Toward a Cognitive Neuroscience of Metacognition

Arthur P. Shimamura

Department of Psychology (#1650), University of California, Berkeley, Berkeley, California 94720
E-mail: aps@socrates.berkeley.edu

The relationship between metacognition and executive control is explored. According to an analysis by Fernandez-Duque, Baird, and Posner (this issue), metacognitive regulation involves attention, conflict resolution, error correction, inhibitory control, and emotional regulation. These aspects of metacognition are presumed to be mediated by a neural circuit involving midfrontal brain regions. An evaluation of the proposal by Fernandez-Duque et al. is made, and it is suggested that there is considerable convergence of issues associated with metacognition, executive control, working memory, and frontal lobe function. By integrating these domains and issues, significant progress could be made toward a cognitive neuroscience of metacognition.

Metacognition refers to evaluation and control of one’s cognitive processes. In this way, metacognition often suggests conscious or volitional control of thoughts, memories, and actions. Early research in this area focused on metamemory—the evaluation of memory processes (e.g., tip-of-the-tongue phenomenon, feeling of knowing) and awareness of mnemonic strategies that could facilitate remembering (see Nelson, 1992). Much of this work was sparked by developmental studies in which the ability to evaluate one’s memory processes and to be aware of mnemonic strategies improved during the course of early childhood (Brown, 1978; Flavell & Wellman, 1977). Over the years, metacognitive research has blossomed to include a variety of issues and domains—including life-span approaches, cognitive neuropsychology, theory of mind, and educational psychology (for review, see Mazzoni & Nelson, 1998; Metcalfe & Shimamura, 1994; Reder, 1996).

Fernandez-Duque, Baird, and Posner (this issue) offer a useful framework for the analysis of metacognition. Metacognition is viewed in terms of executive control processes, such as those involved in selective attention, conflict resolution, error detection, and inhibitory control. By relating metacognition to executive control, Fernandez-Duque et al. bring together an extraordinarily rich and useful pool of data not otherwise considered important for the analysis of metacognition (for similar conceptualizations see Shimamura, 1996; Umilta & Stablum, 1998). Fernandez-Duque et al. (this issue) review findings from basic cognitive psychology, cognitive neuroscience, and developmental psychology. These findings suggest a strong relationship between metacognitive regulation and executive control. They emphasize the biological bases of metacognition and suggest that midfrontal brain regions are part of a neural circuit that enables metacognitive regulation.

Metacognition: Monitoring and Control of Information Processing

In a highly influential model, Nelson and Narens (1990, 1994) described metacognition as the interplay between two levels of analysis—an object level and a meta-level. By this view, processes within the object-level are monitored by the meta-level. That is, metacognitive monitoring involves the flow of information from the object level to the meta-level. The role of the meta-level is to evaluate what is being monitored, and based on this evaluation, control object-level processing by a feedback flow of information. Thus, memory evaluations, such as judgments of learning (e.g., “how well did I learn the material?”) or feelings of knowing (“how well will I perform on a test of the material?”), can be construed as aspects of metacognitive monitoring. Metacognitive control can be construed in terms of regulating information processing, such as allocating more study time or initiating certain retrieval strategies.

As reviewed by Fernandez-Duque et al. (this issue), there is evidence to suggest that the frontal cortex contributes to metacognition. Indeed, impairments in metacognitive monitoring can be observed in patients with frontal lobe damage. In one study, Shimamura and Squire (1986) assessed the accuracy of feeling-of-knowing judgments in amnesic patients. Subjects rated their feeling of knowing for the answers to trivia questions that could not be recalled. Accuracy of their feeling-of-knowing ratings was assessed by correlating the ratings with subsequent performance on a recognition test. Patients with Korsakoff syndrome but not other amnesic patients exhibited poor metacognitive evaluations. As not all amnesic patients exhibited impaired metacognition, it was concluded that metacognitive deficits are not an obligatory feature of amnesia. The metacognitive deficit in patients with Korsakoff syndrome was attributed to more widespread impairment, such as damage to the frontal cortex. Later Janowsky, Shimamura, and Squire (1989a) found evidence for a metacognitive impairment in patients with frontal lobe lesions who did not exhibit amnesia.

