Modeling Meta-Cognition in a Cognitive Architecture

Ron Sun (rsun@rpi.edu)
Cognitive Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

Xi Zhang (xzf73@mizzou.edu)
Department of CS, University of Missouri, Columbia, MO 65211, USA

Robert Mathews (psmath@lsu.edu)
Psychology Department, Louisiana State University, Baton Rouge, LA 70803, USA.

Abstract

This paper describes how meta-cognitive processes may be captured within a cognitive architecture CLARION. Existing cognitive architectures often lack built-in meta-cognitive mechanisms. However, meta-cognitive processes are important, in that they are an essential part of cognition and without them, cognition may not function properly. We contend that meta-cognitive mechanisms should be an integral part of cognitive architectures, and thus they have been developed as a part of CLARION. It is demonstrated how human data of meta-cognitive experiments may be simulated using CLARION. The simulations show that meta-cognitive processes can be adequately captured within the CLARION framework.

Introduction

According to Flavell (1976), meta-cognition refers to "one's knowledge concerning one's own cognitive processes and products or anything related to them." Meta-cognition includes "the active monitoring and consequent regulation and orchestration of these processes in relation to the cognitive objects or data on which they bear, usually in the service of some concrete goal or objective."

Moreover, meta-cognitive processes have often been portrayed as explicit processes that involve deliberate reasoning (Metcalfe and Shimamura 1994, Mazzoni and Nelson 1998). However, recently, evidence has been mounting that meta-cognitive processes may not be entirely explicit. For example, Reder and Schunn (1996) argued that there were likely implicit processes, for the simple reason of avoiding using up limited cognitive resources (such as attention) and interfering with regular processes. Thus, they argued that, while strategies themselves might be explicit, and/or explicitly learned, the selection (and use) of strategies was implicit. We have reasons to believe that meta-cognitive knowledge is neither necessarily explicit, nor necessarily implicit. Meta-cognition is likely a combination of implicit and explicit processes, the same as regular cognitive processes, as has been argued amply before (Sun 1999, 2002, Sun et al 2001, Reber 1989).

In relation to computational cognitive modeling, it is worth noting that in existing cognitive architectures, there usually lack built-in meta-cognitive mechanisms. However, meta-cognitive mechanisms are important. They are an essential part of cognition, and without them, cognition may not function properly (Metcalfe and Shimamura 1994, Mazzoni and Nelson 1998). Therefore, we believe that meta-cognitive mechanisms should be an integral part of cognitive architectures.

In this paper, we will develop a theoretical framework of meta-cognition in the context of an overall architecture of the mind — the CLARION cognitive architecture. The architecture is then used to construct models of specific meta-cognitive processes, which are then used to capture experimental data related to meta-cognition. Such simulations serve to validate the models (to some preliminary extent).

Meta-Cognitive Experiments and Data

In the task of Metcalfe (1986), subjects were given a sheet of paper that described a story. They were asked to solve the puzzle in the story. They were told to write down a number between 0 and 10, whereby 0 meant that they were "cold" about the problem (i.e., they had no idea at all about the solution) and 10 meant that they were certain that they had the right solution. They were supposed to do so every 10s at the sound of a click. When the subjects had achieved a solution, they were to write it down on a piece of paper. 134 subjects (undergraduate introductory psychology students) were tested.

43 subjects got the right solution and 44 subjects came up with wrong answers. In general, subjects who came up with the correct solution gave lower warmth ratings than did subjects with incorrect solutions. ANOVA (correct $\times$ incorrect) showed $F(1,50) = 2.81, p = 0.09$, when the last three ratings were used. However, if we only look at the last two warmth ratings before reaching a solution, this effect was significant, $F(1,50) = 6.48, p < 0.5$. If we only look at the last warmth rating before reaching a solution, this effect was also significant, $F(1,68) = 15.00, p < 0.5$.

