The Role of the Basal Ganglia–Anterior Prefrontal Circuit as a Biological Instruction Interpreter

Andrea STOCCO [a,1], Christian LEBIERE [b]
Randall C. O'REILLY [c] and John R. ANDERSON [b]
[a] *Institute for Learning and Brain Sciences, University of Washington, Seattle, WA*
[b] *Department of Psychology, Carnegie Mellon University, Pittsburgh, PA*
[c] *Department of Psychology, University of Colorado at Boulder, Boulder, CO*

Abstract. Intelligent and versatile behavior requires the capability of adapting to novel and unanticipated situations. When facing novel and unexpected tasks, a fast and general solution consists in creating new declarative task representations, and subsequently acting upon them. Although this mechanism seems straightforward in general terms, it poses significant difficulties to be implemented in a biological model, and the exact neural substrates of this process are still unknown. Based on the analysis of two different computational models, we hypothesized that the brain circuit for interpreting instructions would comprise the aPFC (holding dependencies among specialized cortical areas) and the basal ganglia (orchestrating the exchange of information among regions). To verify this hypothesis, we designed and ran an fMRI experiment where participants had to perform changing tasks that consisted of different combinations of atomic cognitive operations. Both models and experimental data suggest that the aPFC is critical in representing abstract knowledge that reflects planned cognitive operations. This is consistent with the late appearance of aPFC in the evolution of the human brain, and its role in enabling human intelligence and culture. On the other hand, results and simulations show that the effect of this cortical region is made possible by the contribution of the basal ganglia circuit, which works as a general-purpose interpreter of declarative knowledge.

Keywords. Instructions; Basal Ganglia; Cognitive Models; Neural Networks.

Introduction

One of the hallmarks of intelligent behavior is the capability of directing one's own behavior on the basis of predefined, declarative representations. This capability is useful because declarative knowledge is usually more flexible to manipulate than other types of knowledge, and can be more easily communicated. Humans routinely exhibit this type of intelligent behavior when they are engaged in complex tasks such as planning or problem solving. Perhaps the most striking example of this behavior is following instructions, i.e. the capability of traducing abstract representations of behavior into action. This process is akin to interpreting a programming language

[1] Corresponding Author: Andrea Stocco, Institute for Learning and Brain Sciences, University of Washington, Seattle, WA 98195. Email: stocco@uw.edu.

statement in computer science. Computationally, this process requires some mandatory computational steps that are independent of the implementation of the interpreter itself; in particular, instructions need to be translated into operations and structures that match the underlying hardware.

In this paper, we provide converging and computational evidence that a particular circuit in the human brain is responsible for interpreting instructions. In particular, we present two different models of the task, developed in two different modeling frameworks, together with preliminary results from a neuroimaging experiment. The model and the data suggest that the circuit involved in interpreting instructions comprises the anterior regions of the prefrontal cortex and a set of medial nuclei collectively known as the basal ganglia.

1. The Task

Instructed behavior is seldom investigated in cognitive psychology, and data from the instructional phase of experiments routinely discarded. Thus, we developed a novel task that was used for both testing our models and collecting experimental data from participants. The task consists in solving a series of arithmetic problems, each of which is combination of three operations, such as "divide x by 3", "multiply y by 2", and "multiply x and y". Each problem required exactly two input numbers (x and y) and always contained one binary and two unary operations. In order to ensure that intermediate and final results were always integer numbers, participants were instructed to use the quotient as the result of a division, and discard the remainder (e.g., 7 / 2 = 3). The three operations were randomly selected from a set of five, each of which was associated to an alphabetical letter $L = \{A, B, C, D, E\}$. Table 1 illustrates the operations used in the experiment and provides some examples.

Each trial consisted of three consecutive phases: (a) An instruction phase, where the problem was presented; (b) An execution phase, where the two input numbers where presented and calculations were performed; and (c) A response phase, were participants indicated whether a certain number was the solution to the problem or not. The structure of a sample trial is illustrated in Figure 1.

Instructions were presented as a string of letters and variables such as *AExDy*. Instructions were in prefix notation, so that the above problem was interpreted as $A(E(x), D(y))$, that is, $(x / 3) \times (y + 1)$ (see Table 1).

