Multi-Modal Cognitive States: Augmenting the State in Cognitive Architectures

B. Chandrasekaran

Laboratory for AI Research
Department of Computer Science & Engineering
The Ohio State University
Columbus, OH 43210
Email: Chandra@cse.ohio-state.edu

Abstract

Different streams of AI idealize different aspects of human cognition. Idealization of intelligence as an embodied activity, involving an integration of cognition, perception and the body, places the tightest constraints on the design space for AI artifacts, forcing AI to deeply understand the design tradeoffs and tricks that biology has developed. I propose that a step in the design of such artifacts is to broaden the notion of cognitive state from the current linguistic-symbolic, Language-of-Thought framework to a multi-modal one, where perception and kinesthetic modalities *participate* in thinking. This is in contrast to the roles assigned to perception and motor activities as modules external to central cognition in the currently dominant theories in AI and Cognitive Science. I develop the outlines of this proposal, and describe the implementation of a bi-modal version in which a diagrammatic representation component is added to the cognitive state.

AI and Idealization of Human Cognition

Before I present my proposal on the multi-modal cognitive state, I wish to digress a little by discussing possible relations between AI system design and cognitive science, or as I prefer to reword it, between AI and human or biological cognition. There is *always* a connection, since all of AI is based *idealization* of on one or other aspect of biological cognition. Different AI approaches idealize different aspects of human cognition as *the* characteristic idealization of intelligence. The logic approach to AI is based on a model of cognition as *reasoning*, idealizing the linguistic/logical aspect of deliberative thought. Proposed alternatives to logic, such as frames and scripts, use a different idealization, *viz*., cognition is what memory does. Newell and Simon's means-ends methods of problem solving and their successor, the Soar architecture (Newell, 1990), are modeled after human deliberative problem solving and learning. In a different vein, Brooks' robotics work (Brooks, 1986) is based on a model of human physical behavior as arising less from reasoning and problem solving and more from a specific kind of hierarchical organization of a behavioral repertoire.

The different idealizations are helpful in building AI systems of specific types. For building knowledge-based systems, or a system to prove theorems in Group Theory, perhaps modeling intelligence as goal-directed logical or quasi-logical reasoning is sufficient. It makes no sense to build it using Brooks' subsumption architecture. Consider, on the other hand, the design of a *coffee-making robot*: a robot that would, among other things, make coffee and bring it up in response to a voice command in English; and, if in the process it finds that the refrigerator is out of milk, would drive to the supermarket, buy milk and finish coffee making.

The most rigid constraints on models of cognition are placed by the coffee-making, spoken-language-understanding robot mentioned above. Unlike Group Theory theorem provers, this robot would have to coordinate in real time perception and motor action with language understanding and problem solving. It would benefit from having memories that contain perceptual representations, such as mental maps, in addition to traditional, linguistically represented knowledge. Unlike Brook's robots, this robot would have to understand language, reason, and learn at various levels. Further, the fact that everything the robot does is with respect to achieving goals in the world means that all its reasoning is subject to whether something will be *good enough* for the task at hand, rather than whether the solution is "correct" with respect to the abstract version of the problem. Almost all the issues that have been raised as objections or extensions to GOFAI arise naturally in this context: situatedness, use of the external world as representation, active perception, and embodiment are just some examples. On the other hand, impressive as such a robot might be as an AI achievement, in the task proving theorems in Group Theory, it might still fall short of the achievement of theorem provers that have been built solely for that task domain.

The more general and flexible we want the AI system to be, the more numerous are the constraints placed on the proper idealization of cognition. But there is no sharp and complete characterization of the idealization of cognition.

Should robots scratch their heads when they are thinking and feel a bit lost? Maybe scratching the head plays an important role in social aspects of cognition, which in turn might affect what they learn and how. While no one in AI today would seriously argue for a head-scratching robot as a better design, there is also no sharp dividing line between where cognition starts and perception and motor processes end.

Cognition, Architecture, Embodiment and Multimodality of Thought

Generality and flexibility are hallmarks of intelligence, and this has led to a search for *cognitive architectures*, exemplified by Soar (Newell, 1990) and ACT-R (Anderson, 1996). Different task-specific cognitive systems may be programmed or modeled by encoding domain- and task-specific knowledge in the architecture. These architectures are also based on idealizations of biological intelligence. Abstracting from human cognition, they typically posit a working memory (WM), a long term memory (LTM), mechanisms to retrieve from LTM and place in WM information relevant to the task, mechanisms that help the agent set up and explore a problem space, and mechanisms that enable the agent to learn from experience. The specific mechanisms proposed and the representational formalisms on which they work together constitute the architecture designer's theory of cognition. Because of their origin from an idealization of human cognition, it is not surprising that Soar and ACT-R are useful both to build AI agents as well to build cognitive models.