Warmth rating reflects meta-cognitive monitoring — keeping an eye on one's own cognitive processes. However, the difference in warmth rating is highly counter-intuitive — we would normally expect that subjects who came up with the correct solution gave higher warmth ratings than did subjects with incorrect solutions, but the result was the exact opposite. The question is how this result should be explained; in particular, we would want to know how this result should be explained mechanistically (computationally), within the general framework of a cognitive architecture.

Meta-Cognition in CLARION

The Overall Architecture. Overall, CLARION is an integrative architecture, consisting of a number of distinct subsystems, with a dual representational structure (implicit and explicit) in each. Its subsystems include the action-centered subsystem (the ACS), the non-action-centered subsystem (the NACS), the motivational subsystem (the MS), and the meta-cognitive subsystem (the MCS). The role of the ACS is to control actions, regardless of whether the actions are for external physical movements or internal mental operations. The role of the NACS is to maintain general knowledge, either implicit or explicit. The role of the MS is to provide underlying motivations for perception, action, and cognition, in terms of providing impetus and feedback (e.g., indicating whether outcomes are satisfactory or not). The role of the MCS is to monitor, direct, and modify the operations of the ACS dynamically as well as the operations of all the other subsystems.

Each of these subsystems consists in turn of two levels of representation (a dual representational structure): Generally, the top level encodes explicit knowledge and the bottom level encodes implicit knowledge. Notice the fact that the relatively inaccessible nature of implicit knowledge may be captured by subsymbolic, distributed representation provided, for example, by a backpropagation network (Sun 2002). This is because distributed representational units in the hidden layer(s) of a backpropagation network are capable of accomplishing computations but are subsymbolic and generally not individually meaningful (Sun 1994). This characteristic of distributed representation, which renders the representational form less accessible, accords well with the relative inaccessibility of implicit knowledge (Reber 1989, Seger 1994, Cleeremans et al 1998). In contrast, explicit knowledge may be captured in computational modeling by symbolic or localist representation, in which each unit is more easily interpretable and has a clearer conceptual meaning (Sun 1994). This characteristic of symbolic or localist representation captures the characteristic of explicit knowledge being more accessible and more manipulable (Sun 1994). [1]

Figure 1 contains a sketch of this basic architecture of a cognitive agent, which includes the four major subsystems. The following three subsections will sketch, one by one and in some more detail, these subsystems of CLARION.

The Action-Centered Subsystem. The overall algorithm for action decision making in the action-centered subsystem (the ACS) of CLARION is as follows, where the bottom level is named the IDNs (the Implicit Decision Networks) and the top level the ARS (the Action Rule Store):

1. Observe the current state x.
2. Compute in the bottom level (the IDNs) the "value" of each of the possible actions (a_i's) associated with the state x: $Q(x, a_1)$, $Q(x, a_2)$, ..., $Q(x, a_n)$. Stochastically choose one action according to Q values.

[1]See Sun (1994, 2002) for full arguments.

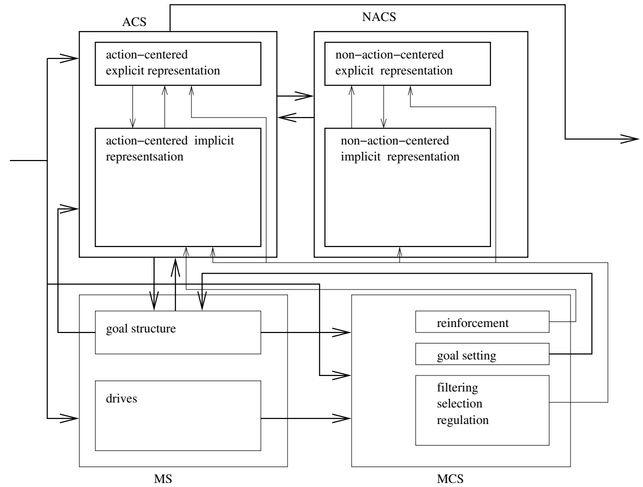

Figure 1: The CLARION architecture.