Table 1. The five operations used in the experiment

Operation	Meaning	Examples	
$A(x, y)$	$x \times y$	$A(4, 2) = 4 \times 2 = 8$;	$A(2, 3) = 2 \times 3 = 6$
$B(x, y)$	x / y	$B(8, 2) = 8 / 2 = 4$;	$B(6, 3) = 6 / 3 = 2$
$C(x)$	$x \times 2$	$C(4) = 4 \times 2 = 8$;	$C(3) = 3 \times 2 = 6$
$D(x)$	$x + 1$	$D(7) = 7 + 1 = 8$;	$D(3) = 3 + 1 = 4$
$E(x)$	$x / 3$	$E(9) = 9 / 3 = 3$;	$E(6) = 6 / 3 = 2$

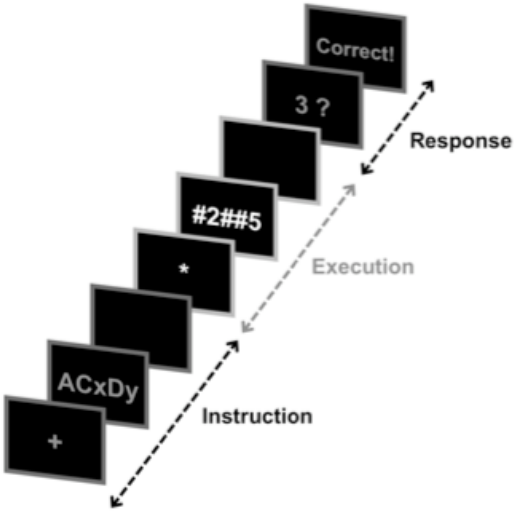

Figure 1. Structure of a sample trial in the experiment.

2. Models for Interpreting Instruction

To explore the nature of the processes involved in interpreting instructions we developed two computational cognitive models that could perform the task. The two models exemplify two complementary and converging approaches. The first model was developed within an integrated cognitive architecture that allows symbolic encoding and decoding of declarative knowledge by production rules. The second model, on the other hand, was built upon an existing lower-level neural network model of the basal ganglia-prefrontal circuit.

2.1. The ACT-R Model

The higher-level model of the instruction task was implemented in ACT-R [1], a cognitive architecture that has been particularly successful in modeling human learning and memory and, more recently, neuroimaging data [2]. ACT-R includes declarative knowledge, represented as dictionary-like arrays of slot-value pairs called *chunks*, and procedural knowledge, represented as production rules.

Chunks are permanently stored in a long-term memory but, in contrast to most production systems, can be accessed only when available in buffers serving as interface with memory and sensory modules [1]. Buffers have a limited capacity of one chunk only, and can only be accessed by production rules. Figure 2 illustrates the relationship between modules, buffers, and procedural knowledge.

Production rules specify the chunk patterns across the various buffers in both the condition and action sides. Production rules can typically variabilize only the slot value of a chunk, and only under specific circumstances (effectively, involving no search) can they use a variable to refer to a specific slot (and not its value).

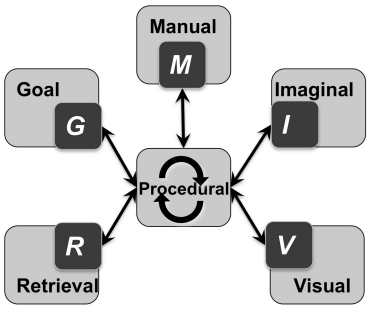

Figure 2. Overview of the ACT-R architecture [1]. Modules are in light grey; buffers in dark grey.

The ACT-R model can execute the entire task, including visually parsing the screen and performing simulated motor responses. During the instruction phase, the model encodes each problem as a series of three consecutive steps. Each step is created by scanning the instruction string right to left, recursively finding the first unattended letter; retrieving the associated operation; and determining whether to apply the operation it to either *x*, *y*, or both. During the execution phase, the model simply retrieves the three steps in order, executing the corresponding operations and updating the values of *x* and *y* at the conclusion of each step.

In ACT-R, all the task information must be either available in the buffers or retrieved prior to being used. Thus, some choices had to be made on how to distribute the relevant task information. These choices are usually constrained both by the specific computations available in a module and its established mapping to a brain region [1]. For instance, the intermediate values of *x* and *y*, together with the current step's position in the series, were stored in a chunk in the imaginal buffer. This is consistent with the imaginal buffer's association with the parietal cortex, a brain region critically involved in visuo-spatial working memory and mathematical cognition [1-3].