Another aspect of metacognitive evaluation is source monitoring (see Johnson, Hashtroudi, & Lindsay, 1993). In source monitoring tasks, subjects must evaluate contextual information, such as remembering when or where some event occurred or who presented some information. Schacter, Harbluk, and McLachlan (1984) presented amnesic patients fictitious facts (e.g., Bob Hope’s father was a fireman) using a male or female source. Later, patients were asked to recall the fact and to determine the source of the fact. Some patients not only failed to remember the source of a presented fact but also failed to remember the entire learning episode. This “source amnesia” was correlated with performance on tests sensitive to frontal lobe damage. In another study (Janowsky, Shimamura, & Squire, 1989b), patients with frontal lobe lesions exhibited significant problems in source monitoring. In this study, patients were given the answers to trivia questions (e.g., Angel Falls is located in Venezuela). Later, they could recall facts but failed to remember that the answers were presented during the experimental session.

Recently, functional neuroimaging studies have affirmed the role of the frontal cortex in source monitoring and retrieval. In an event-related fMRI study, Nolde, Johnson, and D’Esposito (1998) asked subjects to remember whether a previously
studied item was originally presented as a word or picture. Left prefrontal activation was associated with the retrieval of source information. This finding has been replicated in another fMRI study in which subjects were required to remember whether a previously presented word had been presented on the left or right side (Rugg, Fletcher, Chua, & Dolan, 1999).

These findings, as well as those reviewed by Fernandez-Duque et al. (this issue), suggest that the frontal cortex contributes to metacognition (e.g., feeling-of-knowing judgments, source monitoring). The nature of this contribution, however, is not clear. Metacognitive disorders can be interpreted in terms of early, historical views in which the frontal cortex is considered responsible for abstract reasoning, planning, and problem solving (see Goldstein, 1936; Halstead, 1947). Such complex, high-level characterizations of metacognition do not lend themselves easily to contributions of specific cognitive (or brain) components that mediate performance on metacognitive tasks. As such, the path toward a cognitive neuroscience of metacognition has been rather obscure. On the other hand, investigations of executive control have assessed and defined specific components—such as selecting stimulus information, maintaining information in working memory, and manipulating information processing (Pashler, 1998; Petrides, 1998; Roberts, Robbins, & Weiskrantz, 1998; Shimamura, in press; Smith & Jonides, 1999). Thus, the linking of metacognition to aspects of executive control offers opportunities to define better cognitive components of metacognition.

Metacognition and Aspects of Executive Control

To what extent does metacognition involve aspects of executive control? The Nelson–Narens (1990) notion of metacognitive control is quite similar to the notion of executive control. As suggested by Fernandez-Duque et al. (this issue), supervisory models, such as the one proposed by Norman and Shallice (1986) and integrated in Baddeley’s model of working memory (Baddeley, 1986), have features that resemble metacognitive control. In particular, both metacognitive control and executive control share the primary feature of enabling top-down modulation of cognitive processes.

Surprisingly, the relationship between metacognition and executive control has not been fully appreciated. Even less appreciated (even by Fernandez-Duque et al., this issue) is the relationship between metacognition and working memory. Working memory refers to the processes and representations involved in the temporary activation or storage of information. According to Baddeley (1986, 1999), working memory is represented by a central executive that controls information in three storage buffers—the phonological loop, the visuospatial sketchpad, and most recently, the episodic buffer. In this and other views of working memory (see Cowan, 1988; Shimamura, in press), executive control can be defined as processes involved in the selection, activation, and manipulation of information in working memory. In terms of the Nelson–Narens (1990, 1994) model, object-level information that is being monitored is in working memory, and top-down control of that information involves meta-level control.

In his seminal volume on working memory, Baddeley (1986) coined the term *dys-executive syndrome* to refer to a neurological disorder in which the central executive is impaired. Baddeley (1986) suggested that damage to the frontal cortex often leads
to this disorder. In the description of the dysexecutive syndrome, Baddeley (1986) sharpened understanding of the role of the frontal cortex in terms of its role in executive control of working memory. Recently, the notion of executive control has been refined and developed even further (see Petrides, 1998; Shimamura, 2000b; Smith & Jonides, 1999).

Shimamura (in press) identified four aspects of executive control—selecting, maintaining, updating, and rerouting (see Table 1). Selecting refers to the ability to focus attention to stimulus events or activate memory representations. In conflict situations, such as the Stroop task, control must enable selection of certain stimulus features while at the same time filtering others. Maintaining refers to the ability to keep active information in working memory. Immediate memory tasks, such as digit span tasks, assess this process. Updating refers to the ability to modulate and rearrange activity in working memory. For example, in some tasks, such as the \( n \)-back task (see below), subjects must not only maintain information in working memory but must also manipulate or update information. Rerouting refers to the ability to switch from one cognitive process or response set to another. Shimamura (in press) proposed that these four aspects of executive control are arranged by level of complexity, from the most rudimentary aspect of control, selecting, to the most demanding aspect, rerouting. This view draws on analyses and theories developed by others, such as D’Esposito, Aguirre, Zarahn, and Ballard (1998); Petrides (1998); and Smith and Jonides (1999).