3. Find out all the possible actions (b_1, b_2, ..., b_m) at the top level (the ARS), based on the current state x and the existing rules in place at the top level. Stochastically choose one action.
4. Choose an appropriate action a, by stochastically selecting the outcome of either the top level or the bottom level.
5. Perform the action a, and observe the next state y and (possibly) the reinforcement r.
6. Update the bottom level in accordance with an appropriate algorithm based on the feedback information.
7. Update the top level using an appropriate algorithm (for extracting, refining, and deleting rules).
8. Go back to Step 1.

In each network (encoding implicit knowledge), actions are selected based on Q values. A Q value is an evaluation of the "quality" of an action in a given state: $Q(x, a)$ indicates how desirable action a is in state x. At each step, given the state x, we compute the Q values of all the actions (i.e., $Q(x, a)$ for all a's). We then use the Q values to decide probabilistically on an action to be performed, through a Boltzmann distribution of Q values (this method is also known as Luce's choice axiom; Watkins 1989). The *Q-learning* algorithm (Watkins 1989), a reinforcement learning algorithm, is used for learning Q values (i.e., learning implicit knowledge at the bottom level). Q values are gradually tuned, on-line, through successive updating, to enable sequential behavior to emerge. Q-learning is implemented in backpropagation networks (Sun 2002, 2003). [2]

Next, explicit knowledge at the top level (the ARS) is captured by action rules. An action rule indicates an action to be performed in a particular state.

Action rules may be learned in a variety of ways. For example, to capture a bottom-up learning process (Karmiloff-Smith 1996, Sun 20002), the *Rule-Extraction-Refinement* algorithm (RER) learns action rules at the top level using information in the bottom level. The basic idea of bottom-up learning is as follows: If an

[2]As Q-learning is not directly relevant to this work, see Sun (2003) for further details.

action chosen (by the bottom level) is successful (i.e., it satisfies a certain criterion), then an explicit action rule is extracted. Then, in subsequent interactions with the world, the rule is refined by considering the outcome of applying the rule: If the outcome is successful, the condition of the rule may be generalized to make it more universal; if the outcome is not successful, then the condition of the rule should be made more specific and exclusive of the current case. An agent needs a rational basis for making these above decisions, which is based on *information gain* (Sun 2002, 2003). The details of the operations used in the algorithm (including extraction, generalization, and specialization) and the numerical criteria measuring whether a result is successful or not (used in deciding whether or not to apply some of these operators) can be found in Sun (2002, 2003). (Since they are not directly relevant to this work, we will not get into the details.)

For integrating the outcomes of the two levels, at each step, with probability P_{TL}, we use the outcome from the top level. With the probability $P_{BL}(= 1 - P_{TL})$, we use the outcome of the bottom level. The selection probabilities may be variable, determined through a process known as "probability matching": that is, the probability of selecting a component is determined based on the relative success ratio of that component.

The Non-Action-Centered Subsystem. At the bottom level of the NACS, the "associative memory" networks (the AMNs for short) encode implicit non-action-centered knowledge. Associations are formed by mapping an input to an output. The regular back-propagation learning algorithm can be used to establish such associations between pairs of inputs and outputs (Rumelhart et al 1986).

On the other hand, at the top level of the NACS, a general knowledge store (the GKS) encodes explicit non-action-centered knowledge (cf. Sun 1994). In this network, concepts are represented by *chunks*, which are specified through dimensional values (as in the ACS). Links between chunks encode associations between pairs of chunks (concepts), known as *associative rules*.

On top of rule-based reasoning, similarity-based reasoning may be employed in the NACS. An agent may compare a known (given or inferred) chunk with another chunk. If the similarity between them is sufficiently high, then the latter chunk is inferred.

