PHILOSOPHICAL PSYCHOLOGY, VOL. 17, NO. 3, SEPTEMBER 2004

Routledge
Taylor & Francis Group

Desiderata for cognitive architectures

RON SUN

ABSTRACT *This article addresses issues in developing cognitive architectures—generic computational models of cognition. Cognitive architectures are believed to be essential in advancing understanding of the mind, and therefore, developing cognitive architectures is an extremely important enterprise in cognitive science. The article proposes a set of essential desiderata for developing cognitive architectures. It then moves on to discuss in detail some of these desiderata and their associated concepts and ideas relevant to developing better cognitive architectures. It argues for the importance of taking into full consideration these desiderata in developing future architectures that are more cognitively and ecologically realistic. A brief and preliminary evaluation of existing cognitive architectures is attempted on the basis of these ideas.*

1. Introduction

As we have already known, a cognitive *architecture* is the overall, essential structure and process of a broadly-scoped domain-generic computational cognitive model, used for a broad, multiple-level, multiple-domain analysis of cognition and behavior (Newell, 1990; Sun, 2002). A cognitive architecture provides a concrete framework for more detailed modeling of cognitive phenomena, through specifying essential structures, divisions of modules, relations between modules, and a variety of other aspects (Sun, 1999). The analysis of cognition through cognitive architectures is to be performed mainly at the computational level. Cognitive architectures are believed to be essential in advancing understanding of the mind (Anderson, 1983; Anderson & Lebiere, 1998; Newell, 1990; Sun, 2002), and therefore, developing cognitive architectures is an extremely important enterprise in cognitive science.

However, many issues and confusions exist in this field that cry out for serious conceptual clarification, so that further progress can be made. Some of these issues are, for example:

- Basic cognitive assumptions. Right now, almost invariably, each cognitive architecture is based on a radically different set of assumptions, and develops its own world view. Is it possible that we come up with a common set of assumptions and establish a baseline from which different architectures may be developed and compared?

Ron Sun, Cognitive Science Department, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA, email: rsun@rpi.edu

ISSN 0951-5089/print/ISSN 1465-394X/online/04/030341–33 © 2004 Taylor & Francis Ltd
DOI: 10.1080/0951508042000286721

- Essential dichotomies. Various cognitive dichotomies have been proposed before: implicit versus explicit, procedural versus declarative, automatic versus controlled, and so on. What are the essential dichotomies? How should we analyze them in a *process*-based (or mechanistic) way, and thereby develop cognitive architectures on that basis?
- Memory modules. There have been many debates concerning the structure of memory in psychology and cognitive science. It is far from clear what essential subsystems of memory are and thus how memory should be divided up.
- Methodological approaches. Many methodological approaches exist: quantitative data fitting, qualitative demonstration, theoretical arguments, philosophical thought experiments, and so on. What should we rely on in developing cognitive architectures? Some of them? (In that case, which ones?) Or all of them?

Many, many other issues exist as well. See, for example, Newell (1990) and Anderson and Lebiere (2003) for additional issues. In this article, I will focus mainly on structural issues (that is, issues such as cognitive dichotomies, modularity of cognition, memory subsystems, and so on). I hope to clarify some of these issues regarding cognitive architectures, and lay the foundation for future computational model development (Sun, 2002).

In the remainder of this article, Section 2 discusses the general idea of cognitive architectures. Section 3 presents essential desiderata for cognitive architectures, including behavioral and cognitive desiderata. Sections 4 to 9 present a gradually expanding discussion of some of these desiderata and, in the process, develop more detailed considerations and research questions. Section 10 connects these ideas to phenomenological philosophy, which serves as the foundation of the above discussion. Section 11 briefly reviews and critiques existing cognitive architectures based on the desiderata developed thus far. Section 12 summarizes the article.

2. What is a cognitive architecture?

As I stated before, a cognitive architecture is the overall, essential structure and process of a domain-generic computational cognitive model, used for a broad, multiple-level, multiple-domain analysis of cognition and behavior. In particular, it deals with componential processes of cognition in a structurally and mechanistically well defined way. Its function is to provide an essential framework to facilitate more detailed modeling and understanding of various components and processes of the mind. In this way, an architecture serves as an initial set of assumptions to be used for further development. These assumptions, in reality, may be based on either available scientific data (for example, psychological or biological data), philosophical thoughts and arguments, or *ad hoc* working hypotheses (including computationally inspired such hypotheses). An architecture is useful and important precisely because it provides a comprehensive initial framework for further model development in many task domains.

Because of initial assumptions made in an architecture, further development of computational models is constrained in many ways, but is also open to many

possibilities in a variety of other ways. For example, a model may be limited or constrained because of the prior determination of a modular structure, but new possibilities emerge in the area of modular interactions due to the architectural division into modules.

Let us examine some ideas in more detail. We need to distinguish architectures from innate structures of the mind. An innate structure can, but need not, be specified in an initial architecture—it may be specified in more detailed modeling later on (Sun, 2002). An innate structure does not have to be involved in the overall structure/process or the overall division of a computational model; that is, it may reside within an individual module or otherwise in a less conspicuous place. Furthermore, as currently practiced, architectural details may not be innate. For one thing, if one is mainly interested in modeling an adult cognitive agent (in other words, not in modeling developmental processes), an architecture may contain certain structures that are not innate, that have resulted from ontogenetic development under the influence of physical, social, and cultural environments (see Anderson, 1993, for instances). Another case in which some non-innate structures are introduced is that one may choose to model some cognitive tasks in an overly representational way, as often happened in the early days of cognitive science, so that external structures are introduced into a cognitive model as an architectural constraint [1].

Although we need to recognize the differences between architectures and innate structures, too much separation of the two is problematic. Clearly, these above cases of non-innate structures are to be avoided as much as possible. To avoid the pitfalls, we should take a minimalistic approach in architecture development. That is, we should start with a minimal architecture, so as to avoid these non-innate structures. But how should we measure minimality in architecture development?

An architecture can be minimal in two different senses. According to one sense of minimality, an architecture should include only minimal *initial* structures and employ learning as a means for developing further structures upon them, bootstrapping all the way to a full-fledged cognitive model. In so doing, it is important that we are careful to devise only minimal *initial* learning capabilities that are capable of "bootstrapping," in accordance with whatever phenomena that we aim to model (Sun *et al.*, 2001). The other sense of minimality implies reducing the internal structures and representations to a minimal level (not limited to initial structures), while still capturing the phenomena that we intend to model (Bickhard, 1993; Sun, 2000). One way we can accomplish this sense of minimality is through the use of environmental cues, structures, and regularities. Putting it another way, we may place many structures back into the world, instead of placing them (mainly) in the head of an agent (Bickhard, 1993; Hutchins, 1995; Sun, 2000). The avoidance of overly sophisticated initial structures, and thus the inevitable use of learning, may often help to avoid overly representational models (Sun, 2000) (because of learning can help to exploit environmental structures). The avoidance of overly representational models, conversely, often entails a simpler initial structure (since overly representational models are usually complex), as well as the involvement of learning (because otherwise the cost of the manual analysis of complex structures

across the internal and the external world may be prohibitive). We should aim to achieve both of these two kinds of minimality in developing cognitive architectures, and thereby avoid the pitfalls above, so that the resulting architectures contain roughly the innate structure of a cognitive agent.

For example, as will be discussed later, one such minimal structure that is needed is the two-level structuring (derived from the dual-representation hypothesis of Sun, 1994, 2002). I believe that such two-level structuring is minimal and necessary, because there is no other way to come up with such a structure based on currently known (or imaginable) learning methods, and also because it is the basis on which many other structures are built, both innately and developmentally (as will be discussed later).

3. Essential desiderata for cognitive architectures

It is fair to say that research in multiple fields, scattered across multiple scientific disciplines, seems to be converging towards the understanding that the classical treatment of cognition has been an impasse. Rethinking of approaches, methodologies, concepts, arguments, and facts is needed. Such rethinking is evident in a number of "new" approaches that focus on the interaction between a cognitive agent and its world (Bickhard, 1993; Sun, 2000) or the interaction amongst cognitive agents (Hutchins, 1995). From studies of human cognition, motivation, and development, through consciousness, sociality, and language, to artificial intelligence (e.g. cognitive robotics), we are witnessing the resurgence of research whereby the importance of interaction is better appreciated. The time is ripe to emphasize this perspective also in cognitive architecture development (Sun, 2002), and this emphasis is one fundamental aspect of what I want to accomplish with the present article.