The two most critical parts of the model are the chunks representing the problem steps and the production rules that interpret them. Problem steps were maintained in a special module that mimics the computations of the existing goal module. A new module was created because the goal module is associated with internal control states and not with declarative templates for future actions [2]. No established association exists between this novel module that processes instructions and a brain region, but some speculations are possible. Its role in holding higher-level representations that tie together lower-level actions suggest an association with the anterior prefrontal cortex (aPFC), which has been often associated with similar functions [4,5]

The model's second key component is the production rules that interpret instructions. These rules differ from standard ACT-R rules in that they use variables to indicate slot names, and not only slot values. This procedure is needed to properly instantiate operations are referring to either *x* or *y*. The execution of production rules has been associated with the basal ganglia [2], and basal ganglia activity has been successfully predicted either simply counting the number of production rules fired per time unit [1-3], or by counting the number of variable bindings per time unit [6]. Thus, the model predicts that the activity of the basal ganglia should reflect the increased number of variables in the *Execution* phase.

2.2. The Conditional Routing Model

The ACT-R model provides only indirect evidence of the neural basis of interpreting instructions. More compelling evidence can be obtained by modeling the process of following instructions within a framework that directly deals with the underlying biological circuits.

Interpreting instructions requires frequent updating of representations in working memory, a process that is mediated by a neural loop that connects various cortical areas with the prefrontal cortex through the basal ganglia. Several models of this circuit exist (e.g., [7-9]). The conditional routing model by Stocco, Lebiere, and Anderson [9] is both consistent with the known biology of the circuit and provides a biological explanation for some of the computations required by an ACT-R model—in particular, for the variable binding process.

The basal ganglia comprise a number of interconnected nuclei that route signals form the entire cortex to the frontal lobes. The heart of the model is the simulated striatum, which receives afferents from the entire cortex and is the entry point of the circuit. The striatum is modeled as a flat structure of projection neurons, the so-called striatal matrix, controlled by a set of interneurons. Biologically, interneurons have a high tonic activity maintaining a constant inhibition on projection neurons [10]. In the model, projection neurons have a high threshold θ that is calculated to match the expected incoming signals from the cortex and the inhibitory interneurons:

$$\theta \approx \Sigma_i w_i E(x_i) \tag{1}$$

where w_i is weight of the synapses formed with pre-synaptic neuron i, and $E(x_i)$ is the rate-coded expected activation value of i. Variable binding is permitted by the particular two-level organization of the model striatum. The striatal matrix is divided into regions that reflect the organization of the cortex. Thus, every cortical region is represented by a corresponding patch on the striatal matrix. Each path also has an internal organization, with sub-compartments representing different parts of the cortex the original cortical regions projects to. This two-level organization can be imagined as a matrix of source-destination pairs of cortical regions, and the entire striatum can be imagined as a switchboard [9]. Figure 3 provides a visual rendition of this organization.

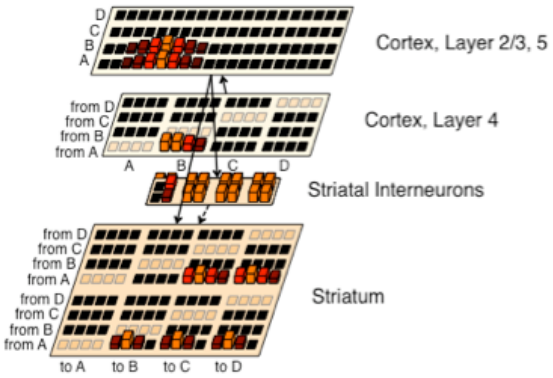

Figure 3. Organization of the striatum and the cortex in the routing model [9].

Consistent with neurophysiology [10], neurons in the striatum are mostly silent, with only a minority of them actually active at any time. In our model, the active neurons correspond to the active combinations of sources and destinations. Ignoring local computations that occur within striatal neurons, the final state of the striatum is the block product $\mathbf{v} \otimes \mathbf{M}$ of the initial vector $\mathbf{v}$ of activations in the source cortical area, and the switchboard matrix of allowed destinations $\mathbf{M}$. The block product is a special case of tensor product—a powerful mechanism for variable binding in neural networks [11]. In this case, the variable is the destination cortical region, which is bound to the value $\mathbf{v}$, i.e. the original content of the source region. Notice the similarity between this mechanism and ACT-R's production rules, where variables are used to bind the contents of a particular destination buffer to the values held in a source buffer.