The first aspect of executive control—selecting—involves the focus of attention to specific sensory or memory features. As reviewed by Fernandez-Duque et al. (This issue; see also MacLeod & MacDonald, in press), analysis of the Stroop test provides a useful paradigm in which to assess the ability of selecting specific stimulus features in the presence of conflicting information. Several neuroimaging findings suggest increased activation of anterior cingulate areas (Bench, Frith, Grasby, & Friston, 1993; Peterson, Bradley, Skudlarski, Gatenby, & Zhang, 1999). Oddly, however, activation in this brain region occurs for both congruent and incongruent trials. Nevertheless, these findings do suggest that aspects of selection depend upon activation of the anterior cingulate (see also Posner, Petersen, Fox, & Raichle, 1988).

Maintaining information in working memory has also been associated with activation in frontal cortex. Impairment of immediate memory has been observed in patients with frontal lobe lesions on a variety of span tests, such as those involving numeric, spatial, color, or auditory stimuli (Baldo & Shimamura, in press; Chao & Knight, 1998; Janowsky, Shimamura, Kritchevsky, & Squire, 1989; Ptito, Crane, Leonard, Amsel, & Caramanos, 1995). Moreover, neuroimaging studies have affirmed the role

### TABLE 1

<table>
<thead>
<tr>
<th>Executive process</th>
<th>Related concept</th>
<th>Benchmark task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selecting</td>
<td>Selective attention</td>
<td>Stroop</td>
</tr>
<tr>
<td>Maintaining</td>
<td>Short-term memory</td>
<td>Digit span</td>
</tr>
<tr>
<td>Updating</td>
<td>Monitoring</td>
<td>( n )-back</td>
</tr>
<tr>
<td>Rerouting</td>
<td>Set shifting</td>
<td>Task switching</td>
</tr>
</tbody>
</table>
of the prefrontal cortex in maintaining both object and spatial information in working memory (Awh, Jonides, Smith, Schumacher, & Koepppe, 1996; Jonides et al., 1993; McCarthy, Blamire, Rothman, Gruetter, & Shulman, 1993; Smith, Jonides, Koepppe, Awh, Schumacher, & Minoshima, 1995). Interestingly, Smith et al. (1995) found posterior left-hemisphere activation for object short-term memory, but posterior right-hemisphere activation for spatial short-term memory. These findings suggest neural circuitry involving the participation of prefrontal regions with posterior cortical regions.

When information in working memory requires changes in activation, executive control must enable updating. The \( n \)-back task is a working memory task that involves updating. In this task, subjects are presented a series of stimuli, such as single letters. For each presentation, subjects respond whether or not the letter presented on the current trial is the same letter that was presented \( n \) trials ago. For example, in a 2-back task, subjects respond “yes” if the current stimulus is the same one presented two trials earlier \( (n - 2) \) trial. Thus, on each trial in a 2-back task subjects must (1) maintain in working memory the current stimulus and the last two stimuli, (2) evaluate the current stimulus with the \( n - 2 \) stimulus, (3) respond “yes” or “no,” (4) dump the \( n - 2 \) stimulus, and (5) continue maintaining the \( n - 1 \) and \( n \) stimuli for the next trial. It is clear that this task requires extensive manipulating and updating of information, rather than simply selecting or maintaining of information. Increased prefrontal activation has been observed on this task in neuroimaging studies (Awh et al., 1996; Carlson, Martinkauppi, Raemae, & Salli, 1998; Cohen, Perlstein, Braver, & Nystrom, 1997).

Updating is also critical on certain memory retrieval tasks. For example, in the FAS verbal fluency task (Benton & deHamsher, 1976; Milner, 1964), subjects are given 1 min to retrieve words that begin with a specific letter (F, A, and S). Such fluency tasks require subjects to monitor previously responded items and prevent (inhibit) prior from future responses. Failure to update responses in working memory leads to perseverations. Indeed, patients with frontal lobe lesions exhibit significant impairment on fluency tasks (Baldo & Shimamura, 1998; Benton & deHamsher, 1976; Jones-Gotman & Milner, 1977; Milner, 1964). Moreover, in a PET study (Frith, Friston, Liddle, & Frackowiak, 1991) verbal fluency performance was associated with activation in the prefrontal cortex (see also Thompson-Schill et al., 1998).