The Motivational Subsystem. The motivational subsystem (the MS) is concerned with drives and their interactions (Toates 1986). That is, it is concerned with why an agent does what it does. Simply saying that an agent chooses actions to maximizes gains, rewards, or payoffs leaves open the question of what determines gains, rewards, or payoffs. The relevance of the motivational subsystem to the main component, the ACS, lies primarily in the fact that it provides the *context* in which the goal and the payoff of the ACS are set. It thereby influences the working of the ACS, and by extension, the working of the NACS. Since this part is not directly relevant to this work, we will not get into more details (see Sun 2003 for further details).

The Meta-Cognitive Subsystem. On the other hand, meta-cognition taps into one's knowledge concerning one's own cognitive processes and outcomes. It involves active monitoring as well as regulation/control of these processes, usually in the service of some goal (Flavell 1976). This notion of meta-cognition is operationalized within CLARION. The meta-cognitive subsystem (the MCS) monitors and controls/regulates cognitive processes for the sake of improving cognitive performance. Control and regulation may be in the forms of setting goals for the ACS (on the basis of motivational states), setting essential parameters of the ACS and the NACS, interrupting and changing on-going processes in the ACS and the NACS, and so on. Control and regulation may also be carried out through setting reinforcement functions (on the basis of motivational states).

Specifically, in this subsystem, several types of meta-cognitive processes are available, for different meta-cognitive monitoring and control/regulation purposes. Among them, there are the following types:

(1) behavioral aiming:
setting of reinforcement functions
setting of goals

(2) information filtering:
focusing of input dimensions in the ACS
focusing of input dimensions in the NACS

(3) information acquisition:
selection of learning methods in the ACS
selection of learning methods in the NACS

(4) information utilization:
selection of reasoning methods in the ACS
selection of reasoning methods in the NACS

(5) outcome selection:
selection of output dimensions in the ACS
selection of output dimensions in the NACS

(6) cognitive mode selection:
selection of explicit processing, implicit processing, or a combination thereof (with proper integration parameters), in the ACS

(7) setting parameters of the ACS and the NACS:
setting of parameters for the IDNs
setting of parameters for the ARS
setting of parameters for the AMNs
setting of parameters for the GKS

Structurally, the subsystem may be subdivided into a number of modules. The bottom level consists of the following (separate) networks: the goal setting network, the reinforcement network, the input selection network, the output selection network, the parameter setting network (for setting learning rates, temperatures, etc.), and so on. In a similar fashion, the rules at the top level (if they exist) can be correspondingly subdivided. See Figure 2 for a diagram of the MCS.

The monitoring buffer contains several sections of information: the ACS performance section, the NACS performance section, the ACS learning section, the NACS

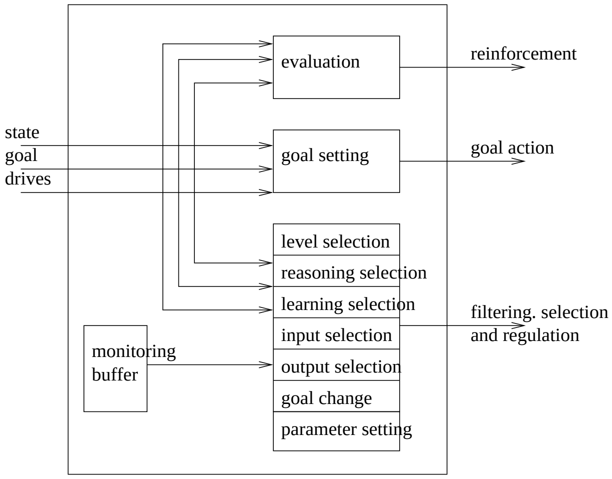

Figure 2: Structure of the meta-cognitive subsystem.

learning section, and other sections. Each section contains information about both the bottom level and the top level of a subsystem. Most relevant to this work, in each "performance" section, the information about a subsystem includes not only the strengths of the top conclusions, but also the *relative strengths* of the top conclusions, which concern how distinguished or certain the top conclusions are in relation to other competing ones. [3] A relative strength is defined as:

$$RS_i = \frac{S_i}{\sum_j S_j} \quad (1)$$

where S stands for strength, and RS for relative strength.