First, let us examine a few overall desiderata for cognitive architectures:

- Ecological realism. A realistic cognitive architecture needs to take into account the essential functions of cognitive agents (humans in particular) in their natural environments. Puzzle solving, cryptography, and geometric theorem proving should not be our essential goals because they are not ecologically most important activities for cognitive agents. What we should focus on is *everyday activities* of cognitive agents, in their natural ecological environments (Gibson, 1979; Lorenz, 1950). This is precisely what we mean by ecological realism. (We shall look into characteristics of everyday activities later.) Likewise, we cannot ignore the fact that cognitive agents are always situated in sensory environments, they continuously have to cope with many contingencies, they have concurrent, often conflicting, needs and goals, and they are embodied in physical structures that are limited in their movements, perceptions, and actions.
- Bio-evolutionary realism. A cognitive model of human intelligence should be reducible to a model of animal intelligence. There are reasons to believe that, rather than discontinuous, cognitive processes of animal species and humans

form a continuum. Large discontinuity is not warranted. The existence of large discontinuity may indicate flaws in our very conception of human cognitive processes. Often, in cognitive models, overly elaborate representations and mechanisms dominate that cannot possibly be reduced to existing, or even potentially plausible, models of animal cognition (cf. Rosenbloom *et al.*, 1993). This phenomenon is symptomatic of a serious methodological problem: viewing high-level cognition as paramount to human cognition and thereby ignoring ecological realism as an essential requirement of cognitive modeling (as pointed out before in Sun, 2000, 2002). This requirement complements and supplements ecological realism. We may term this requirement bio-evolutionary realism, as it reflects the natural evolutionary history of biological species leading up to humans (Newell, 1980).

- Cognitive realism. We should aim to capture *essential* characteristics of human behavior and cognitive processes, as we understand them from psychology, philosophy, and neuroscience. However, we should not, and cannot, capture every minute variation in human performance and cognition. Rather, in our development of cognitive architectures, we should attempt to abstract away from details of the voluminous data that have been accumulated in many scientific disciplines relevant to the understanding of cognition, and focus only on fundamental, characteristic traits of human behavior and cognition. There are, in fact, many well-known cognitive and behavioral characteristics that we can identify, which we will discuss in more detail later.
- Eclecticism of methodologies and techniques. Any unnecessary or premature commitment to any specific approaches, methodologies, or paradigms can only be detrimental to the progress of the study of cognition and the development of cognitive architectures. We want to take a broad-based approach, and be as all-encompassing as possible when we evaluate and incorporate prior research results, methods, and techniques. Future cognitive architecture development should take an integrative approach, incorporating as much as possible various prior perspectives, approaches, and results that are productive and useful (Sun, 2002).

In relation to human everyday activities (which I alluded to earlier), let us discuss some behavioral characteristics commonly exhibited in such activities, which we should attempt to capture in cognitive architectures:

- Reactivity. In human everyday activities, behavioral responses are mostly generated without involving elaborate computation (such as comparing all possible alternatives at length). Reactivity of human behavior entails relatively fixed responses, so that an individual does not have to re-compute responses every time a response is needed (Savage, 2003). Such reactivity is also direct and immediate; that is, it is "non-representational" (without involving overly elaborate and overly explicit mediating conceptual representations). For detailed characterization of this aspect of human behavior, see, for example, Agre and Chapman (1990) and Clark (1997).
- Sequentiality. Human everyday activities are mostly sequential: they are carried

out one step at a time, stretched out temporally. Temporal dependencies and structures are essential to such activities and they are the basis of various behavioral responses (Stanley *et al.*, 1989; Sun, 1999; Willingham *et al.*, 1989). Heidegger (1927) provided a philosophical treatment of the temporal nature of human everyday activities.

- Routineness. Human everyday activities are very much routinized and thus largely made of routines, or habitual sequences of behavioral responses. We may look at the matter this way, if we are committed to the assumptions of (1) sequentiality and (2) relatively fixed (or habitualized) reactivity, then we also have to be committed to the assumption of routineness of human activities. However, note the gradual adaptation, or learning, of these routines as well—generally, they are formed gradually and subject to constant modification. Therefore, overall, we may view human everyday activities as consisting of forming, changing, and following routines. See, for example, Heidegger (1927) and Agre and Chapman (1990) for various discussions of routines.
- Trial-and-error adaptation. Learning of reactive routines is mostly, and essentially, a trial-and-error adaptation process. Manifestations of such adaptation have been variously studied under the rubric of law of effect (Thorndike, 1911), classical and instrumental conditioning (Rescorla & Wagner, 1972; Shanks, 1993; Sutton & Barto, 1981), and probability learning (Wasserman *et al.*, 1993). There are reasons to believe that this type of learning is the most essential to human everyday activities and cognition (see Sun, 1999, 2002).

Now, turning to the *cognitive* characteristics of human everyday activities, there is likewise a list of essential characteristics that should be captured in cognitive architectures:

- Dichotomy of implicit and explicit processes. Generally speaking, implicit processes are inaccessible, "holistic," and imprecise, while explicit processes are accessible and precise (Dreyfus & Dreyfus, 1987; Reber, 1989; Smolensky, 1988; Sun, 1994, 1999, 2000). This dichotomy is closely related to some other well-known dichotomies: the dichotomy of symbolic versus subsymbolic processing, the dichotomy of conceptual versus subconceptual processing (Smolensky, 1988), and the dichotomy of the conscious versus the unconscious (Sun, 1999). It can also be justified psychologically, by the voluminous empirical studies of implicit and explicit learning, implicit and explicit memory, implicit and explicit perception, and so on. These empirical dichotomies denote more or less the same thing, and thus they all serve as justifications for a general distinction between implicit and explicit cognition.
- Synergistic interaction. Recently, there have been some emerging indications of synergy between implicit and explicit cognition. I hypothesized (see Sun, 1994, 1999, 2002) that the reason for having the two separate components, the implicit and the explicit, or any other similar combination of components, was that these different systems could (potentially) work together synergistically, supplementing and complementing each other in a variety of different ways.

This is because these two components have qualitatively different characteristics, thus generating better overall results when they are combined (Breiman, 1996). See, for example, Mathews *et al.* (1989), Sun (1999), Sun *et al.* (2001), and Dreyfus and Dreyfus (1987) for more discussions, demonstrations, and arguments in favor of the notion of synergy.

- Bottom–up learning. The interaction between the two sides of the dichotomy with regard to learning includes top–down (explicit learning first and implicit learning later), bottom–up (implicit learning first and explicit learning later), and parallel learning (simultaneous implicit and explicit learning). However, there are reasons to believe that the most important and the most essential in human everyday activities is bottom–up learning. There are various indications of this possibility, including (1) various philosophical arguments, such as Heidegger (1927), Dewey (1958), and Merleau-Ponty (1963), in which the primacy of direct interaction with the world in an implicit way is emphasized; and (2) psychological evidence of the acquisition and the delayed explication of implicit knowledge (for example, Bowers *et al.*, 1990; Karmiloff-Smith, 1986; Mandler, 1992; Siegler & Stern, 1998; Stanley *et al.*, 1989; Sun, 2002). (More discussions of this point will come later.)
- Modularity. Some cognitive faculties are specialized and separate, either because they are functionally encapsulated (i.e. their knowledge and processes do not transfer into other domains) or because they are physically (neurophysiologically) encapsulated. It is relatively easy to justify modularity *teleologically*, which is one of the ways for containing the growth of complexity. Modular structures can be formed evolutionarily so as to simplify learning ontogenetically (or to bypass learning altogether in some cases). Modular structures can be used to guarantee efficiency for important or critical behaviors and routines (whether they are *a priori* or learned). For various notions, as well as justifications, of modularity, see, for example, Fodor (1983), Timberlake and Lucas (1993), Cosmides and Tooby (1994), and Pinker (1994).

It might seem a bad move to start one's theory (i.e. cognitive architectures in this case) with such a broad set of desiderata as those listed above. However, if one is to propose a truly generic cognitive model, it is basically impossible to avoid a broad set of desiderata (as Newell, 1990, discovered with respect to his SOAR model). Furthermore, this set of desiderata is merely the starting point for developing broadly-scoped, comprehensively formulated cognitive architectures. Such a starting point is by all means necessary. In addition to these, we may incorporate even more desiderata down the road, such as externally driven versus internally driven processing and their interactions (e.g. Merleau-Ponty, 1963; Piaget, 1971), issues related to human categorization, and neurophysiological considerations (e.g. Damasio, 1994), which are also theoretically interesting and cognitively fundamental in some ways.

As observed by Vere (1992), because a cognitive architecture aspires to provide an integrative theory of cognition, it is invariably subjected to the "attack of the killer bees"—each subfield or each small domain to which the architecture is applied is

"resolutely defended against intruders with improper pheromones." He proposed that we should "create a sociological environment in which work on integrated cognitive systems can prosper." To do so, "systems entering the cognitive decathlon are judged … based on a cumulative score of their performance in each cognitive 'event'." In this way, contestants do not have to beat all of the narrower systems in their one specialty event, but compete against other well-rounded cognitive systems. This seems to be the appropriate approach towards the development of cognitive architectures.

Below, I will discuss a few of these above *cognitive* desiderata in more detail. In particular, I will address two important points regarding cognitive characteristics: (1) dichotomy of implicit and explicit cognitive processes, and (2) modularity of cognition, along with their respective associated further issues, such as interaction and synergy resulting from the implicit/explicit dichotomy, bottom–up learning, development of modularity, and modularity of memory. Through this set of gradually expanding discussions, I hope to arrive at some more detailed, finer-grained further desiderata for developing cognitive architectures, and some essential research issues in developing future cognitive architectures.