An important aspect of their representational commitment is that the cognitive state, roughly characterized as the content of the WM, is *symbolic*, or to use a more precise term, predicate-symbolic. That is, the knowledge in LTM as well as representations in WM are compositions of symbol strings where the symbols stand for individuals, relations between individuals, or various ways of composing relational predicates, in some domain of interest. For example, in a blocks world, a state representation might be ON(A,B) & Left(B,C). In this the designs of these architectures share the commitment to symbolic cognitive state representation with almost all of AI (knowledge representation) and Cognitive Science (the Language of Thought hypothesis).

Our coffee-making robot would need not only cognition as these architectures model, but they would also need to perceive and perform motor activities. The relationship of cognitive architecture as currently conceived to perception and motor systems is given in Figure 1. Together, the boxes on the right within the gray polygon correspond to an architecture such as Soar or Act-R. The perception modules supply information about the world in the predicate-symbolic form, and the Action module executes an action specification in the predicate symbolic form, such as Move(A, Table), produced by cognition. The perception and action modules are of course essential in this way for the agent to work in the world, but they don't do any "thinking." That is performed by cognition using predicate-symbolic representations.

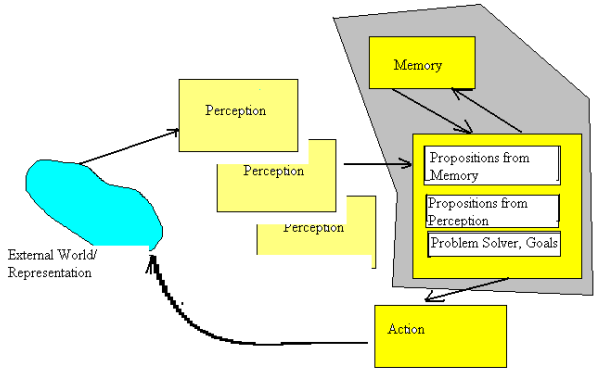

Fig. 1. In the current frameworks, Perception and Action are modules external to Cognition. They do not participate in thinking.

In contrast, consider the phenomenology of our inner selves. We often solve problems imagining spatial situations and performing what feel like internal perceptions, such as in the problem, "Imagine taking a step forward, a step to the right, and a step back. Where are you with respect to the starting point?". Most of us experience "seeing" in an internal image that the final point is one step to the right of the starting point. This phenomenology is independent of the controversy surrounding the "true" nature of mental images. The logic of problem solving is as if a perception is performed on an image Similarly, a musical composer might solve problems in composition by "hearing" and modifying mental auditory images. In fact, Beethoven is said to have composed a symphony after he became deaf – the problem solving involved in composition must have involved internal auditory images. Deciding if one could make it through a narrow restricted passage, such as a bent tube, requires an internal kinesthetic image and manipulating it. In short, in this and similar examples, a perceptual representation, distinct from a predicate-symbolic representation, seems to play a role in thinking, not just providing information about the external world. Our coffee-making robot might have a need for similar internal images to help its thinking.

Not only is cognitive state multi-modal as described above, memory often needs to support such multi-modality as well. Asked if John was standing closer to Bill than to Stu during an episode the previous night at a party, we might recall an image of their relative locations – something akin to a schematic diagram – and *construct* the answer from the diagram, rather than "retrieve" it from memory. We may not especially have noticed the relative closeness at the time of the episode and stored in memory a symbolic

representation such as Closer-to(John: Bill, Stu). In fact, a diagrammatic memory might answer a variety of questions that might not have been specifically anticipated, such as "Was John close enough to Bill to have been able to whisper to him?." A diagrammatic memory component is potentially able to support the generation of a large number of predicate-symbolic representations, including relational predicates not defined at the time of the episode. Again, our robot could benefit from having such a memory, e.g., maps of location of the supermarket.

