. (2003) Working memory: looking back and looking forward. *Nat. Rev. Neurosci.* 4, 829–839
2 Cowan, N. (2008) What are the differences between long-term, short-term, and working memory? *Prog. Brain Res.* 169, 323–338
3 Baddeley, A. (2012) Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63, 1–29
4 Postle, B.R. (2006) Working memory as an emergent property of the mind and brain. *Neuroscience* 139, 23–38
5 Buchsbaum, B.R. and D'Esposito, M. (2008) The search for the phonological store: from loop to convolution. *J. Cogn. Neurosci.* 20, 762–778
6 Sreenivasan, K.K. *et al.* (2014) Revisiting the role of persistent neural activity during working memory. *Trends Cogn. Sci.* 18, 82–89
7 Fuster, J.M. (1997) Network memory. *Trends Neurosci.* 20, 451–459
8 Lewis, D.J. (1979) Psychobiology of active and inactive memory. *Psychol. Bull.* 86, 1054–1083
9 Nader, K. and Einarsson, E.O. (2010) Memory reconsolidation: an update. *Ann. N. Y. Acad. Sci.* 1191, 27–41
10 Albright, T.D. (2012) On the perception of probable things: neural substrates of associative memory, imagery, and perception. *Neuron* 74, 227–245
11 Gavornik, J.P. and Bear, M.F. (2014) Learned spatiotemporal sequence recognition and prediction in primary visual cortex. *Nat. Neurosci.* 17, 732–737
12 Wandell, B.A. *et al.* (2002) Common principles of image acquisition systems and biological vision. *Proc. IEEE* 90, 5–17
13 Harris, J.A. *et al.* (2001) The cortical distribution of sensory memories. *Neuron* 30, 315–318
14 Murray, E.A. *et al.* (2007) Visual perception and memory: a new view of medial temporal lobe function in primates and rodents. *Annu. Rev. Neurosci.* 30, 99–122
15 Bussey, T.J. and Saksida, L.M. (2005) Object memory and perception in the medial temporal lobe: an alternative approach. *Curr. Opin. Neurobiol.* 15, 730–737
16 Graham, K.S. *et al.* (2010) Going beyond LTM in the MTL: a synthesis of neuropsychological and neuroimaging findings on the role of the medial temporal lobe in memory and perception. *Neuropsychologia* 48, 831–853
17 Barense, M.D. *et al.* (2012) Intact memory for irrelevant information impairs perception in amnesia. *Neuron* 75, 157–167
18 Squire, L.R. and Kandel, E.R. (2009) *Memory: from Mind to Molecules*, Roberts & Co
19 Hebb, D.O. (1955) *The Organization of Behavior; A Neuropsychological Theory*, Wiley
20 Fahle, M. *et al.* (2002) *Perceptual Learning*, MIT Press
21 Marom, S. (2010) Neural timescales or lack thereof. *Prog. Neurobiol.* 90, 16–28
22 Mongillo, G. *et al.* (2008) Synaptic theory of working memory. *Science* 319, 1543–1546
23 Perrett, D.I. *et al.* (2009) Seeing the future: natural image sequences produce 'anticipatory' neuronal activity and bias perceptual report. *Q. J. Exp. Psychol.* 62, 2081–2104
24 Miller, E.K. (1994) Neocortical memory traces. *Behav. Brain Sci.* 17, 488–489
25 Craik, F.I. (2002) Levels of processing: past, present. and future? *Memory* 10, 305–318
26 Lockhart, R.S. (2002) Levels of processing, transfer-appropriate processing, and the concept of robust encoding. *Memory* 10, 397–403
27 Feldman, J.A. and Ballard, D.H. (1982) Connectionist models and their properties. *Cogn. Sci.* 6, 205–254
28 MacDonald, M.C. and Christiansen, M.H. (2002) Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychol. Rev.* 109, 35–54
29 Ericsson, K.A. and Kintsch, W. (1995) Long-term working memory. *Psychol. Rev.* 102, 211–245
30 Buonomano, D.V. and Maass, W. (2009) State-dependent computations: spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci.* 10, 113–125
31 McClelland, J.L. *et al.* (2010) Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends Cogn. Sci.* 14, 348–356
32 Kiebel, S. *et al.* (2008) A hierarchy of time-scales and the brain. *PLoS Comput. Biol.* 4, e1000209
33 Rabinovich, M. *et al.* (2008) Transient dynamics for neural processing. *Science* 321, 48–50
34 Yamashita, Y. and Tani, J. (2013) Self-organized functional hierarchy through multiple timescales: neuro-dynamical accounts for behavioral compositionality. In *Computational and Robotic Models of the Hierarchical Organization of Behavior* (Baldassarre, G. and Mirolli, M., eds), pp. 47–62, Springer
35 Hasson, U. *et al.* (2004) Intersubject synchronization of cortical activity during natural vision. *Science* 303, 1634–1640
36 Hasson, U. *et al.* (2010) Reliability of cortical activity during natural stimulation. *Trends Cogn. Sci.* 14, 40–48