2.2.1. Instructions and the Control of Variable Binding

The very structure of the model suggests one natural way of interpreting instructions. In the routing model, the execution of an operation simply consists in the proper transfer of signals between cortical regions. For example, updating the values of x and y after an operation consists in copying the representation held in the prefrontal region that retrieves arithmetic facts to the cortical region that temporarily holds either x or y. This transfer is directed by the proper activation of cells in the striatum. In fact, any internal operation can be properly represented as a switchboard matrix that shares the same organization of the striatum.

Following this logic, we expanded the routing model by adding a novel cortical area that shares the switchboard organization of $\mathbf{M}$, so that variable bindings in the striatal matrix can be properly controlled by the activation of the corresponding cells in the region. In addition to having a switchboard organization, neurons in this region need to have a very low tonic activity; this is required so that their expected activation value $E(x)$ is low, minimizing the effect in calculating the thresholds in Equation (1). and making it easy to bring the activation of projection neurons above the threshold θ. In fact, we ran a number of simulations showing that this mechanism is sufficient to make the model execute arbitrary operations such as the instructed arithmetic operations required by the task.

One can wonder about the biological plausibility of such a hypothetical region. In fact, the anterior part of the prefrontal cortex (aPFC), and in particular the frontal pole, possesses exactly the necessary computational characteristics. Specifically, the aPFC receives massive projections from the frontal lobe, and these projections are topologically organized, thus providing an organization that resembles the frontal projections to the striatum Also, this region is usually silent during the execution of most tasks, with its most polar part actually deactivates during a task [5], thus satisfying the condition of a low expected value. Finally, its projections seem to innervate a large part of the head of the caudate nucleus, the most frontal part of the basal ganglia [12].

3. Neurocognitive Evidence

So far, two different computational models have been presented that provide evidence that the process of interpreting instructions can be achieved by the joint workings of the

anterior prefrontal cortex and the basal ganglia. Before testing this prediction, it is worth examining whether it is consistent with the existing experimental evidence.

There is mounting evidence for the role of the aPFC in holding higher-level representations, such as those needed in analogical and meta-cognitive tasks [13], or in tasks that require branching of different goals [4].

To the best of our knowledge, the involvement of the basal ganglia in interpreting instructions has not been tested directly. Many converging lines of research, however, have singled out the basal ganglia as a potential basis for flexible behavior in general. For instance, there is an obvious connection between the function of the basal ganglia and the regulation and updating of working memory. Patients with either Parkinson's or Huntington's disease are impaired in tasks tapping different forms of working memory [14], and working memory-related activity in the basal ganglia has been reported in a number of neuroimaging studies [15,16]. Individual differences in working memory performance are also related to genetic differences in the expression of dopamine receptors in the basal ganglia [17], and high working memory capacity individuals show greater modulation of basal ganglia activity with increasing task demands [18]. Other evidence comes from tasks that require strategic reasoning to cope with changes in task rules. These tasks are often used in investigations of so-called executive functions. One such example is the Wisconsin Card Sorting Task, which requires participants to sort cards according to rules they need to discover by trial and error, and are continuously changed by the experimenter. Again, Parkinson patients are unable to correctly perform this task [19].

In summary, the basal ganglia are recruited in a number of different tasks that share the common property of requiring flexible restructuring of behavior, either because new task rules come into play of because the trial difficulty changes. Furthermore, individual differences in performance in these kinds of paradigms are reliably associated with individual differences in the basal ganglia, either at the level of functional responses or at the level of neuroanatomy.

4. The Experiment

The models' predictions were tested in a neuroimaging study. Ten participants were recruited to perform the task previously described while lying in a 3T fMRI scanner. Their brain activity was recorded at a rate of a full volume acquisition every 2 seconds, with 34 oblique slices acquired for each volume. Each participant solved 80 problems, divided into four blocks of 20 trials each. Unlike most fMRI experiments, each problem was self-paced.