4. An essential dichotomy

The distinction between implicit and explicit processes has been made in many theories of cognition, for example, in Anderson (1983), Keil (1989), Reber (1989), Damasio (1994), and Sun (1994, 2002). It is believed that both types of processes are essential to cognitive agents.

Anderson (1983) proposed the distinction between declarative and procedural knowledge, to account for changes in performance resulting from extensive practice, based on data from a variety of skill learning studies (ranging from arithmetic to geometric theorem proving). For Anderson, the initial stage of skill development is characterized by the acquisition of declarative knowledge (explicit verbal knowledge concerning a task). During this stage, the learner must explicitly attend to this knowledge in order to successfully perform a task. Through practice, a set of implicit procedures develop that allow the task to be performed without using declarative knowledge. When the skill is proceduralized, it can be performed with no access to explicit knowledge and often without concurrent conscious awareness of details involved. Similar distinctions have been made by other researchers based on different data sets.

Several other distinctions made by other researchers capture a very similar difference between different types of processing. For example, Smolensky (1988) proposed a distinction between conceptual (accessible) and subconceptual (inaccessible) processing. According to this framework, explicit knowledge is based on conceptual processing (and thus accessible) and implicit knowledge is based on subconceptual processing (and thus inaccessible). Dreyfus and Dreyfus (1987) proposed the distinction of analytical and intuitive thinking, and believed that the transition from the former to the latter was essential to the development of complex cognitive skills (on the basis of phenomenological analysis of chess playing at

different stages of learning chess). This transition is very similar to the declarative-to-procedural transition as advocated by Anderson (1983, 1993), although the two processes are not identical. Taken together, the distinction between explicit and implicit processes is supported in many ways.

The distinction of implicit and explicit processes has been *empirically* demonstrated in the implicit learning literature. There have been three common tasks used in implicit learning research. The *serial reaction time* tasks (Willingham *et al.*, 1989) probe subjects' ability to learn a repeating sequence. On each trial, one of the four lights on a display screen was illuminated. Subjects were to press the button corresponding to the illuminated light. The lights were illuminated in a repeating 10-trial sequence. It was found that there was a rapid and significant reduction in response time to repeating sequences relative to random sequences. The reduction in response time was attributed to the learning of the sequence. However, subjects might not be able to explicitly report the repeating sequence, and were sometimes even unaware that a repeating sequence was involved.

On the other hand, the *process control tasks* (Berry & Broadbent, 1988) examine subjects' ability to learn a relation between the input and the output variables of a controllable system, through interacting with the system dynamically. Subjects were required to control an output variable by manipulating an input variable. In one instance of the task, subjects were to manage a (simulated) sugar production factory and the goal was to reach and maintain a particular production level, through controlling the size of the workforce. Although they often did not recognize the underlying relations explicitly, subjects reached a certain level of performance in these tasks.

Similarly, in the *artificial grammar learning tasks* (Reber, 1989), subjects were presented strings of letters that were generated in accordance with a finite state grammar. After memorization, subjects showed an ability to distinguish new strings that conformed to the artificial grammar used to generate the initial strings from those that did not. Although subjects might not be explicitly aware of the underlying grammars (barring some fragmentary knowledge), when they were asked to judge the grammaticality ("well-formedness") of novel strings, they performed significantly beyond the chance level.

In all, these tasks share the characteristic of performance being implicit to a significant extent. There are many other tasks that are similar in this regard, such as various concept learning, automatization, and instrumental conditioning tasks (see Sun, 2002, for further details). Together, they clearly demonstrate the distinction between implicit and explicit processes. (It is worth noting that in social psychology, unlike in cognitive science, there have already been a large number of dual-process models dealing specifically with social phenomena. See, for example, Chaiken & Trope, 1999; Smith & DeCoster, 2000; and many others.)

5. Interaction in the dichotomy

Empirical research has shown that human cognition depends on the *interaction* of two types of processes. For example, Mathews *et al.* (1989) suggested that "subjects

draw on two different knowledge sources to guide their behavior in complex cognitive tasks"; "one source is based on their explicit conceptual representations"; "the second, independent source of information is derived from memory-based processing, which automatically abstracts patterns of family resemblance through individual experiences." Likewise, Sun (1994) pointed out that "cognitive processes are carried out in two distinct levels with qualitatively different mechanisms," although "the two sets of knowledge may overlap substantially." Reber (1989) pointed out that nearly all complex cognition in the real world (as opposed to small, controlled laboratory settings) involved a mixture of explicit and implicit processes interacting in some ways, and the relationship between the two might be complex.

Various demonstrations of interaction exist using artificial grammar learning, process control, and other tasks. For instance, Stanley *et al.* (1989) and Berry (1983) found that under some circumstances concurrent verbalization (which generated explicit knowledge) could help to improve subjects' performance in a process control task (i.e. the synergy effect; Sun *et al.*, 2001). Reber and Allen (1978) similarly showed in artificial grammar learning that verbalization (i.e. explicit processes) could help performance. In the same vein, although no verbalization was used, Willingham *et al.* (1989) showed that those subjects who demonstrated more explicit awareness of the regularities in the stimuli (i.e. those who had more explicit knowledge) performed better in a serial reaction time task, which likewise pointed to the helpful effect of explicit knowledge. Ahlum-Heath and DiVesta (1986) also found that verbalization led to better performance in learning Tower of Hanoi [2].

As variously demonstrated by Berry and Broadbent (1988), Stanley *et al.* (1989), and Reber *et al.* (1980), verbal instructions (given prior to learning) can facilitate or hamper task performance too. One type of instruction was to encourage explicit search for regularities that might aid in task performance. Reber *et al.* (1980) found that, depending on the ways stimuli were presented, explicit search might help or hamper performance. Berry and Broadbent (1988) found that explicit search might help or hamper performance depending on the saliency of regularities. Owen and Sweller (1985) found that explicit search hindered learning. Another type of instruction was explicit how-to instruction that told subjects specifically how a task should be performed, including providing information concerning regularities in stimuli. Stanley *et al.* (1989) found that such instructions helped to improve performance significantly. However, Dulaney *et al.* (1984) showed that correct and potentially useful explicit knowledge, when given at an inappropriate time, could hamper learning. All of these findings point to the complex interaction between implicit and explicit processes.

6. The dichotomy and bottom–up learning

Let me address the idea of bottom–up learning, a particular aspect of the interaction between implicit and explicit processes. "Bottom–up learning" concerns how complex reasoning can arise from the simple adaptive behavior, how abstract concepts can be based on simple, concrete, reactive action patterns, how consciousness can emerge from unconsciousness, and so on.

Admittedly, most of the work that makes the distinction between two types of knowledge assumes a *top–down* approach; "proceduralization" leads to skilled performance. In Anderson (1983), proceduralization is accomplished by converting explicit declarative knowledge into implicit production rules, which are subsequently refined through practice. In Anderson (1993), this is accomplished by maintaining explicit memory of instances, which is utilized in performance through analogical processes, and by creating production rules from these instances after repeated use. However, these models were not developed to account for learning in the absence of, or independent from, preexisting explicit domain knowledge.

Several lines of research demonstrate that cognitive agents may learn skills (routines in everyday activities) without first obtaining a large amount of explicit knowledge. In research on implicit learning, Berry and Broadbent (1988), Willingham *et al.* (1989), and Reber (1989) expressly demonstrated a *dissociation* between explicit knowledge and skilled performance, in a variety of tasks including process control tasks (Berry & Broadbent, 1988), artificial grammar learning tasks (Reber, 1989), and serial reaction time tasks (Willingham *et al.*, 1989). Berry and Broadbent (1988) indicated that the human data in process control tasks were not consistent with exclusively top–down learning models, because subjects could learn to perform a task without being provided *a priori* explicit knowledge and without being able to verbalize the rules they used to perform the task. This shows that skills are not necessarily accompanied by explicit knowledge, which would not be the case if top–down learning is the only way to acquire skills. Willingham *et al.* (1989) similarly demonstrated that implicit knowledge was not always preceded by explicit knowledge in human learning, and that implicit and explicit learning were not necessarily correlated. Rabinowitz and Goldberg (1995) showed that there could be parallel learning separately. There have been indications that explicit knowledge may arise from implicit skills in many circumstances. Stanley *et al.* (1989) found that the development of explicit knowledge paralleled but lagged behind the development of implicit knowledge. Reber and Lewis (1977) made a similar observation.

Similar claims concerning the development of implicit knowledge prior to the development of explicit knowledge have also been made in other areas. The implicit memory research (e.g., Schachter, 1987) demonstrates a dissociation between explicit and implicit knowledge/memory, in that an individual's performance can improve by virtue of implicit "retrieval" from memory and the individual can be unaware of the process. This is not amenable to the exclusively top–down approach. Instrumental conditioning also reflects a learning process that is not entirely consistent with the top–down approach, since the process can be non-verbal and non-explicit (without conscious awareness) and lead to forming action sequences without *a priori* explicit knowledge. Such conditioning is applicable to both simple organisms as well as humans (Gluck & Bower, 1988; Thorndike, 1911; Wasserman *et al.*, 1993). In developmental psychology, Karmiloff-Smith (1986) proposed the idea of "representational redescription." During development, low-level implicit representations were transformed into more abstract and explicit representations and thereby made more accessible. This process is not top–down either, but in the exactly opposite direction.