37 Honey and Christopher, J. *et al.* (2012) Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* 76, 423–434
38 Lerner, Y. *et al.* (2011) Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* 31, 2906–2915
39 Hasson, U. *et al.* (2008) A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* 28, 2539–2550
40 Murray, J.D. *et al.* (2014) A hierarchy of intrinsic timescales across primate cortex. *Nat. Neurosci.* 17, 1661–1663
41 Stephens, G.J. *et al.* (2013) A place for time: the spatiotemporal structure of neural dynamics during natural audition. *J. Neurophysiol.* 110, 2019–2026
42 Leopold, D.A. *et al.* (2003) Very slow activity fluctuations in monkey visual cortex: implications for functional brain imaging. *Cereb. Cortex* 13, 422–433
43 He, B.J. (2014) Scale-free brain activity: past, present, and future. *Trends Cogn. Sci.* 18, 480–487
44 Chaudhuri, R. *et al.* (2015) A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Biorxiv* 017137
45 Baria, A.T. *et al.* (2013) Linking human brain local activity fluctuations to structural and functional network architectures. *Neuroimage* 73, 144–155
46 Smith, B.L. (2002) Effects of speaking rate on temporal patterns of English. *Phonetica* 59, 232–244
47 Lerner, Y. *et al.* (2014) Temporal scaling of neural responses to compressed and dilated natural speech. *J. Neurophysiol.* 111, 2433–2444
48 Sheinberg, D.L. and Logothetis, N.K. (2001) Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *J. Neurosci.* 21, 1340–1350
49 Li, W. *et al.* (2006) Contour saliency in primary visual cortex. *Neuron* 50, 951–962
50 Op De Beeck, H. and Vogels, R. (2000) Spatial sensitivity of macaque inferior temporal neurons. *J. Comp. Neurol.* 426, 505–518

51 Zipser, K. *et al.* (1996) Contextual modulation in primary visual cortex. *J. Neurosci.* 16, 7376–7389
52 Gütig, R. and Sompolinsky, H. (2009) Time-warp-invariant neuronal processing. *PLoS Biol.* 7, e1000141
53 Tank, D.W. and Hopfield, J.J. (1987) Neural computation by concentrating information in time. *Proc. Natl. Acad. Sci. U.S.A.* 84, 1896–1900
54 Squire, L.R. (1992) Declarative and nondeclarative memory: multiple brain systems supporting learning and memory. *J. Cogn. Neurosci.* 4, 232–243
55 Henke, K. (2010) A model for memory systems based on processing modes rather than consciousness. *Nat. Rev. Neurosci.* 11, 523–532
56 Cowan, N. *et al.* (2005) On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. *Cogn. Psychol.* 51, 42–100
57 Engle, R.W. *et al.* (1999) Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *J. Exp. Psychol. Gen.* 128, 309–331
58 Kim, H. (2010) Dissociating the roles of the default-mode, dorsal, and ventral networks in episodic memory retrieval. *Neuroimage* 50, 1648–1657
59 Rugg, M.D. and Vilberg, K.L. (2013) Brain networks underlying episodic memory retrieval. *Curr. Opin. Neurobiol.* 23, 255–260
60 King-Casas, B. *et al.* (2005) Getting to know you: reputation and trust in a two-person economic exchange. *Science* 308, 78–83
61 Montague, P.R. *et al.* (2002) Hyperscanning: simultaneous fMRI during linked social interactions. *Neuroimage* 16, 1159–1164
62 Tomlin, D. *et al.* (2006) Agent-specific responses in the cingulate cortex during economic exchanges. *Science* 312, 1047–1050
63 Buckner, R.L. and Carroll, D.C. (2007) Self-projection and the brain. *Trends Cogn. Sci.* 11, 49–57
64 Goldberg, I.I. *et al.* (2006) When the brain loses its self: prefrontal inactivation during sensorimotor processing. *Neuron* 50, 329–339
65 Mitchell, J.P. (2009) Social psychology as a natural kind. *Trends Cogn. Sci.* 13, 246–251
66 Vincent, J.L. *et al.* (2008) Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. *J. Neurophysiol.* 100, 3328–3342
67 Aggleton, J.P. (2012) Multiple anatomical systems embedded within the primate medial temporal lobe: implications for hippocampal function. *Neurosci. Biobehav. Rev.* 36, 1579–1596
68 Scoville, W.B. and Milner, B. (1957) Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry* 20, 11–21
69 Holdstock, J.S. *et al.* (1995) The performance of amnesic subjects on tests of delayed matching-to-sample and delayed matching-to-position. *Neuropsychologia* 33, 1583–1596
70 Baddeley, A. and Wilson, B.A. (2002) Prose recall and amnesia: implications for the structure of working memory. *Neuropsychologia* 40, 1737–1743
