Neuron
Perspective

Neurocognitive Architecture of Working Memory

Johan Eriksson,[1,2,*] Edward K. Vogel,[3] Anders Lansner,[4,5] Fredrik Bergström,[1,2] and Lars Nyberg[1,2,6]
[1]Department of Integrative Medical Biology, Umeå University, 901 87 Umeå, Sweden
[2]Umeå Center for Function Brain Imaging (UFBI), Umeå University, 901 87 Umeå, Sweden
[3]Department of Psychology, Institute for Mind and Biology, University of Chicago, Chicago, IL 60637, USA
[4]Department of Computational Biology, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden
[5]Department of Numerical Analysis and Computer Science, Stockholm University, 106 91 Stockholm, Sweden
[6]Department of Radiation Sciences, Umeå University, 901 87 Umeå, Sweden
*Correspondence: johan.eriksson@umu.se
http://dx.doi.org/10.1016/j.neuron.2015.09.020

A crucial role for working memory in temporary information processing and guidance of complex behavior has been recognized for many decades. There is emerging consensus that working-memory maintenance results from the interactions among long-term memory representations and basic processes, including attention, that are instantiated as reentrant loops between frontal and posterior cortical areas, as well as sub-cortical structures. The nature of such interactions can account for capacity limitations, lifespan changes, and restricted transfer after working-memory training. Recent data and models indicate that working memory may also be based on synaptic plasticity and that working memory can operate on non-consciously perceived information.

Introduction

Working memory maintains information in an easily accessible state over brief periods of time (several seconds to minutes). This feature is required for future goal-directed behavior and allows us to act beyond the confines of the here and now. As such, working memory is taxed by numerous laboratory and everyday cognitive challenges. The research literature on working memory is enormous, and in this Perspective we will not provide a comprehensive review. Rather, we aim to present a condensed summary of key facts and features to illustrate the ‘‘neurocognitive architecture’’ of working memory (Box 1; for related accounts, see D’Esposito and Postle, 2015; Fuster, 2009; Jonides et al., 2008).

Although there is no complete consensus on its definition, a basic feature in most conceptualizations of working memory is short-term maintenance of information in the absence of sensory input. Here, too, this definition is at the heart of our treatment of working memory. Information maintenance is considered to be the result of an interaction between basic building blocks of working memory (Figure 1A), notably a selective attention process (Figure 1B) that operates on perceptual information and related long-term memory (LTM) representations. Thus, here and elsewhere, attention is understood to be a cornerstone of working-memory processes (e.g., Baddeley and Hitch, 1974; Cowan, 1995; see Kastner and Ungerleider, 2000; Petersen and Posner, 2012). Figure 1B exemplifies maintenance of object information. First, in orange, the encoding of information into working memory is the result of interactions among selective attention processes and perceptual object representations that trigger related LTM object representations. Working-memory representations are vulnerable to distraction and interference. Therefore, when the perceptual input no longer is present, sustained attention along with a rehearsal process is crucial for maintaining the information in working memory (red contours in Figure 1B). If all the information to be maintained can ‘‘fit’’ within the focus of attention, an active maintenance process fulfills maintenance through reverberating signals between regions that provide attentional/‘‘top-down’’ signals (e.g., fronto-parietal areas; see Kastner and Ungerleider, 2000) and regions specifically related to the current content of working memory (i.e., perceptual and LTM representations). If there is more to maintain than fits within the focus of attention, an additional rehearsal process needs to complement the active maintenance process. Finally, at the retrieval phase (white contours), as in a delayed-match-to-sample task, selective attention and pattern completion processes become engaged to match the perceptual information provided at the retrieval stage with information maintained in working memory. Figure 1C exemplifies the situation in which ‘‘manipulation operations’’ are performed on the information currently maintained in working memory. This could be mental arithmetics, e.g., as in the computation-span working-memory task, or it could involve updating of the current content of working memory (e.g., O’Reilly, 2006). The concept of working memory includes the prospective use of information, which has been promoted as a major motivation for using the term ‘‘working,’’ rather than ‘‘short-term,’’ memory (e.g., Fuster, 2009). Purposeful use of the temporarily maintained information depends on the objective (goal) and structure of the task, as well as the context in which the task is performed. Together, these aspects provide the scaffold on which working memory proceeds. Accordingly, task set, prospective planning, and other cognitive control operations are integral parts of working-memory processing (purple field in Figure 1).