Mandler (1992) proposed a similar process: From perceptual stimuli, relatively abstract "image schemas" were extracted that coded several basic types of movements. Then, on top of such image schemas, concepts were formed utilizing information therein. An infant gradually formed "theories" of how his/her sensorimotor procedures work and thereby gradually made such processes explicit and accessible. Similarly, Keil (1989) viewed conceptual representations as composed of an associative component and a "theory" component, and developmentally, there was a shift from associative to theory-based representations. "Theories" developed from associative information that was already available.

In all, data and theories both indicate that learning may proceed from implicit to explicit knowledge (as well as the reverse). Thus, bottom–up learning can be justified on both empirical and theoretical grounds (Sun, 2002; Sun *et al.*, 2001). This issue is also related to the "symbol grounding" problem: bottom–up learning enables conceptual structures of an agent to be grounded in both the subsymbolic processes of the agent as well as the interactions between the agent and the world (see Sun, 2000, for a detailed discussion of this issue).

7. Modularity

Another important desideratum is modularity. Instead of having one general-purpose machinery (or a small number of them) that is universally applicable, there may be a large number of specialized pieces of machinery each of which deals with a particular aspect or a particular functionality. On this view, the mind is more like a Swiss army knife than a general-purpose blade (Cosmides & Tooby, 1994). This "Swiss army knife" theory of mind has some significant bearing on computational modeling of cognitive agents, as we need to take modular structures into consideration.

Although the notion of modularity has been well known, it is not very clearly understood and delineated. Obviously, delineating modules or establishing a taxonomy of modules is not a simple matter. We need some clarification of the notion of modularity first (and then, of course, detailed model building to be carried out later).

7.1. Notions of modularity

To disentangle the notion, we can distinguish several different senses of the term *module*. First of all, there is the notion of *functional modules*, which are functionally encapsulated, such as early vision, hearing, or visceral processes. Modularity implies that certain processes can only perform a certain range of functions. These processes will only need a limited range of stimuli. The limitedness of stimuli implies *domain specificity*; that is, each module only handles a particular domain (a particular type and range of stimuli). Fodor (1983) was mostly concerned with this type of module. According to Fodor, these modules are cognitively impenetrable (Fodor, 1983), or informationally encapsulated; that is, they are inaccessible to other modules or central systems. For example, according to him, the human language faculty is such

a module. In terms of language, there are indeed some evidence of double dissociation of linguistic processes and other cognitive processes; however, the evidence is far from conclusive.

Second, there is also the notion of *anatomical modules* that are biological and anatomically encapsulated; that is, each of them is located in an isolatable anatomical region of the brain and they work in relative independence of each other. Damasio (1994) dealt with this type of module. This type of module explains the invariability of certain elemental functions.

More complex, higher-level functions can arise from the interaction and combination of different neural circuits (for example, visual processing involves many brain regions and brain circuits). Functions can be accomplished by some (fixed) combinations of neural circuits, i.e. pathways. The fixedness of functions can be due to the fixedness of these pathways (combinations of circuits). A functional module may be spatially distributed, physically shared, or in some other ways not equivalent to an anatomical unit. There have even been discussions of dynamic modules (Tononi & Edelman, 1998).

Lastly, another sense of modules is *domain-specific* modules (Hirschfield & Gelman, 1994; Karmiloff-Smith, 1986). A domain-specific module contains highly specific knowledge and skills, that is, those knowledge and skills that are well developed in a particular domain but do not easily translate into other domains. For example, driving a car is a very specific skill that does not translate into skills for, for example, flying an airplane. This type of domain-specific module may or may not be informationally encapsulated: they may or may not have wide access to information, knowledge, and skills in other modules (thus they are different from, and contain as a subset, functional modules discussed earlier). For example, face recognition may be a domain-specific module, since it is highly developed and specific, but it has to rely on memory retrieval processes and other sources of information.

7.2. Advantages of modularity

Modularity has at least the following advantages, from an agent's stance:

- Computational tractability, in that modularity reduces computational demands (computational complexity of cognition), (1) through the use of separate learning processes in an innately divided architecture, or (2) through automatic decomposition of tasks into various subtasks, or (3) even through innately encoded routines that require (almost) no learning.
- Accuracy and performance in general, because modularity means being free from interference, at least to some extent, from other processes and modules.

These properties offer a definitive evolutionary advantage, and thus it is no surprise that modularity is adopted by cognitive agents through natural selection.

We can also examine the issue from a design stance (i.e. the third person view, for example, used in developing a computational model of a cognitive agent). From such a stance, more advantages can be identified:

- Understandability, in that it is easier to design a modularly structured system, piece by piece, than a whole system at once; thus, modularity can improve the quality of system design.
- Reliability, in that modularity can help to isolate parts of a system to prevent minor problems from spreading and to prevent serious problems from occurring.
- Debuggability, in that it is easier to locate problems when breakdown occurs if a modular architecture is adopted.

However, along with advantages, there may be disadvantages that come with modularity as well (such as the "binding" problem when combining information from multiple visual modules).

Note that the notion of modularity does not necessarily mean that there is no generic mechanism or process that can be applied to many different functions. Domain specificity of modules and generality of mechanisms are not mutually exclusive. For one thing, generic mechanisms can be adopted and then specialized for various specific functions, along a developmental line, into specialized modules (Karmiloff-Smith, 1986). Second, in terms of cognitive modeling, it is more viable, practically speaking, to assume some generic mechanisms and then try to capture various modules by adapting and specializing them and/or by working out some specific combinations. Likewise, generic processes do not necessarily imply centralized representations either, as they can potentially be specialized for representations within individual modules as well as communications among modules.

7.3. Developing modularity

In terms of genesis of modularity, as hypothesized by Wilson (1975),

> When exploratory behavior leads one or a few animals to a breakthrough enhancing survival and reproduction, the capacity for that kind of exploratory behavior and the imitation of the successful act are favored by natural selection … The process can lead to greater stereotyping—'instinct' formation—of the successful new behavior.

It has been proposed (Cosmides & Tooby, 1994) that there are at least the following innate functional modules (or families of instincts), including perceptual/motor-based modules: a spatial perception module, a human face recognition module, a tool-use module, a fear module, an emotion-perception module; and socially-oriented modules: a social-exchange module, a "theory of mind" module, a parenting module, and a mating module; as well as language/communication related modules: a syntax module, a semantics module, a communication module, and so on. Since each of them handles a biologically significant set of stimuli, it is very likely that evolution produces specialized, sharply tuned, and pre-wired modules in a cognitive agent to handle them, to ensure the survival (and the reproduction) of the agent.

Through interacting with the world, individual agents may also develop their

own modules along the way, in order to obtain highly efficient and efficacious skills. This is especially true of "domain-specific modules" mentioned earlier.

8. Modularity of memory

One important kind of modularity is the modularity of memory, that is, the separation of, and the interaction among, different memory modules in a cognitive agent. In the literature, there are all sorts of memory modules, various known as declarative memory, procedural memory, semantic memory, episodic memory, long-term memory, short-term memory, working memory, and so on. Different taxonomies of memory exist. They are concerned with different ways of organization and storage of information, knowledge, concepts, and categories. Let us look into this aspect of cognitive architectures, and develop more fine-grained desiderata for cognitive architectures in relation to memory.

Memory has been an active research area in cognitive science. The kind of research on memory carried out in cognitive science can be traced back to the late 19th century. The usual experimental paradigm consists of the presentation of a list (or several lists) of words or nonsense syllables, and the recall and/or recognition of them later on (immediately afterward or after a substantial delay) by subjects.

8.1. Conflicting taxonomies

Data from experimental work suggested a number of (seemingly) distinct modules in memory. First, based on small, laboratory experiments, the distinction between short-term memory and long-term memory was suggested, in that short-term memory is capacity limited and temporary, while long-term memory is relatively permanent, with unlimited capacity. Short-term memory was viewed, at one point, as a number of slots each of which could hold an item temporarily until displaced by another item. Another characterization (which emerged in the 1970s) was working memory (Baddeley, 1986)—a memory store that consists of three separate short-term memory systems: an articulatory loop, a visual-spatial scratch-pad, and an executive for control, again based on small, laboratory experiments. Cowan (1993) proposed two types of short-term memory: active memory (activated representations in long-term memory) and focus of attention (representations being actively accessed or rehearsed). There has been extensive study of both short-term memory and long-term memory, through (mainly) different designs of laboratory experiments. These studies covers encoding, storage, and retrieval processes (the three major functions of memory), and various issues involved with one or more of these functions: depth of processing, incidental versus intentional learning, forgetting, interference, cue-effectiveness, use of imagery, and so on. Various ideas have been proposed regarding whether there are separate memory systems or whether there is really only one unitary memory system. Experimental results have been ambivalent (Ratcliff & McKoon, 1998).