Other aspects of the MCS, such as setting reinforcement functions, setting goals, information selection, and so on, can be found in Sun (2003). This subsystem may be pre-trained before the simulation of any particular task (to capture evolutionary pre-wired instincts or previously acquired knowledge and skills).

Simulations of Meta-Cognitive Data

In the simulation, we model meta-cognitive monitoring, and thereby provide a detailed computational explanation of the (counter-intuitive) experimental results of Metcalfe (1986).

Model Setup. The explanation of the experiment of Metcalfe (1986) on which this simulation was based was that when a subject came up with multiple plausible explanations and had to evaluate their relative merits, his/her subjective certainty on the conclusions (a meta-cognitive judgment) was relatively low due to the co-existence of multiple plausible explanations. Hence a lower warmth rating was produced. But, in this way, the subject was more likely to come up with a correct (the most plausible) explanation eventually. On the other hand, when a subject came up with only one plausible explanation, there was no need to evaluate multiple possibilities, and thus his/her subjective certainty, and also his/her "warmth" rating, was higher, but that sole explanation was more likely wrong, because of the ambiguity of the situation and the lack of careful evaluation of all possibilities on the part of the subject (Metcalfe 1986).

In Clarion, both the ACS and the NACS were involved in the simulation of this task. The NACS performed inference under the control of the ACS. Through the monitoring buffer, the MCS monitored the progress of inference in the NACS and performed meta-cognitive control when needed.

More specifically, the goal of performing "regular inference" was set up first by the MS (before it all began). The MCS then selected relevant input dimensions to be used in reasoning in the NACS, which excluded all other information not relevant to the task at hand (for example, there was contextual input information, such as time and location, that was not relevant to the task). The MCS also selected reasoning method to be used in the NACS, in this case, "forward chaining with similarity based reasoning".

Among other things, the monitoring buffer in the MCS kept track of how clear-cut conclusions reached by the NACS were. The NACS section of the buffer recorded the relative strengths of the n most highly activated conclusions (as explained earlier). When that part of the buffer reported that there was one conclusion that stood out with a high relative strength, the conclusion was considered certain and its "warmth" level was correspondingly high. Otherwise, the conclusion was less certain, and the "warmth" level was lower. Hence, "warmth" was captured in this simulation by relative strength.

The ACS directed the reasoning of the NACS. The following action rules were implemented in the top level of the ACS for this purpose:

If goal= regular-inference, then perform one-step of inference in the NACS (using method selected by the MCS and information filtered by the MCS).

If goal= regular-inference, and chunk i is a conclusion chunk with $RS_i > threshold_{RS}$ and $\forall_j S_i > S_j$, then retrieve chunk i and report it.

If goal= warmth-reporting, then report the "warmth" of the chosen chunk from the monitoring buffer in the MCS.

RS stood for relative strength, and S for strength. The threshold for relative strengths was set at $threshold_{RS} = 0.5$.

Although the IDNs (the bottom level of the ACS) were present, they had very little effect. This was due to the stochastic selection of levels in favor of the top level (due to the setting of the cross-level integration parameters), which was the result of the task instructions, which led to performing a rather explicit inference task.

At the top level of the NACS (the GKS), relevant knowledge was encoded as associative rules. Some subjects (those who turned out to have higher warmth ratings) had few of these rules, while other subjects (those who turned out to have lower warmth ratings) had more of these rules. In relation to the domain of this experiment, rules were in the following form:

[3] Details of other sections can be found in Sun (2003).

If event A1 happens, then B11 might be the answer
If event A1 happens, then B12 might be the answer
If event A2 happens, then B21 might be the answer
If event A2 happens, then B22 might be the answer

At the bottom level of the NACS, one AMN was present. The network was trained with the same knowledge as embodied by the associative rules in the GKS.