According to this ‘‘component processes’’ view of working memory, no processes (and correspondingly no brain structures) are unique or specific to working memory. Rather, working memory is the result of various combinations of processes that in other constellations can be functionally described in other terms than working memory (Figure 1D; cf. Cowan, 2001; D’Esposito and Postle, 2015; Fuster, 2009; Jonides et al., 2008). It should

Box 1. Current Status of the Field

- Working memory results from the interaction between several component processes, including attention, prospection, and perceptual and long-term memory representations.
- Many brain regions interact during working memory and include "executive" regions in the PFC, parietal cortex, and basal ganglia, as well as regions specialized for processing the particular representations to be maintained, such as the fusiform face area for maintaining face information.
- Persistent neural activity in various brain regions accompanies working memory and is functionally necessary for maintenance and integration of information in working memory.
- Working-memory capacity is limited and may only hold a small amount of information (absolute limits remain controversial); capacity can be increased through "chunking" bits of information into more complex units.
- Working-memory functioning changes across the lifespan with an inverted U-shaped trajectory and can be modified by training.
- Working memory may involve short-term plasticity but does not seem to require structural alterations, such as new protein synthesis, as it works by recruiting already existing synapses and ion channels ("activated LTM").

be emphasized that working memory, as conceptualized here, is a particular state of a representation (temporarily enhanced accessibility), regardless of the kind of representation. That is, working memory can basically involve any kind of representation (verbal, visual, auditory, spatial, etc.), including various procedures or temporally ordered action sequences (e.g., when following a recipe), and by extension engage many different parts of the brain, where these representations are stored. Also, often the information to be encoded into working memory does not exactly match stored representations (e.g., novel configurations of familiar objects in tests of spatial working memory or unfamiliar faces). Therefore, although stored information in LTM can support working-memory maintenance, many tasks will require encoding and maintenance of novel information and, in some cases, even information that has no clear mapping to stored information (Olsson and Poom, 2005). In the latter case, working-memory capacity will be lower and may more or less entirely rely on perceptual representations. Whether or not the encoding of such information and other information into working memory is likely to also foster new LTM representations, and by inference lead to synaptic resculpting, will be further discussed below.

Behavioral Properties of Working Memory

Capacity Limitations

A fundamental property of working memory is that it is highly limited in how much information can be held active simultaneously (Baddeley, 2003; Cowan, 2001; Luck and Vogel, 1997). Most estimates of the average capacity among healthy young adults suggest that working memory has a capacity limit of approximately 3 or 4 simple items (Luck and Vogel, 1997). This limitation highlights a sharp contrast between working memory and LTM, which is thought to have a nearly boundless capacity for storing new information from the environment. While there is broad agreement that working memory can maintain only a small amount of information simultaneously, two factors make a simple statement of a maximum limit very challenging. First, the amount of information that can be held depends strongly on whether the items can be grouped into meaningful units, or "chunks." That is, by clustering information together one can exploit preexisting information about concepts already stored in long-term memory, which allows more efficient storage in working memory, presumably by reducing the number of active elements that must be maintained in working memory. Such chunking can be observed in many domains, from the clustering of letter strings to form acronyms of familiar concepts (Miller, 1956) to the exploitation of visual statistical regularities to form grouped arrays of objects (Brady et al., 2009). Second, objects with high levels of complexity may require additional resources to adequately resolve their details. Thus, working-memory performance may be reduced for such complex items due to insufficient precision at encoding. Indeed, there is evidence for variability in encoding precision even between objects presented within a single array (van den Berg et al., 2012; Fougnie et al., 2012). When considering these factors, it becomes apparent that the functional limits to working-memory performance can vary substantially depending upon the nature of the processing demands imposed by the specific working-memory task. The opportunity to utilize either LTM or grouping tends to increase performance, while the requirement to report fine details of complex objects tends to decrease performance.

Even when grouping cues and high-precision demands are minimized, there is still debate about the nature of the maximum limit. Specifically, is it best characterized as a maximal upper limit on the number of discrete representations that can be maintained? Or is it better described as a finite pool of resources for representations that can be flexibly allocated to any arbitrary number of items? There is extensive evidence across many tasks and memoranda that individuals can remember only 3–4 simple items with near perfect accuracy, with steep drop-offs in performance for arrays that exceed this number (Figure 2A). However, flexible resource models can mimic such limits by positing that all (or most) items from a display are represented in working memory, but that with greater numbers of items the precision/resolution of each representation dwindles (Bays and Husain, 2008; Wilken and Ma, 2004). Thus, errors for arrays that have more than 3 or 4 items may be due to imprecise representations of each of the items rather than being due to items simply not being stored in working memory. Although both models make highly similar predictions regarding task performance across varying working-memory loads (e.g., that mnemonic precision declines with number of items), they differ in one key component: the role of guessing. Discrete models propose that if a subject is tested on an item from an array that is not held in one of the 3–4 slots, he or she will guess its identity. By contrast, continuous models propose that subjects never truly guess because all items from the display are assumed to be represented in working