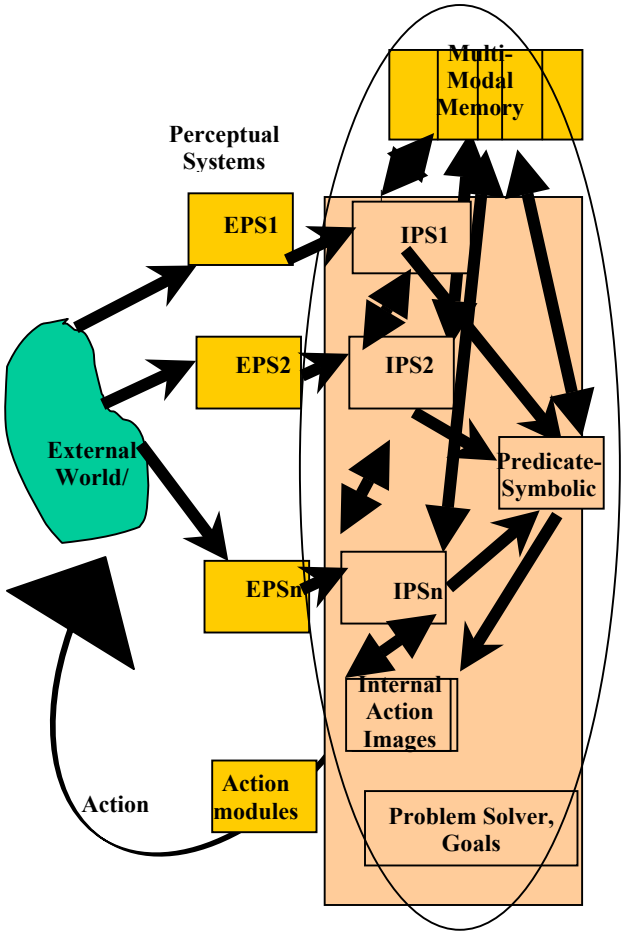

Fig. 2. Schematic of Multi-Modal Cognitive State proposal. The perceptual and kinesthetic modalities are part of the cognitive states as internal images that take part in thinking.

Multimodal Cognitive State

What follows is a highly schematic outline of a proposal for a multi-modal cognitive state and associated mechanisms.

To motivate the ideas, let us look at Figure 2. The boxes on the right under the oval together constitute the augmented state, and associated systems. For each modality, IPS is the component that supports the internal image. The images in IPS can be created in two ways, by composing elements from memory (as when we imagine "an elephant eating a banana"), and when the agent perceives the external world, i.e., as output of EPS. The relational perception operators (such as "one step to the right of") are applied to the images – whether they arise from perceiving the external world or from memory-based operations.

The term "image" to refer to the content of IPS may be misleading – they are not the same as the images that are incident at the input of the perceptual modality, e.g., retinal image for vision, or the spatio-temporal pressure waves at the input to the ear. Instead, these are the *outputs* of the perceptual system. This output supports the perceptual experience in the relevant modality, such as the experience of spatially extended shapes in vision, of sounds in the auditory domain, and so on. Recognition (categorization) of this experience into an object category ("a peacock") is a symbolic output, as is assertion of relations ("block A inside box B"), both of which are produced by operations on IPS, to a first approximation. The reader might still be mystified about what makes IPS a category apart from the traditional symbolic representations. I'll give an example from the diagrammatic domain later, but for now, think of it as the content of the perceptual experience of a person who is looking at a Henry Moore-like abstract sculpture of shapes, or listening to a cascade of sounds. While this person may have linguistic thoughts associated with his experience, the experience is *not reducible* to his linguistic thoughts. After all, one needs to listen to, not just read about, music to experience it. Similarly, the experience is not one of a retinal image of intensities or of pressures on the eardrum – early perception has organized these into perceptual experience of shapes and sounds.

Change of cognitive state. The process of thinking entails changes in cognitive state, in a goal-directed manner. In ACT-R and Soar, cognitive state changes by virtue of rule or operator applications to predicate-symbolic state representations. We may identify this with *inference*, not in the sense of logical inference, but that under the right conditions of matching of information, new information in the form symbol structures can be added. When we have IPS's as cognitive state components, we have additional ways to change cognitive state. One of these is application of a perceptual operator to the contents of an IPS. Thus, if the visual IPS consists of a diagram corresponding to one step forward, one step to the right and one step back, a perception operator can extract the information that the end point is one step to the right of the starting point. This is propositional information that can be added to the symbolic part, changing cognitive state. Conversely, symbolic contents can create or change an IPS. For example, if we now add the information that the person took one step to the left, the IPS is updated with a new diagrammatic element of a line from the previous end point to the starting point. Performing this modification to IPS would require knowledge in the form of an appropriate

diagrammatic element in LTM. The change in the IPS thus changes cognitive state. In general, a change in one of the components can give rise to changes in other components, by *associative evocation*. For example, the word "apple" might evoke the shape of an apple in the visual IPS and a crunching sound in the auditory IPS. Some of these evocations will die out without being useful, but they are available in case they can help in the next steps in problem solving.