71 Duff, M.C. *et al.* (2006) Development of shared information in communication despite hippocampal amnesia. *Nat. Neurosci.* 9, 140–146
72 Hannula, D.E. *et al.* (2006) The long and the short of it: relational memory impairments in amnesia, even at short lags. *J. Neurosci.* 26, 8352–8359
73 Hannula, D.E. *et al.* (2007) Rapid onset relational memory effects are evident in eye movement behavior, but not in hippocampal amnesia. *J. Cogn. Neurosci.* 19, 1690–1705
74 Olsen, R.K. *et al.* (2012) The hippocampus supports multiple cognitive processes through relational binding and comparison. *Front. Hum. Neurosci.* 6, 146
75 Squire, L.R. and Zola, S.M. (1998) Episodic memory, semantic memory, and amnesia. *Hippocampus* 8, 205–211
76 Hannula, D.E. *et al.* (2015) Memory for items and relationships among items embedded in realistic scenes: disproportionate relational memory impairments in amnesia. *Neuropsychology* 29, 126–138
77 Pertzov, Y. *et al.* (2013) Binding deficits in memory following medial temporal lobe damage in patients with voltage-gated potassium channel complex antibody-associated limbic encephalitis. *Brain* 136, 2474–2485
78 Olson, I.R. *et al.* (2006) Working memory for conjunctions relies on the medial temporal lobe. *J. Neurosci.* 26, 4596–4601
79 Ranganath, C. and Blumenfeld, R.S. (2005) Doubts about double dissociations between short- and long-term memory. *Trends Cogn. Sci.* 9, 374–380
80 Brunel, N. (2003) Dynamics and plasticity of stimulus-selective persistent activity in cortical network models. *Cereb. Cortex* 13, 1151–1161
81 Cowan, N. (2001) The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* 24, 87–114
82 Miller, G. (1956) The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 81–97
83 Romo, R. and Salinas, E. (2003) Flutter discrimination: neural codes, perception, memory and decision making. *Nat. Rev. Neurosci.* 4, 203–218
84 Chafee, M.V. and Goldman-Rakic, P.S. (1998) Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memory task. *J. Neurophysiol.* 79, 2919–2940
85 McCabe, D.P. *et al.* (2010) The relationship between working memory capacity and executive functioning: evidence for a common executive attention construct. *Neuropsychology* 24, 222–243
86 Conway, A.R. *et al.* (2005) Working memory span tasks: a methodological review and user's guide. *Psychon. Bull. Rev.* 12, 769–786
87 Harrison, S.A. and Tong, F. (2009) Decoding reveals the contents of visual working memory in early visual areas. *Nature* 458, 632–635
88 Sussillo, D. and Abbott, L.F. (2009) Generating coherent patterns of activity from chaotic neural networks. *Neuron* 63, 544–557
89 Klampfl, S. and Maass, W. (2013) Emergence of dynamic memory traces in cortical microcircuit models through STDP. *J. Neurosci.* 33, 11515–11529
90 Hoerzer, G.M. *et al.* (2014) Emergence of complex computational structures from chaotic neural networks through reward-modulated Hebbian learning. *Cereb. Cortex* 24, 677–690
91 Anderson, J.R. (1983) *The Architecture of Cognition*, Harvard University Press
92 Oberauer, K. (2002) Access to information in working memory: exploring the focus of attention. *J. Exp. Psychol. Learn. Mem. Cogn.* 28, 411–421
93 Larocque, J.J. *et al.* (2014) Multiple neural states of representation in short-term memory? It's a matter of attention. *Front. Hum. Neurosci.* 8, 5
94 Palmer, S. *et al.* (2013) Predictive information in a sensory population. *arXiv* 1307.0225
95 Gawne, T.J. and Richmond, B.J. (1993) How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.* 13, 2758–2771
96 Schneidman, E. *et al.* (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440, 1007–1012
97 Schneidman, E. *et al.* (2003) Synergy, redundancy, and independence in population codes. *J. Neurosci.* 23, 11539–11553
98 Gold, J.I. and Shadlen, M.N. (2007) The neural basis of decision making. *Annu. Rev. Neurosci.* 30, 535–574
99 Huk, A. and Shadlen, M. (2005) Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *J. Neurosci.* 25, 10420–10436
100 Shadlen, M. and Newsome, W. (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiol.* 86, 1916–1936
101 Ogawa, T. and Komatsu, H. (2010) Differential temporal storage capacity in the baseline activity of neurons in macaque frontal eye field and area V4. *J. Neurophysiol.* 103, 2433–2445
102 Shohamy, D. and Turk-Browne, N.B. (2013) Mechanisms for widespread hippocampal involvement in cognition. *J. Exp. Psychol. Gen.* 142, 1159–1170