Another distinction is along the line of semantic memory and episodic memory (Tulving, 1972), suggested by some different experiments. It was suggested that

there were two distinct systems: one for temporal, spatial information of events (the episodic memory) and the other for generic, conceptual information (information of and by concepts).

Yet another distinction is between declarative memory and procedural memory (Anderson, 1983; Tulving, 1983). According to this distinction, a separate memory system is used to store procedural skills, i.e. for the automatic production of skilled performance *without* explicit and controlled access of information, while another memory system is used for storing information of concepts or events, which may allow explicit and controlled access. In general, procedural memory is more rigid and tailored to specific situations, while declarative memory is more general-purpose and flexible; procedural memory can be accessed rapidly, while declarative memory is slower.

There was also the suggestion that memory be divided into explicit memory and implicit memory: For example, procedural memory is implicit while semantic memory and episodic memory are explicit (Tulving, 1983, 1985). Another classification of memory (Roediger, 1990) is a division into declarative (explicit) and procedural (implicit) memory whereby declarative memory is further divided into episodic (working) memory and semantic (reference) memory, and procedural memory into skill, priming, classical conditioning, and other similar memory systems. However, there is no general consensus either regarding whether the distinction of implicit and explicit memory exists or regarding how the memory system should be divided along the explicit/implicit line.

In general, for many issues in memory research, there are contradictory experimental indications; there are thus many mutually conflicting theories. For different experiments or different aspects of memory, even though there may not be direct contradictions in experimental results, there can still be many mutually incompatible, *ad hoc* explanations of them. In this way, explanatory particularism is prevalent [3].

8.2. Conflicting models

In computational modeling, there are equally many mutually conflicting proposals regarding how to capture these memory modules and their distinctions computationally. The simplest representation is the undifferentiated list of features (the feature list), which involves concepts and properties without any structure that connects them. This representation may be adequate if a set of features (i.e. a set of defining features) may be found that are necessary and jointly sufficient for a concept. However, for most concepts in the everyday world, there is no such defining features (Wittgenstein, 1953). To remedy the problem, some more complex (or "advanced") representations were proposed in AI research, which include semantic networks, frames, scripts, schemas, and others (Minsky, 1981; Quillian, 1968). They rely on various structures for (presumably) more precise specifications.

In the area of quantitative psychological models, there have been a variety of them proposed, ranging from instance-based models to prototype-based models, and from spreading activation models to compound-cue models (Ratcliff &

McKoon, 1998). Each of them was quite successful in fitting certain data while failing in fitting some others (Ratcliff & McKoon, 1998).

There are integrated models that combine various memory systems in one framework, addressing the question of the relations between different memory systems. For example, Anderson's cognitive architecture ACT-R (Anderson, 1983; Anderson & Lebiere, 1998) is one of these integrated architectures. ACT-R consists of a production (rule-based) system and a semantic network. The semantic network captures declarative memory, and the production system captures procedural memory. The distinction of short-term and long-term memory is captured through, on the one hand, activation traces as short-term memory and, on the other hand, established nodes and links as long-term memory. The distinction of semantic and episodic memory is not dealt with. The difference in accessibility between implicit and explicit memory is likewise not fully accounted for. There have also been other integrated models, for example, CLARION (Sun, 2002).

In general, in modeling, the difference between implicit and explicit memory has not been adequately addressed (Ratcliff & McKoon, 1998). In particular, mechanistic or process differences between implicit and explicit memory have not been a focus (but see Sun, 2002). The debate between unitary versus multiple memory systems (whereby some are explicit while others implicit) remains ongoing.

8.3. Different methodologies

Although there are many interesting observations, a wealth of data, and even some general principles from memory research, there are problems. First of all, there is a methodological problem: the experiments were conducted mostly in a laboratory setting, with materials as far removed from everyday life as possible (such as nonsense syllables); this methodology may have the advantage of avoiding "contamination" of the experiments from extraneous sources (everyday commonsense knowledge), but it also may lead to neglecting the fact that memory is part of the cognitive processes engaged in, and fine tuned for, everyday activities coping with the world. The artificiality of experimental designs and experimental materials severs the tie between memory and everyday experience of cognitive agents. Thus, experiments may reveal only a partial or even a distorted picture of memory. Even with this problem ignored, the experimental methodology determines that the experiments can only tackle memory by bits and pieces—one disparate issue at a time. Thus we may lose sight of the whole picture. These kinds of experiments are, in some sense, modeled after physics experiments, rather than dealing directly with human existential experience in a holistic and ecologically realistic way [4]. Because of this limiting methodology, every researcher tends to come up with his/her own theory of memory based on his/her own particular perspective and bias. There is a lack of coherence and overall organizing principles.

Second, work on memory tends to view memory as a passive storage device, which simply stores data like a computer memory. However, the human mind is, in fact, far removed from a digital computer (the narrow sense of the word). Dynamics of the human mind is complex and interactive. Memory is not simply *retention*, but

it is also *pretension*, as pointed out by Husserl (1970); that is, it actively participates in intercepting and interpreting the ongoing flow of sensory information and it itself changes and organizes in the process.

Third, current work on memory downplays its active participation in everyday activities and the role of such activities in (and their influence on) memory functions. The everyday activities of a cognitive agent consist of actions and reactions of the agent in the world, for the sake of survival and other needs/goals of the agent, which underlie the memory of the agent. Memory should be viewed as part of the whole of existential experience. This is the path toward a principled, ecological understanding.

In terms of representations used (such as semantic networks), the following shortcomings may be identified from a computational modeling standpoint:

- In semantic networks, conceptual hierarchies require *a priori* determination through hand-coding; in frames, slots need to be determined also through hand-coding. This is not a serious problem if we are only concerned with small toy domains for laboratory experiments. However, in any domain of a realistic size, this poses a serious problem for practical reasons.
- Semantic networks and frames are often filled with *ad hoc* content, that is, hierarchies and structures specifically designed for one particular kind of circumstance or only for the purpose of getting one particular result.
- A related problem is the context-free fixedness of such hierarchies and structures. Usually, conceptual hierarchies are explicitly and manually constructed *a priori*. In human cognition, many concepts (if not all) can be flexible; that is, they can have one or another superordinate concept, depending on (1) the current context (for example, contextual priming; cf. Barsalou, 1983), (2) the current goals, and (3) even personal, idiosyncratic connections. Thus, a more flexible representation is called for [5].

In sum, more research on memory is needed. New research on memory needs to be more ecologically realistic, and pay much more attention to the bigger picture of cognition in the context of everyday activities. On that basis, new cognitive architectures with ecologically realistic memory systems may be devised.

9. Goals and routines

Let us examine some considerations concerning goals as well as (sub)routine structures induced by goals (Anderson, 1993). These considerations are important to the development of cognitive architectures, because sequentiality is an essential behavioral characteristic that needs to be captured in cognitive architectures, and in turn, complex routine/subroutine structures necessary for achieving sequentiality need to be addressed in cognitive architectures as well.

9.1. Importance of routines

Over the years, ethologists, among others, have proposed a number of mechanisms

that provide mechanistic underpinnings for a variety of animal behaviors that are specific to particular stimulus circumstances. Fixed action patterns (FAPs), innate releasing mechanisms (IRMs), and modal action patterns (MAPs) are all instances of such constructs (Savage, 2002). For example, FAPs are self-contained entities in that each has a source of motivational energy that activates a specific sequence of behaviors (a routine or subroutine), under some particular stimulus conditions. According to the models of Lorenz (1950) and Tinbergen (1951), each species was equipped with a sufficient variety of FAPs to ensure an appropriate response in any normal circumstances.

Approximating such motivational constructs (which, notably, were proposed mostly in relation to animals), Anderson (1983, 1993) proposed the use of a goal stack in describing human cognition. A goal stack allows the use of routines (or subroutines) (Sun, 2002; Tyrell, 1993). Once a goal is pushed onto the stack, a routine (or subroutine) for accomplishing the goal is automatically initiated (very much like what was described by Lorenz, 1950), through selecting actions suitable for accomplishing the goal every step of the way. An initiated routine will keep running, until interrupted or terminated. The initiated routine may be terminated by popping the corresponding goal off the stack, when the goal has been accomplished (or when it has been recognized that the goal cannot be accomplished for some reason). During the running of the routine, a higher priority goal may be pushed onto the goal stack when a current state prompts such an action. The current routine can then be suspended, and the routine for the new goal be carried out. At the termination of the new routine, the previous routine may be resumed (or abandoned). A goal may spawn subgoals, by pushing these subgoals onto the stack, which may also cause the suspension of a running routine.