Diagrammatic Representation and Reasoning

To bring this discussion down to earth, let us consider a concrete implementation we have done of a bimodal architecture (work currently being done in collaboration with Unmesh Kurup), where the traditional predicate-symbolic component is augmented with a diagrammatic component.

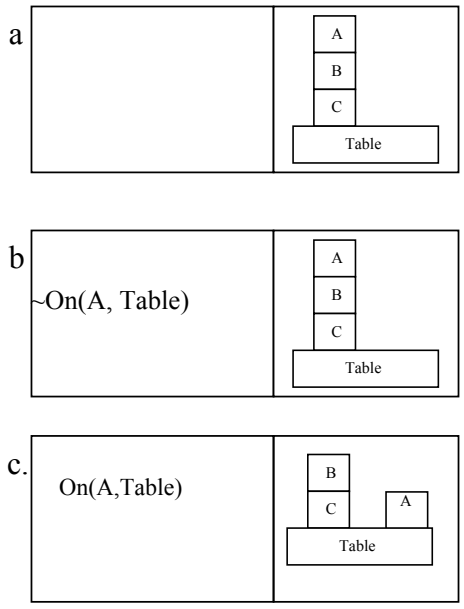

Fig. 3. A series of bimodal states.

DRS (Chandrasekaran, 2004) is the name given to the representation system we have built for representing the diagrammatic component of cognitive state. We only consider diagrams that are configurations of diagrammatic objects each of which is a point, curve, or a region object. These diagrams are not the marks on a piece of paper, i.e., these are not just an array of pixels. The diagram in DRS corresponds to the stage in perception where figure-ground separation has been performed, i.e., the external image has been organized into diagrammatic objects. DRS consists of a set of internal labels for objects and a complete spatial specification of the objects in a convenient form. All that is required is that the spatial specification is available for a set of perceptual routines that take as arguments the appropriate number of diagrammatic objects and return the results, e.g., Inside(point p, region R). Action routines produce diagrammatic objects satisfying given constraints, e.g., curve(point p1, point p2), not-intersect(Region R), which will produce a curve from point p1 to p2 such that it does not go through R. The outputs of the perceptual routines and the constraints for action routines are in the predicate-symbolic form, as in the examples given.

The diagrammatic component is represented as a collection of regions. (The alphabetic symbols are attached to the DRS elements, the spatiality of the alphabetic characters is not part of DRS as it is part of the physical diagram.)

Thus, in a Blocks world problem, a state might look as in Figure 3a. Note that in this example only the diagrammatic component contains a representation. Suppose, as part of problem solving, it was necessary to know if A is on the Table. Perception ON(A,Table) applied to the diagrammatic part will return a negative answer, and the next state would look as in Figure 3b. Suppose now the problem requires that A be on the table, and Move(A, Table) is given as constraint to the action routine, which executes it. The next state might look as in Figure 3c.

As a side note, an advantage of this representation is that the symbolic component doesn't need to be complete at any point; the information needed can be obtained, if it is available by applying perception to the diagrammatic component. *With respect to the spatial aspects of the problem*, the diagrammatic component is complete in a way that the symbolic component is not, and cannot be. In fact, there is no real reason to carry the complete set of symbolic descriptions from state to state. For situated problem solving, the agent can depend on the external world, and the corresponding internal images in the cognitive state, to significantly reduce the complexity of representation.

DRS is a Symbol Structure *and* a Perceptual Representation. DRS has some of the attractive properties of symbol structures, specifically compositionality. We can imagine Block A on the Table by composing the region object corresponding to Block A with the region object for the Table, and placing the former region above and in touch with the latter region. On the other hand, DRS is not a pure symbol structure with only syntactic relations between them. The region objects, the spatial extents of Block A and the Table, *are* a good part of the semantics of the symbols. The fact that the result of composition is itself a spatially fully specified configuration means that perception operations can be applied on it. In a sense, we are having our cake and eating it, since a traditional objection in philosophy and psychology to the idea of mental images was that images are fundamentally different from symbol systems, and the objectors couldn't see how images can have the compositionality property that linguistic symbols have.