Simulation Results. In our simulation, in general, those variants of the models that generated the correct solution gave lower warmth ratings than those that generated incorrect solutions. Thus, the simulation model, as explained earlier within the general framework of CLARION, accounted for the counter-intuitive findings in the experimental data of Metcalfe (1986).

In the simulation data, during the last two episodes, there were significant differences between the two groups of simulated subjects. At the last episode, the average warmth rating of the simulated subjects with correct solutions was 3.327, while that of the simulated subjects with incorrect solutions was 5.241. At the episode preceding the last, the average warmth rating of the simulated subjects with correct solutions was 3.299, while that of the simulated subjects with incorrect solutions was 5.085.

An ANOVA of the warmth data of the last episode showed that there was a significant difference between correct versus incorrect ($F(1, 298) = 92.808, MSE = 273.196, p < 0.0001$). Similarly, an ANOVA of the warmth data of the penultimate episode showed that there was also a significant difference ($F(1, 292) = 78.049, MSE = 232.681, p < 0.0001$).

The earlier explanation of the data pattern of this experiment has been confirmed by the simulation: That is, when a subject initially came up with multiple plausible explanations (when multiple relevant rules were available), his/her subjective certainty on the conclusions was low due to the co-existence of multiple plausible explanations. Thus a lower warmth rating was produced. However, the subject in this case was more likely to come up with a correct explanation, based on evaluations of the relative merits of the different explanations.

Discussions

The simulations in this work help us to better understand issues related to meta-cognition, which may well be one of the central issues of cognition and may have significant implications for further advances in cognitive theories and in further development of cognitive architectures (Reder 1996, Anderson and Lebiere 1998, Lovett et al 2000, Sun et al 2001, Sun 2002). Notably, many cognitive architectures do not include a meta-cognitive component (such as ACT-R; Anderson and Lebiere 1998). Thus, work in this area is not only useful but very much needed.

Compared with other existing cognitive architectures, CLARION clearly has much better developed meta-cognitive mechanisms. For example, in ACT-R, meta-cognitive control and regulation are often accomplished through manually adjusting parameter settings (Lovett et al 2000). Although meta-cognitive reasoning processes may be implemented using ACT-R's production rules, there is no built-in mechanism or process specifically designed to capture meta-cognitive processes in a precise way. We can also compare CLARION with SOAR (Rosenbloom et al 1993). In SOAR, like in ACT-R, meta-cognitive processes may be implemented using production rules, but they are not distinguished from regular processes and delimited in terms of their specific characteristics.

In contrast, CLARION has a set of specially designed mechanisms for monitoring, controlling, and regulating cognitive processes. The meta-cognitive subsystem may filter/select information, adjust cognitive parameters, or intervene in regular cognitive processes. Also, in contrast to these two other cognitive architectures, in CLARION, meta-cognitive processes are architecturally specified to a large extent. They are specifically modeled to the extent that we believe is appropriate. They are not totally undelimited, as in the other architectures. But they are also not completely fixed and thus they are not inflexible. As our understanding of meta-cognitive processes grows, these mechanisms in CLARION may be further refined, and tailored to capture the exact range and scope of human meta-cognitive processes.

As indicated by our earlier discussion of simulation results, CLARION, through simulation, succeeded in explaining, computationally, counter-intuitive results in the experimental data of Metcalfe (1986) (which concerned the lower warmth ratings from the subjects who found the correct answers compared with the subjects who failed to do so). The explanation (based on amount of relevant knowledge) naturally fell out of the processes embodied by CLARION. A number of other meta-cognitive simulations have also been carried out that further confirm our conclusions (Sun 2003).