In addition to the distinction between encoding and executing a set of instructions, the experiment manipulated the amount of practice as a second factor. This manipulation provides an additional means to isolate the specific act of interpreting instructions, which is important when analyzing data with a limited number of participants (see below). Practice was manipulated by having participants perform a subset of the problems before the experiment. During the experiment, half of trials were novel and half came from the subset of practiced trials.

4.1. Results

Because the low number of participants limited the statistical power of traditional analysis, we performed a conjunction analysis, using statistical parameter maps thresholded at a liberal voxel-level value ($p < 0.01$, uncorrected) to isolate regions that are activated in two or more target contrasts.

The ACT-R model predicts that the module corresponding to the aPFC region should be more active in Novel than Practiced trials, in both the Instruction and Execution phases. Thus, we created to statistical parameter maps (one for the Instruction phase, one for the Execution phase) that identified those voxels that were statistically more active during the Novel than during the Practiced trials (i.e., Novel > Practiced). As predicted, the analysis identified a cluster of voxels located in the aPFC region, with smaller cluster located in even anterior position in the frontal lobe. The results of this analysis are illustrated in the top part of Figure 4; the crosshairs highlight the aPFC regions.

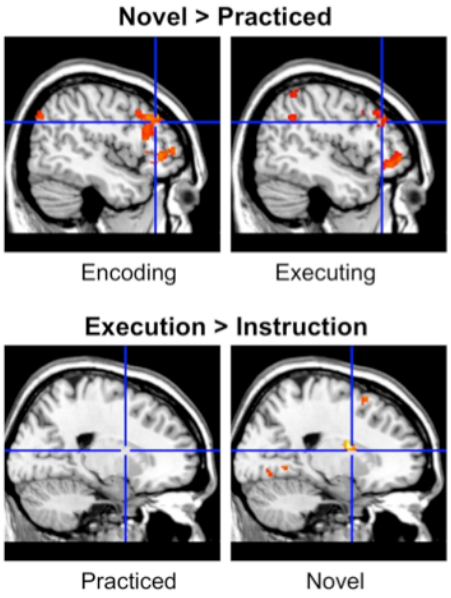

Figure 4. Results of the experiment.

Both the ACT-R and the conditional routing model predict that the basal ganglia should be more active during the Execution phase than during the Instruction phase. Additionally, both models predict that this asymmetry should hold for Novel problems only; Practiced problems can be executed as a routine, without referring to the original instructions, and there is no reason to expect any additional basal ganglia involvement during their execution. To verify this hypothesis, we created two new contrast maps that identify those voxels more active in the Execution than the Instruction phase (i.e., Execution > Instruction) in the Novel and in the Practiced problems, respectively. As predicted, we found one cluster of voxels that was more active during the Execution phase and corresponded to the right striatum; it is indicated by the crosshairs in the bottom part of Figure 4. As predicted this cluster showed up only in the contrast map

obtained from Novel trials; Practiced problems did not show, in fact, any voxel that was more active during the Execution phase. In summary, our preliminary results support our models' predictions and permit to identify two regions crucially involved in interpreting instructions: the aPFC, probably responsible for encoding and accessing abstract representations of cognitive actions, and the basal ganglia, probably responsible for performing the necessary variable bindings while interpreting instructions.

5. Conclusions

This paper has presented two models and a neuroimaging study of how humans interpret instructions. The models and the experimental data suggest that a circuit formed by the basal ganglia and the anterior prefrontal cortex provide the necessary computations to translate abstract representations of behavior into action.

There are at least three reasons why we believe that understanding how the brain interprets instructions is important. First, following arbitrary representations of actions is the core capability that underlies flexible behavior and planning. Thus, it provides one of the foundations of general intelligence.

Additionally, interpreting instructions constitutes an interesting problem because, while its solution is rather simple within symbolic frameworks such as production systems, it is instead rather complex to treat within a connectionist framework. Thus, it provides a challenge for bridging the gap between abstract computations and their biological counterpart.

The third and final reason why we consider this problem worth investigating is that it provides access to the basic operations of the human brain. As suggested in the introduction, the process of interpreting instructions consists in the translation of abstract representations into basic primitive operations. Thus, understanding how this translation mechanism works implicitly provides information about the nature of the primitive computations available in the human brain and their implementation.