Rerouting involves a global shift of information processing—from stimulus registration to response selection. Set or task switching paradigms are benchmark tests of rerouting. For example, Dunbar and Sussman (1995) assessed patients with frontal lobe lesions on a Stroop-like paradigm. Subjects were asked to name pictures of objects and to ignore simultaneously presented words. Patients with frontal lobe lesions did not exhibit any greater interference on this Stroop-like task compared to control subjects. However, in a set shifting variation of the task, subjects were cued at the beginning of each trial as to whether they were to name the picture or read the word. Patients with frontal lobe lesions were particularly slowed in the set shifting condition.

Aspects of rerouting have been assessed in various set switching paradigms, which have become popular in studies of attention (Allport, Styles, & Hsieh, 1994; Monsell, 1996). For example, Duncan et al. (1996) asked subjects to monitor two streams of
stimuli (random series of letters and digits), one presented on the left and one on the right. Initially, they were asked to report only the letters from one side. Later, a cue directed subjects to switch and report the letters presented on the other side. Patients with frontal lobe lesions exhibited difficulty in switching between the two streams. Findings from dual task paradigms also suggest the role the prefrontal cortex in rerouting. In an fMRI study, D’Esposito, Detre, Alsop, and Shin (1995) asked subjects to perform a semantic judgment task (e.g., monitor each time a type of vegetable is presented) and a visuospatial task (e.g., mental rotation) either separately or simultaneously. Significant prefrontal activation was observed when subjects performed the two tasks simultaneously but not when they performed the two tasks separately.

A classic test of frontal lobe function, the Wisconsin Card Sorting Test (Milner, 1964), involves set shifting and rerouting. In this test, stimuli are multidimensional, and subjects must determine which stimulus dimension (color, shape, number) is relevant on any given trial. Patients with frontal lobe lesions fail to appreciate shifts in the relevant dimension and tend to perseverate on previously rewarded dimensions. Similarly, concept reversal tasks (Owen et al., 1993) suggest the role of the frontal cortex in the ability to shift sets. In particular, patients with frontal lobe lesions have difficulty rerouting process from a previously successful or dominant set to a new set.

Theories of Executive Control and Frontal Lobe Function

To the extent that aspects of executive control can inform metacognitive research, it may be helpful to consider theoretical frameworks associated with control processes. Not coincidentally, many of these theories address the role of the frontal cortex in mediating aspects of control. One of the few computational models that have addressed specifically metacognition and frontal lobe function is Metcalfe’s CHARM model (Metcalfe, 1993). In her model, metacognitive evaluations, such as feelings of knowing, are based on a familiarity check that is computed between new information and what is already stored in memory. As a result of this operation, which can be construed as ‘novelty monitoring,’ the degree to which new information is bound into episodic memory is modulated. Interestingly, without this monitoring-control operation, the model performs poorly and similarly to patients with frontal lobe dysfunction.

Kimberg and Farah (1993) proposed a computational model in which executive control is imposed by affecting links that associate information in working memory. These associative links allow selection of appropriate items in working memory based on contextual factors, such as maintaining task-relevant processes. For example, in the Stroop test, the contextual factor involves responding to ink color and not to color names. Without such associative links, performance is less contingent upon top-down (i.e., contextual) control and problems arise that simulate problems associated with frontal lobe dysfunction. For example, this model can account for problems on the Wisconsin Card Sorting Test, Stroop test, and other tests that involve top-down control of stimulus selection.

In a somewhat related view, Cohen and Servan-Schreiber (1992) suggest that context setting can be viewed as the primary feature of selective attention. In this model,
context also refers to task-relevant information, such as determining what stimulus feature to attend to or which respond mode to initiate. As is stated by Cohen and Servan-Schreiber (1992, p. 46), “By this definition, context information is relevant to but does not form part of the content of the actual response.” Based on this view, a connectionist model of prefrontal function is developed in which the maintenance of the context controls aspects of selective attention. Indeed, failure to maintain the context simulates problems associated with frontal lobe dysfunction.