Concluding Remarks

In summary, simulations of meta-cognitive monitoring have been conducted based on the built-in meta-cognitive mechanisms in the cognitive architecture CLARION. The cognitive architecture contains rather detailed descriptions of meta-cognitive mechanisms and thus makes simulations of meta-cognitive processes easier to construct, less ad hoc, and more uniform. That is, meta-cognitive processes are, more or less, architecturally specified in CLARION. We showed that this approach is viable for cognitive modeling.

The afore-described simulations captured rather accurately the experimental data of Metcalfe (1986). To some extent, the afore-described simulation, along with other meta-cognitive simulations using CLARION, validated our approach (as embodied by CLARION).

Acknowledgments

This work is supported in part by Army Research Institute contract DASW01-00-K-0012.

References

J. Anderson and C. Lebiere, (1998). *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, NJ.

A. Cleeremans, (1997). Principles for implicit learning. In: D. Berry (Ed.), *How Implicit Is Implicit Learning?* 195-234. Oxford, England: Oxford University Press.

J. Flavell, (1976). Metacognitive aspects of problem solving. In: B. Resnick (ed.), *The Nature of Intelligence*. Erlbaum, Hillsdale, NJ.

A. Karmiloff-Smith, (1986). From meta-processes to conscious access: evidence from children's metalinguistic and repair data. *Cognition*. 23. 95-147.

M. Lovett, L. Daily and L. Reder, (2000). A source activation theory of working memory: cross-task prediction of performance in ACT-R. *Cognitive Systems Research*, Volume 1, Issue 2, 99-118.

G. Mazzoni and T. Nelson, (eds.) (1998). *Metacognition and Cognitive Neuropsychology*. Erlbaum, Mahwah, NJ.

J. Metcalfe, (1986). Dynamic metacognitive monitoring during problem solving. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 12, 623-634.

J. Metcalfe and A. Shimamura, (eds.) (1994). *Metacognition: Knowing about Knowing*. MIT Press, Cambridge, MA.

T. Nelson and L. Narens, (1990). Meta-memory: a theoretical treatment and new findings. In: G. Bower, (ed.) *The Psychology of Learning and Motivation*. Vol.26, 125-140. Academic Press, New York.

A. Reber, (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*. 118 (3), 219-235.

L. Reder, (ed.) (1996). *Implicit Memory and Metacognition*. Erlbaum, Mahwah, NJ.

L. Reder and C. Schunn, (1996). Metacognition does not imply awareness: strategy choice is governed by implicit learning and memory. In: L. Reder, (ed.) *Implicit Memory and Metacognition*. Erlbaum, Mahwah, NJ.

P. Rosenbloom, J. Laird, and A. Newell, (1993). *The SOAR Papers: Research on Integrated Intelligence*. MIT Press, Cambridge, MA.

D. Rumelhart, J. McClelland and the PDP Research Group, (1986). *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, MIT Press, Cambridge, MA.

C. Seger, (1994). Implicit learning. *Psychological Bulletin*. 115 (2), 163-196.

M. Stadler, (1995). Role of attention in implicit learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 15, 1061-1069.

R. Sun, (1994). *Integrating Rules and Connectionism for Robust Commonsense Reasoning*. John Wiley and Sons, New York, NY.

R. Sun, (1999). Accounting for the computational basis of consciousness: A connectionist approach. *Consciousness and Cognition*, 8, 529-565.

R. Sun, (2002). *Duality of the Mind*. Lawrence Erlbaum Associates, Mahwah, NJ.

R. Sun, (2003). *A Tutorial on CLARION*. http://www.cogsci.rpi.edu/~rsun/tutorial.pdf

R. Sun, E. Merrill, and T. Peterson, (2001). From implicit skills to explicit knowledge: a bottom-up model of skill learning. *Cognitive Science*, 25 (2), 203-244.

F. Toates, (1986). *Motivational Systems*. Cambridge University Press, Cambridge, UK.

C. Watkins, (1989). *Learning with Delayed Rewards*. Ph.D Thesis, Cambridge University, Cambridge, UK.