There are alternatives though. In robotics, there have been various proposals concerning "layered architectures" (see, for example, Gat, 1998). These architectures in general share the same basic idea of dividing the action control of a robot into three (or more) components: for instance, (1) the controller, which takes actions reactively in response to environmental input in accordance with some behavioral routines, (2) the sequencer, which selects among different behavioral routines to be carried out by the controller, and (3) the deliberator, which plans out future courses of actions and directs the sequencer to act accordingly. The three layers are quite different in characteristics. The controller is stateless, or has only limited (mostly transient) memory. The sequencer has past state information, on which basis it selects behavioral routines in the controller. The deliberator maintains information about the past and the future, and plans actions accordingly. Partially due to this difference in the amount of information they possess, their speeds vary. The controller is the fastest in making action decisions, while the deliberator is the slowest due to the amount of information it has to deal with. Thus, the division of labor among the three components is useful in maintaining both fast responses (through using the controller) and behavioral flexibility (through using the deliberator). The sequencer may be viewed as the interface between the two, carrying out the plans of the deliberator.

In future cognitive architectures, we need to develop more sophisticated goal

structures and (sub)routine mechanisms. For instance, in a more sophisticated model, goals may emerge from competitions among different needs and desires, goals may change in various ways, including in a stack-like fashion, as well as other possibilities, such as switching to a new goal altogether, and so on. Routines may have both of the following two properties: persistence and interruptibility. The interplay of these two properties needs to be explored and developed in full (Tyrell, 1993).

9.2. Formation of routines

The initiation of routines (e.g. setting goals), the routines themselves, and the termination of routines can all be learned, in addition to being pre-wired using predetermined rules. Routines (and their initiation and termination) may be learned through experience, including autonomous exploration, instructions, imitations, extraction, and other means (see, for example, Sun & Sessions, 2000).

If it is advantageous to invoke a routine (or subroutine) (i.e. to switch to a different action policy for a period of time), then a goal module may learn to set a specific goal, for example, in an attempt to maximize reinforcement, and thereby change the overall state other modules experience. Due to this change, in other modules, a different routine suitable for the current situation may be learned or invoked. Similarly, if it is advantageous to terminate a routine (or subroutine) (i.e. to switch to a different action policy), the goal module may learn to reset the current goal, for example, in an attempt to maximize reinforcement.

10. Phenomenological considerations

Let us turn to some ideas from phenomenological philosophy, which are in fact foundations of what we have been discussing and, as such, serve to justify the foregoing discussions.

10.1. Comportment

One term that Heidegger (1927) chose to describe the basic activities of an agent, the interaction of an agent with its everyday world, is *comportment*. As he put it, "comportments have the structure of directing-one-selves-toward, of being-directed-toward" (Heidegger, 1927). This term denotes the two-way interaction between an agent and its world. We may use this notion as a foundation for understanding the interaction and the mutual dependency between an agent and its world, especially at a subconceptual level (at an implicit cognitive level).

Comportment is direct and unmediated. Thus, it is free from representationalist baggages. Put it another way, comportment does not necessarily involve, or presuppose, explicit representations, and all the problems and issues associated with explicit representations. To the contrary, all representations and relations between mental states and their objects presuppose it as a basis: direct and unmediated

comportment is in fact the condition of possibility of all mental representations. Understanding and modeling comportment is thus the foundation of any ecologically realistic approach toward cognitive science that aims to understand cognitive processes through understanding an agent's interaction with its everyday world (Sun, 2002).

Comportment, according to Heidegger, "makes possible every intentional relation to beings" and "precedes every possible mode of activity in general," prior to explicit beliefs, prior to explicit knowledge, prior to explicit conceptual thinking, and even prior to explicit desire. Comportment is thus primary, in exactly this sense. The traditional mistake of representationalism lies in the fact that they treat explicit knowledge and its correlates as the most basic instead, and thus they turn the priority upside-down; and in so doing, "every act of directing oneself toward something receives [wrongly] the characteristics of knowing" (Heidegger, 1927; see also Bickhard, 1993).

What we need to do to gain a better understanding of comportment beyond mere philosophical speculation is to look into the *development* of comportment (Sun *et al.*, 2001). In particular, we should examine its development in the ontogenesis of an individual agent, which is the most important means by which an agent develops its subconceptual behavioral routines, or comportment, although some of the structures (such as modularity) might be formed evolutionarily, *a priori*, as discussed before.

10.2. Conceptual thinking

However, we also need to go one step further, on the basis of behavioral routines. Simply put, it is not enough to have only (implicit, subconceptual) routines for everyday activities; an agent also needs to develop conceptual thinking to some degree, in order to supplement simple subconceptual reactive responses: for example, to reason before actions, to plan in order to guide reactivity, or to be precise and determinate beside being exploratory.

Conceptual (symbolic) thinking is a derivative way of thinking; symbolic structure is a derivative kind of representation. This point of view has been argued by many philosophers and philosophically minded scientists since Heidegger's time. The reason that conceptual representations and reasoning are derivative is because the opposite side, that is, subconceptual reactive coping in everyday activities, is of utmost importance in an agent's existence in its everyday world. Such coping provides a necessary and minimally sufficient means for an agent to survive in the world.

Let us look into this point. First, subconceptual reactive coping can conceivably provide a minimally sufficient means for survival; just notice the simple organisms that are flourishing on the earth, from bacteria to invertebrates to simple vertebrates, which by no means have any high-level thinking ability beside simple evolved coping mechanisms. Only on the basis of these subconceptual coping activities, in certain species, high-level explicit conceptual thinking arises. Reactive coping thus constitutes the foundation of all other activities. Second, explicit

conceptual thinking may have many advantages, but it is not clear that it alone can sustain an agent in its everyday world. Explicit conceptual representations may be computationally too costly, and it may not be possible to articulate some intricate processes, which is nevertheless necessary for explicit conceptual representations (Bickhard, 1993; Sun, 2000). Therefore, without being able to rely completely on conceptual thinking, reactive coping is a necessary means for agents.

Conceptual thinking is "derived" from low-level mechanisms, because it is secondary in several (different but related) senses: evolutionarily, phylogenetically, ontogenetically (developmentally), and ontologically. It is evolutionarily and phylogenetically secondary, because it was a more recent product of evolution, and has conceivably been evolved from lower-level mechanisms (Wilson, 1975). It is ontogenetically secondary, because in the case of humans, it is usually developed slowly and at a later stage of individual development, after the development of fundamental coping skills (Karmiloff-Smith, 1986, Mandler, 1992). It is ontologically secondary, because it is only a special case of a generic competence for individual actions and individual learning.

10.3. Conceptual thinking versus comportment

Explicit conceptual thinking often leads to a detached and reflective stance (Dreyfus, 1992, p. 45); that is, an agent can step back from the involvement and the reactive engagement in everyday activities and reflect on thoughts abstractly, in a contemplative way. Such a detached and reflective stance is made possible by explicit representations "derived" from ongoing activities, which enable an agent to treat its own thoughts as objects of thinking (i.e. to become detached), instead of being immersed in the coping itself [6].

When an agent is involved completely in everyday routine activities, participation in these activities is "transparent" to the agent, in the sense that there is no explicit conceptual thinking required of the agent. When conceptual representations are developed to some degree, participation in these activities becomes less transparent, since conceptual reasoning starts to intrude and is engaged from time to time.

Let us examine some scenarios in the context of Heidegger's description of equipment and breakdown. According to Heidegger, in everyday routines, things that an agent encounters is considered as "equipment," that is, things that are used for accomplishing something else. Pieces of equipment fit together with each other, into an "equipmental whole." So each piece of equipment functions in a nexus of other pieces of equipment. Together, they constitute the everyday existential world of an agent. Equipment is not primarily understood through conceptually characterizing its shape, function, or other isolated properties. As demonstrated by Wittgenstein's (1953) analysis of common concepts (such as "games"), in general, there is no simply way to characterize a piece of equipment, or any other concepts, in this fashion. As shown by Dreyfus (1991), even a simple piece of equipment such as a chair defies such a characterization. Equipment can be understood, in a primordial sense, from its role in the equipmental whole and its utilization in the everyday

activities of an agent. Equipment is transparent and "ready-to-hand," available for use by an agent in a direct and unmediated way in everyday activities, without the need to involve conscious awareness, focus of attention, or conceptual representations. (Notice the parallel between this view and Gibson's view on direct perception; see Gibson, 1979.)

When normal circumstances are changed and routines interrupted, reactive routines can be disengaged. This is termed breakdown by Heidegger (1927). When breakdown occurs, the ready-to-hand (*Zuhandenheit*) equipment turns into the present-at-hand (*Vorhandenheit*). The agent has to use some other means for dealing with these things, as they are no longer directly available as equipment. Conceptual processes may be brought in. Conceptual reasoning may be used to varying degrees in dealing with breakdown. According to Heidegger, it can go from deliberate activities (as opposed to purely implicit, reactive routines as in the case of routine everyday coping), through full deliberation, to theoretical reflection, and finally, to pure contemplation (Dreyfus, 1991; Heidegger, 1927). For example, deliberate activities involves the use of explicit concepts and references, so that certain previously transparent things become explicit; they are particularly useful when minor disturbances occur in the equipmental whole. On the other hand, full deliberation, involving reflective planning and means–end analysis ("if–then" analysis according to Heidegger), is useful when serious disturbances (for example, the malfunctioning of "equipment") occur whereby new ways of dealing with situations need to be devised (Heidegger, 1927). Theoretical reflection further requires an agent to hold back from the involved practical activities in the world and take a "theoretical" stance, which is a complete changeover from the involved stance in everyday activities. Pure contemplation is the stance that is completely free from any interest or involvement with the world.