Learning. In principle, the same learning mechanisms as used in Soar and ACT-R can be used to learn the DRS components as well. The symbol structure corresponding to relevant parts of DRS can be stored in LTM, along with the parameters that can be used to specify the way some primitive shapes may be put together to generate the shape of each object. For 3-D shapes, a number of alternative families of primitives have been proposed: Marr and Nishihara's generalized cylinders and Biederman's primitive shapes are two examples. There are many unsettled issues in the specifics of learning that require further research, but the general forms of the solution are becoming clear.

Other Modalities. DRS provides a feel for the type of representations for other modalities, but first we need to discuss what other modalities actually exist. In addition to sensory modalities such as audition and touch, it appears that at least two forms of sensory modality-independent spatial representation exist. One is egocentric, a sense of space in which we have a more accurate sense of objects near us than of those farther away. Such an experience of space is essential for us to navigate the physical world without hurting ourselves. The second is an abstract sense of space, such as mental maps that we use to reason about routes. They are sensory modality-independent because even though the visual modality plays a large role in constructing these spatial representations, other modalities, such as audition, and even the kinesthetic one, can help construct such models. For example, we often use the direction of sound or extend our hands to try to touch nearby objects in the dark, in order to construct a model of the immediate space around us.

How Multimodality Benefits Agents

In situated cognition, access to the external world obviates the need to carry around in one's short-term memory information about the world and reason about changes – to the extent changes are made in the world, the world is its own representation and the consequences can be picked up by perception from the world. This feature of situatedness can be modeled by architectures such as in Fig. 1 – the perception modules can be accessed to get the information. However, when we later need specific information about events we experienced, being able to store the memory in something like a perceptual form, recalling the perceptual form later, applying internal perceptions and answering specific questions can provide economy of storage, since an appropriate perceptual abstraction can stand for a potentially large number of propositions, as discussed earlier. In addition to the economy issue, a perceptual representation in memory has the additional advantage that it may be used to answer queries about relations that were defined to the agent after the time of experience, so she could not possibly extract those propositions at that time to store away in memory.

The real benefits come during reasoning without access to the external world, i.e., reasoning by imagining alternatives in the problem space, just as traditional all-symbolic problem solvers do, but where the imagined states have perceptual components. This is what a composer does as she explores the design space of the composition – she needs to experience how a piece of the composition might sound, how a modification of the score might improve it, and so on. A painter has to imagine to some degree the intended painting in his mind's eye. In problem solving involving spatial components, or elements for which spatial analogs exist, the problem solver similarly has to imagine, possibly as a schematic diagram, alternate possibilities, and assessing such states would require applying internal perceptions. Not all this can be done purely symbolically – purely symbolic descriptions of perceptual representations involve qualitative abstractions of quantitative information, and such qualitative abstractions throw away information that may be needed for perceptions. Given the locations of three individuals on a surface, no qualitative abstraction of the locations or relations will suffice to answer all the possible questions about the relative locations. If we abstract the original information as, e.g., Left(A,B) & Left (B,C), we won't be able to answer questions such as, "Is A more to the left of B than B is to C?."

The power of human cognition arises at least partly from the seamless integration of language-like thinking based on symbolic abstractions that transcend perceptual modalities, and efficient but modality-restricted perceptual representations and processes. This role of perceptual representations in the process of thinking is also suggestive of the evolutionary development of human-level cognition as built on top of perceptual machinery.

Related Proposals

Two significant proposals, one from psychology and the other from neuroscience, are related to the proposal that I make. The first is Barsalou (1999) on perceptual symbol systems. The second is the work by Damasio (1994), in which he locates the basis of thinking on perceptual and body images, which are in turn realized in biological systems as neural activation patterns. Neither of the proposals is, or is intended to be, computational, i.e., unlike my proposal it is hard to directly turn these proposals into AI system implementations. Nevertheless, there are many points of contact and reverberations between my and their proposals.