6. Acknowledgments

This research was made possible by award FA9550-08-1-0404 from the Air Force Office of Scientific Research (AFOSR) to John Anderson, Randall C. O'Reilly, and Christian Lebiere; by support from the Army Research Laboratory's Robotics Collaborative Technology Alliance to Christian Lebiere; and by a special award from the Brain Imaging Research Center (BIRC) of Pittsburgh to Andrea Stocco.

References

[1] J.R. Anderson, *How can the human mind occur in the physical universe?* Oxford University Press, New York, NY, 2007.

[2] J.R. Anderson, J.M. Fincham, Y. Qin, and A. Stocco, A central circuit of the mind. *Trends in Cognitive Sciences* **12** (2008), 136-143.

[3] J.R. Anderson, Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science* **29** (2005), 313-341.

[4] E. Koechlin, G. Basso, P. Pietrini, S. Panzer, and J. Grafman, The role of the anterior prefrontal cortex in human cognition. *Nature* **399** (1999), 148-51.
[5] S.J. Gilbert, S. Spengler, J.S. Simons, J.D. Steele, S.M. Lawrie, C.D. Frith, and P.W. Burgess, Functional specialization within rostral prefrontal cortex (Area 10): A meta-analysis. *Journal of Cognitive Neuroscience* **18** (2006), 932–948.
[6] A. Stocco and J.R. Anderson, Endogenous control and task representation: An fMRI study of algebraic problem solving. *Journal of Cognitive Neuroscience* **20** (2008), 1300-1314.
[7] M.J. Frank, B. Loughry, and R.C. O'Reilly, Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective & Behavioral Neuroscience* **1**, (2001) 137-160.
[8] F.G. Ashby, S.W. Ell, V.V. Valentin, and M.B. Casale, FROST: a distributed neurocomputational model of working memory maintenance. *Journal of Cognitive Neuroscience*, **17** (2005), 1728-1743.
[9] A. Stocco, C. Lebiere, and J. R. Anderson, Conditional routing of information to the cortex: A model of the basal ganglia's role in cognitive coordination. *Psychological Review* **117** (2010), 540-574.
[10] J.M. Tepper, and J.P. Bolam, Functional diversity and specificity of neostriatal interneurons. *Current Opinion in Neurobiology* **14** (2004), 685-692.
[11] P. Smolensky, Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* **46** (1990) 159-216.
[12] A. Di Martino, A. Scheres, D.S. Margulies, A.M.C. Kelly, L.Q. Uddin, Z. Shehzad, B. Biswal, J.R. Walters, F.X. Castellanos, and M.P. Milham, Functional connectivity of human striatum: A resting state fMRI Study. *Cerebral Cortex* **18** (2008), 2735-2747.
[13] E. Ferrer, E., E. O. O'Hare, and S. A. Bunge, Fluid reasoning and the developing brain. *Frontiers in Neuroscience*, **3** (2009), 46-51.
[14] M.G. Packard and B.J. Knowlton, Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience* **25** (2002) 563-593.
[15] T.S. Braver, J.D. Cohen, L.E. Nystrom, J. Jonides, E.E. Smith, and D.C. Noll, A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage* **5** (1997) 49-62.
[16] F. McNab and T. Klingberg, Prefrontal cortex and basal ganglia control access to working memory. *Nature Neuroscience* **11** (2008) 103-107.
[17] Y. Zhang, A. Bertolino, L. Fazio, G. Blasi, A. Rampino, R. Romano, M.-L. T. Lee, T. Xiao, A. Papp, D. Wang, and W. Sadée, Polymorphisms in human dopamine D2 receptor gene affect gene expression, splicing, and neuronal activity during working memory. *Proceedings of the National Academy of Sciences* **104** (2007), 20552-20557.
[18] C.S. Prat, T.A. Keller and M.A. Just, Individual differences in sentence comprehension: a functional magnetic resonance imaging investigation of syntactic and lexical processing demands. *Journal of Cognitive Neuroscience* **19** (2007), 1950-1963.
[19] O. Monchi, M. Petrides, V. Petre, K. Worsley and A. Dagher, Wisconsin Card Sorting revisited: Distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *The Journal of Neuroscience* **21** (2001), 7733–7741.