Through analyzing equipment and breakdown, Heidegger demonstrated that subconceptual routines ("automated" dealing with ready-to-hand equipment) are more fundamental ontologically than isolated, context-free objects and their properties that are knowable only through (conceptual-level) reflection and contemplation. This analysis reversed the usual philosophical priority placed on explicit conceptual understanding, and placed conceptual processes in their proper places with respect to the processes of cognition. In reality, conceptual processes can only occur on the background of everyday routine activities, in addition to the fact that they are generated from such activities. Because of their background and their origin, conceptual processes can be mixed in with everyday routine activities in various ways, under proper circumstances.

However, on the other hand, conceptual thinking has important roles to play too in cognition. The importance of conceptual thinking, while exulted by representationalists, is often mistakenly ignored by advocates of situated cognition and autonomous agents. In pursuing their causes, many researchers in these areas (situated cognition and autonomous agents) may have overstated their cases in downplaying conceptual thinking. This tendency goes all the way back to Heidegger: the role of conceptual and analytic thinking was downplayed and sometimes ignored by Heidegger in his work.

To see the importance of conceptual thinking, we can again use the dichotomy of embodied skills versus explicit knowledge (as a form of the dichotomy of implicit versus explicit processes). The question is: how can an agent develop a set of skills that are highly specific and highly efficient but, at the same time, can be readily applied to a variety of different situations? This is because agents in the world must deal with novel situations and changing environments. This dilemma is difficult to resolve, although humans seem to possess the ability to achieve an appropriate balance between the two sides. It seems that an agent needs, in addition to highly specific embodied skills, sufficiently general and explicit knowledge that are transferable and explicitly manipulatable (Sun & Peterson, 1998; Sun *et al.*, 2001).

Not only does an agent need generic and explicit conceptual knowledge, as opposed to mere embodied skills, for the sake of generalization, it also needs such knowledge for the sake of conceptual problem solving, creativity, and other non-routine activities that require the uniquely human ability of analytic reasoning and explicit conceptualizing. For example, when planning a trip ahead of time, explicit knowledge of places and routes are needed, since reactive coping is out of question; when scheduling a large project, explicit and generic knowledge is also needed, in order to justify the schedule or optimize the schedule; so on and so forth. Though one may claim that some of these activities may or may not require conceptual thinking, such thinking can certainly be of a great deal of help to a cognitive agent, and may even be indispensable for truly complex situations.

In fact, a balance of the two—specific, subconceptual, embodied skills and generic, explicit, conceptual knowledge—is believed to be essential to cognitive agents in a sufficiently complex world: as mentioned before, on one hand, there are ample psychological data that point to the distinction between the two types and the need for both (for example, Keil, 1989; Reber, 1989; Seger, 1994; Sun *et al.*, 2001). On the other hand, there are philosophical arguments for such a distinction/balance as well (Dewey, 1958; Dreyfus & Dreyfus, 1987; Heidegger, 1927; Sun, 2000), which over the years have become increasingly convincing.

11. An evaluation of the state of the art

Thus far, we explored the desiderata for developing cognitive architectures, which, I believed, needed to be brought to light and explicitly examined, in order to advance the state of the art in cognitive architecture research. I have identified some desiderata for developing a cognitively and ecologically realistic cognitive architecture. Among these, the most important are the dichotomy and the interaction of implicit and explicit processes, modularity, memory systems, and goals/routines. I have gone to great length in elaborating on these points.

Now let us take a quick look at some existing cognitive architectures in light of this discussion. Let us briefly examine a few representative cognitive architectures: ACT-R, SOAR, EPIC, PRODIGY, DEM, COGNET, and CLARION.

First of all, ACT-R (Anderson, 1993; Anderson & Lebiere, 1998) has been examined earlier. On the positive side, the model is arguably the most successful cognitive architecture in existence. It succeeded in capturing a wide variety of

human data in many different task domains, ranging from skill learning to language production. It has been testing a variety of goal mechanisms, including goal stacks. Recently, there have been attempts at adding various perceptual and motor modules to the model. On the other hand, as explained before, it has a certain degree of modularity. It employs the division between procedural and declarative memory. However, in my view, its modularity is not sufficient: It does not have a clear-cut (process-based or representation-based) distinction between implicit and explicit processes, it does not subdivide memory to a sufficient extent, and so on. Furthermore, it does not sufficiently explore the *interaction* between implicit and explicit processes and, relatedly, it does not address *bottom–up* learning, both of which have been shown to be important to cognition (Sun, 2002).

SOAR (see Rosenbloom *et al.*, 1993) is based on the ideas of problem spaces, states, and operators. Prominently in the model, there is a goal stack. When there is an outstanding goal on the goal stack, different productions propose different operators and operator preferences for accomplishing the goal. Learning consists of *chunking*—the creation of a new production that summarizes the process leading up to achieving a subgoal, so as to avoid impasses subsequently. SOAR, like ACT-R, lacks sufficient modularity in its architecture. For instance, in SOAR, there is no sufficiently clear representation-based or process-based difference between implicit and explicit cognition (see Sun, 2002 for arguments). There is no distinction between procedural and declarative memory either. Thus there is no bottom–up learning, or top–down learning. There have been attempts at adding various perceptual and motor modules to the model though.

Like SOAR, PRODIGY (Minton, 1990) involves search through a problem space to achieve goals. The search is based on means–ends analysis: finding an operator that reduces the difference between the current state and the goal. The model encodes control knowledge for the selection of operators and their associated parameters. Learning consists of constructing control rules based on previous problem solving experiences. There is a certain degree of modularity in control knowledge. However, it does not make the implicit/explicit distinction. There is no mechanism for either top–down or bottom–up learning.

Drescher (1991) developed an architecture that attempted to implement the Piagetian constructivist view of development, known as the Dynamic Expectancy Model (DEM). It builds on sensory-motor input/output and creates schemas on that basis. Schemas are formulated as context–action–outcome triples. The learning mechanism is based on statistics collected during interaction with the world. New schemas are created and their contexts identified and tuned through statistical means. The model also builds abstractions out of primitive actions. However, the model does not make the distinction between implicit and explicit knowledge and does not account for the distinction of bottom–up and top–down learning. The model deals only with low-level procedural learning (sensory-motor interaction). As is, it lacks other modules.

EPIC (Meyer & Kieras, 1997) is focused on capturing multi-task performance. It includes a production rule system as the central processor, and a set of detailed perceptual and motor processors. The fundamental assumption of the model is that

most capacity limitations are a result of the limitations of peripheral processors, rather than the central processor. In a way, the model adopts the declarative/procedural distinction. It adopts some other forms of modularity as well, including that of different peripheral processors. However, it does not include the dichotomy of implicit and explicit processes, and does not deal with the interaction between the two types of processes (including bottom–up learning). In fact, there is no learning in the model at all.

COGNET was also developed to handle multiple tasks (Zachary *et al.*, 1996). It consists of a problem context, a set of tasks, an attention manager, and a task execution process. Different tasks are evaluated in the current context and selected for execution. Given its handling of concurrent tasks, the model allows certain modularity. However, to instantiate such modularity, there is much work to be done—details of different modules and their interactions need to be specified before simulation is possible. There is no built-in division between declarative and procedural knowledge or between explicit and implicit processes. There is no mechanism for either top–down or bottom–up learning, or other interactions between implicit and explicit processes.

Now turning to a relatively new cognitive architecture, CLARION has been developed in Sun (1999, 2002) and Sun *et al.* (2001). It employs a variety of modularity, including that between implicit and explicit processes, as well as that between various memory components, and so on. As a result, it directly addresses the interaction between implicit and explicit processes, and in particular bottom–up learning.

CLARION has a dual-representational structure. That is, it consists of two levels: the top level captures *explicit* processes and the bottom level *implicit* processes. CLARION provides a concrete instantiation of the notions of a *fundamental* reactive coping mechanism and a derivative conceptual reasoning capability. Conceptual representations are derived, literally, from reactive routines (the bottom level of CLARION), through bottom–up learning, within the context of ongoing activities in the everyday world. CLARION shows that this is not only possible, but also advantageous. The advantages include: minimizing learning mechanisms, synergy in learning (for example, speeding up learning), synergy in performance (for example, improving performance, improving transfer, and so on), and facilitating multi-agent interactions [7]. CLARION provides a way of studying the *interaction* between two types of knowledge in an integrated but dichotomized architecture. Thus, CLARION points to the way of achieving a proper balance of explicit knowledge and implicit skills (i.e. conceptual and subconceptual processes).