In the models proposed by Cohen and Servan-Schreiber (1992) and Kimberg and Farah (1993), executive control is implemented by reinforcing task-relevant processes or associations (i.e., context setting). Such models can be viewed as refining information processing by heightening or selecting appropriate information processing. Significant success in modeling executive control and frontal lobe function has been attained by this kind of mechanism. On the other hand, others have also emphasized the role of inhibitory control of extraneous or irrelevant information processing (see Knight, Scabini, & Woods, 1989; Knight, Staines, Swick, & Chao, 1999; Shimamura, 1995, in press). The contribution of inhibitory control in metacognition is also prominent in the view proposed by Fernandez-Duque et al. (this issue).

To what extent does executive control involve not only the reinforcing of task-relevant processing but also the inhibiting of task-irrelevant processing? As shown by the computational models of Cohen and Servan-Schreiber (1992) and Kimberg and Farah (1993), implementation of an inhibitory control mechanism is not necessary to simulate some aspects of frontal lobe function. Also, in terms of behavioral outcome, it is extremely difficult to differentiate the selection of appropriate responses from the inhibition of inappropriate responses. In the end, the two mechanisms lead to similar behavioral outcomes. There is, however, some physiological evidence to suggest that the prefrontal cortex is involved in inhibiting activation in posterior cortex. Knight et al. (1989) showed that patients with frontal lobe lesions exhibit posterior evoked potentials that are greater than those observed in control subjects. That is, sensory evoked potentials appeared to be disinhibited as a result of frontal lobe damage.

One computational model that implements both selective and inhibitory control is Grossberg’s (1982, 1999) general model of neural control called adaptive resonance theory. In this model, control is implemented by a top-down filtering mechanism that enhances task-relevant information in working memory and inhibits similar but irrelevant information. This kind of filtering is analogous to center-on, surround-off receptive fields. That is, cortical processing is sharpened both by the enhancement of relevant activity (center-on) and by the inhibition of irrelevant activity (surround-off). Although this mechanism may be viewed as a ubiquitous operation that controls many aspects of neuronal processing, on a larger scale the prefrontal cortex may implement such a dynamic filtering mechanism more globally to control task-relevant information processing not amenable to interactions within local circuits (for a similar view, see Shimamura, in press).

Concluding Remarks

Interestingly, Fernandez-Duque et al. (this issue) integrate emotional and cognitive regulation in their analysis of metacognition. They make this connection based on
findings of increased activation in the anterior cingulate for both cognitive regulation and emotional regulation. It is unclear, however, whether this brain region serves all forms of task setting or selective attention or whether there is some special link between emotional and cognitive control.

Another brain region, the orbitofrontal cortex, is also involved in emotional control. Patients with orbitofrontal lesions exhibit disinhibition of emotional responses and inappropriate social behavior. This problem is exemplified dramatically in the case of Phineas Gage, the 19th-century railroad foreman who sustained a horrendous accident involving an iron rod that passed through his orbitofrontal cortex (see Macmillan, 1986, 2000; and Shimamura, 2000, for a related case). Patients with lesions in this area exhibit emotional outbursts, inappropriate responses to social situations, and risky decision-making behavior (Bechara, Damasio, Damasio, & Anderson, 1994; Rule, Shimamura, & Knight, 1999; Stone, Baron-Cohen, & Knight, 1998). To what extent can issues of emotional regulation be linked to metacognitive research? Perhaps, as suggested by Fernandez-Duque et al. (this issue), the same brain activations involved in emotional regulation are also involved in cognitive regulation. Another possibility is that different areas in frontal cortex control different forms of processing. In other words, there are various metalevel systems that monitor and control different aspects of information processing (for further detail, see Shimamura, in press).

Finally, there is incontrovertible evidence suggesting a trend toward a cognitive neuroscience perspective for many if not all aspects of human cognition. The advent of event-related fMRI (D’Esposito, Zarahn, & Aguirre, 1999) has brought behavioral paradigms used in neuroimaging studies closer to those commonly used in basic cognitive research. Moreover, the spatial resolution in identifying regions of activation is extremely impressive and replicable patterns of activation have been observed across laboratories. Findings from functional neuroimaging studies provide information concerning brain regions that are correlated with certain cognitive function. Of course, such techniques are helpful only to the extent that careful behavioral methodology is implemented. Also, findings from patients with circumscribed brain injury are important as findings of functional dissociations suggest that some brain areas not only are related to cognitive function but are necessary for appropriate behavioral responses. It is clear—as suggested by Fernandez-Duque et al. (this issue)—that significant progress can be made toward a cognitive neuroscience of metacognition by relating aspects of metacognition with aspects of executive control.

ACKNOWLEDGMENTS

This research was supported by NIH Grants MH48757 and NS17778 to A. P. Shimamura.

REFERENCES