Concluding Remarks

A question that was used to motivate submissions to this symposium asked if and how AI can benefit from research in cognitive science, or more generally, biological cognition. Echoing others who have made similar observations, I noted that building an embodied AI that has

to interact with humans and the physical world and perform complex tasks over a wide range and in real time would force AI to confront certain issues that it has been able to avoid with its Turing-inspired focus on abstract thinking. The coffee-making robot I described earlier would be a goal that might well replace Turing's Test as a driving challenge for AI research. What can biological cognition, or more generally biological embodied cognition, tell us about how one might go about designing such a robot?

Based on the phenomenology of the content of human thought, I proposed that one step in that direction is to think of thinking less as the pure domain of linguistic-symbolic representations and processes and more one where perceptual and body representations play a more direct role in the production of thought and memory. Specifically, I proposed that the notion of cognitive state be generalized to a multi-modal representation, in which linguistic symbolic content is one of the "images," along with images in the various perceptual and kinesthetic modalities. I provided some arguments for why and how this might help, and gave additional details of a bi-modal state, where the perceptual modality is diagrammatic. As a proposal for aspects of cognitive architecture, the proposal spans AI and Cognitive Science.

I hasten to add that I don't pretend that the multi-modal cognition proposal is remotely adequate to build the coffee-making robot. A number of other ideas on the interface between cognition, perception and motor processes would be needed. One source of ideas is neuroscience. What is needed is not mimicking of the details of biological implementation, but to abstract from these details computational principles that apply to a larger class of embodied systems.

Here's a simple example to illustrate this point. Our coffee-making robot would need to do the equivalent of grasping things as part of its daily routine. The robot might not have a vision system similar to ours, nor a hand with a structure similar to our hand, but it would need to have appendages with which to pick up things and some perception system to locate and identify the objects in the environment. So, at what level of abstraction can the study of the biological system help in the design of our robot?

Colby (1999) provides evidence for four distinct spatial reference frames used in parietal cortex for performing actions in space: Grasp-related, Arm-centered and reaching-related, head-centered, and eye-centered spatial representations. One further learns that in the parietal cortex specific action-oriented outputs of perception are directly connected to motor processes, completely bypassing central cognition. For a robot designer, the relevant computational abstraction of this biological trick for real-time performance is relates to the need to avoid central cognition becoming the bottleneck for control. The designer needs to analyze the aspects of shape and space that need to be monitored by the motor processes, and let the robot's perception system generate this information independent of the perceptual experience it provides to cognition, and let the motor processes access this information directly. This way of stating the computational abstraction transcends the specifics of the human visual and motor systems, but states a general principle. My belief is that studies of biological systems can result in many more such abstract principles to add to the bag of AI design tricks.

Acknowledgments

This paper was prepared through participation in the Advanced Decision Architectures Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory under Cooperative Agreement DAAD19-01-2-0009. I acknowledge the contributions of Bonny Banerjee and Unmesh Kurup in helping implement the proposal for the bi-modal case involving diagrammatic representations.

References

Anderson, J.R. and C.J. Lebiere, 1996, *The Atomic Components of Thought*. 1998: Lawrence Erlbaum AssociatesBarwise, J. and J. Etchemendy, Heterogeneous Logic, in *Logical Reasoning with Diagrams*, G. Allwein and J. Barwise, Editors. Oxford University Press: New York. p. 179-200.

Barsalou, L. W., 1999. "Perceptual symbol systems." *Behavioral and Brain Sciences* **22**: 577-660.

Brooks, R.A. "A robust layered control system for a mobile robot," *IEEE Journal of Robotics and Automation*, vol. 2, pp. 14-23, 1986.

Chandrasekaran, B, Unmesh Kurup, Bonny Banerjee, John R. Josephson and Robert Winkler. 2004. "An Architecture for Problem Solving with Diagrams," in *Diagrammatic Reasoning and Inference*, Alan Blackwell, Kim Marriott and Atsushi Shomojima, Editors, Lecture Notes in Artificial Intelligence 2980, Berlin: Springer-Verlag, 2004, pp. 151-165.

Colby, Carol L. "Parietal cortex constructs action oriented spatial representations," in *The Hippocampal and Parietal Foundations of Spatial Cognition*, N. Burgess, K. J. Jeffery, J. O'Keefe (Eds.), pp. 104-126, Oxford University Press, 1999.

Damasio, Antonio R., *Descartes' Error: Emotion, Reason, and the Human Brain*, ISBN: 0399138943, Putnam Publishing Group, 1994.

Newell, A., 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.