At the bottom level, CLARION captures everyday reactive coping: routines are gradually tuned to deal, in a direct and unmediated way, with everything in the world that the agent encounters, i.e. with all of the "equipment," without necessarily involving conceptual representations and reasoning (Heidegger, 1927). Reactive routines can be effective and efficient in a stable everyday world. On the other hand, as explained earlier, full deliberation, involving reflective planning and means–end analysis, is useful when serious disturbances occur whereby new ways of dealing with situations need to be devised (Heidegger, 1927). In CLARION, this can be

accomplished through extensive use of conceptual representations, at the top level of the model, in which various possibilities can be explicitly reasoned about and explicit sequences of actions and temporal projections onto future states can be established.

Going one step further, theoretical reflection requires that an agent hold back from the involved practical activities in the world and take a "theoretical" stance. In CLARION, it means a complete disengagement from usual sorts of reactive routines and even usual sorts of conceptual processes, as well as from normal everyday activities themselves, and instead being involved in different sorts of processes and representations. This is because it involves looking at things from a different perspective—a theoretical perspective—that aims to investigate things not in terms of their everyday use but in terms of their theoretical interest. Scientific research is, in a way, an example of such activities.

CLARION provides, in computational terms, a natural way for dynamically acquiring conceptual representations for conceptual thinking, without pre-coding (by hand) into the model of a cognitive agent all the requisite knowledge of conceptual thinking. This is accomplished via bottom–up learning as explained earlier. The implementation details can be found in Sun and Peterson (1998) and also in the appendix.

On the negative side, CLARION's mechanisms for goals and routines need to be further developed. As is, they are somewhat rudimentary. In addition, details of memory systems in CLARION need to be further developed too. For one thing, they need to be fleshed out.

Beside these afore-reviewed cognitive architectures, there are of course many more in existence. See, for example, Pew and Mavor (1998) for detailed discussions of some other existing cognitive architectures. Newell (1990) presented a set of criteria of his own, including general issues such as flexibility of behavior, integration of knowledge, dynamic interaction, and adaptation. Anderson and Lebiere (2003) attempted an evaluation of their cognitive architecture ACT-R, along with connectionist models, in accordance with Newell's criteria. Langley and Laird (2002) provide yet another survey of cognitive architectures.

Of course, ultimately, what is important for a cognitive architecture is its ability to account for data in quantitatively precise ways, and to provide interesting interpretations based on quantitative match. Therefore, a good quantitative match is the most important desideratum of all. In this regard both ACT-R and CLARION fare well, better than other competitors (see Anderson, 1993; Anderson & Lebiere, 1998; Sun, 2002; Sun *et al.*, 2001).

12. Concluding remarks

In summary, basic assumptions behind cognitive architectures need to be examined. In this article, I have argued for a set of essential desiderata as the basis for developing future cognitive architectures. The following desiderata have been identified:

- ecological realism

- bio-evolutionary realism
- cognitive realism
- eclecticism of methodologies and techniques

And in terms of cognitive and behavioral characteristics:

- reactivity
- sequentiality
- routineness
- trial-and-error adaptation
- dichotomy of implicit and explicit cognition
- synergistic interaction
- bottom–up learning
- modularity

and so on. Detailed discussions have been carried out concerning some of these points in this article.

A cursory examination of existing cognitive architectures has shown that these desiderata have been, or can be, satisfied to various extents. However, to take into full consideration these desiderata and to develop them to the full extent, much more work is needed. In order to better develop cognitive architectures in the future, we need to address the very issue of a set of basic desiderata for cognitive architectures and base our further efforts on that foundation. Addressing this fundamental issue is a necessary step that will lay the foundation for further progress.

Existing work on cognitive architectures has accomplished a great deal in terms of taking into consideration cognitive and behavioral characteristics of cognitive agents, and in terms of matching human data in a precise and detailed way. In this regard, ACT-R and CLARION, as reviewed earlier, arguably provide two useful examples and, possibly, a starting point for the further development of cognitive architectures.

By no means is this set of desiderata the final word. It is but a starting point in an expectedly long process of discussions and debates. It is hoped that this process can get started quickly and thus further progress can be made quickly in developing better cognitive architectures. This list may need to be expanded, revised, or completely revamped, but we need to start somewhere.

Acknowledgments

This work has been supported in part by Army Research Institute contract DASW01-00-K-0012. Thanks are due to Robert Mathews for discussions, and to Xi Zhang, Yizchak Naveh-Benjamin, and Paul Slusarz for implementing and running simulations. Thanks are also due to the editor, Cees van Leeuwen, and the reviewers, who provided helpful comments on an earlier version.

Notes

[1] For example, in a navigation task, a reactive navigator may be, instead, modeled as a AI planning system in which a complete internal model of the world, along with a set of elaborate planning rules, becomes part of the architecture.

[2] However, as Reber (1989) pointed out, verbalization and the resulting explicit knowledge might also hamper (implicit) learning, under some circumstances, especially when too much verbalization induced an overly explicit learning mode in subjects performing a task that was not suitable for learning in an explicit way, for example, when learning a rather complex artificial grammar. Similarly, in a minefield navigation task, Sun *et al.* (2001) reported that too much verbalization induced overly explicit learning that was detrimental to performance.

[3] For example, Pashler (1998) pointed out that "the broad three-part distinction [sensory store, short-term memory, long-term memory] has received an extraordinary amount of criticism over the years. The original, relatively primitive versions of the model require important modification, principally, abandoning the idea of serial information flow and the suggestion that short-term memory is unitary and exclusively verbal. Furthermore, the model must be read with the understanding that the proposed memory structures are not claimed to have, nor are they likely to have, the exclusive function of memory storage."

[4] Recall the origin of the experimental methodology in the late 19th century, when physics was the model of clarity that all disciplines tried to imitate.

[5] It may be argued that this problem can be remedied by introducing some new components in a model that can modify the connections and slots on the fly in accordance with various dynamic factors. Although this is theoretically possible, semantic networks are notoriously complex and difficult to modify, even statically, let alone dynamically.

[6] I should note the derivative character of this detached stance, in the very same sense as its source, conceptual thinking, is derivative. Heidegger vehemently opposed the use of this detached stance in philosophical thinking, especially as used in Husserlian phenomenology. "The achieving of phenomenological access to the beings which we encounter, consists rather in thrusting aside our interpretative tendencies, which keep thrusting themselves upon us and running along with us, and which conceal not only the phenomenon of such 'concern', but even more those things themselves as encountered of their own accord in our concern with them" (Heidegger, 1927). Dreyfus put it this way in his interpretation of Heidegger: "The bare objects of pure disinterested perception are not basic things we can subsequently use, but the debris of our everyday practical world left over when inhibiting action" (Dreyfus, 1992). He pointed out that neither everyday activities nor detached thinking can be conceived as "a relation between a self-sufficient mind and an independent world," to forgo the inextricable involvement of one with another.

[7] All of these points have been studied in experimental work on CLARION. See Sun and Peterson (1998) and Sun *et al.* (2001).

References

AGRE, P. & CHAPMAN, D. (1990). What are plans for? In P. MAES (Ed.) *Designing autonomous agents*. New York: Elsevier.

AHLUM-HEATH, M. & DIVESTA, F. (1986). The effect of conscious controlled verbalization of a cognitive strategy on transfer in problem solving. *Memory and Cognition*, *14*, 281–285.

ANDERSON, J. & LEBIERE, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

ANDERSON, J. & LEBIERE, C. (2003). The Newell test for a theory of mind. *Brain and Behavioral Sciences*, *26*(5), 587–640.

ANDERSON, J.R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

ANDERSON, J.R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

BADDELEY, A. (1986). *Working memory*. New York: Oxford University Press.

BARSALOU, L. (1983). *Ad hoc* categories. *Memory and Cognition*, *11*, 211–227.

BERRY, D. (1983). Metacognitive experience and transfer of logical reasoning. *Quarterly Journal of Experimental Psychology*, *35A*, 39–49.

BERRY, D. & BROADBENT, D. (1988). Interactive tasks and the implicit–explicit distinction. *British Journal of Psychology*, *79*, 251–272.

BICKHARD, M. (1993). Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence*, *5*, 285–333.

BOWERS, K., REGEHR, G., BALTHAZARD, C. & PARKER, (1990). Intuition in the context of discovery. *Cognitive Psychology*, *22*, 72–110.

BREIMAN, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

CHAIKEN, S. & TROPE, Y. (Eds) (1999). *Dual process theories in social psychology*. New York: Guilford Press.

CLARK, A. (1997). *Being there*. Cambridge, MA: MIT Press.

COSMIDES, L. & TOOBY, J. (1994). Beyond intuition and instinct blindness: toward an evolutionarily rigorous cognitive science. *Cognition*, *50*, 41–77.

COWAN, N. (1993). Activation, attention, and short-term memory. *Memory and Cognition*, *21*, 162–167.

DAMASIO, A. (1994). *Descartes' error: emotion, reason and the human brain*. New York: Grosset/Putnam.

DEWEY, J. (1958). *Experience and nature*. New York: Dover.

DRESCHER, G. (1991). *Made-up minds*. Cambridge, MA: MIT Press.

DREYFUS, H. (1992). *Being-in-the-world*. Cambridge, MA: MIT Press.

DREYFUS, H. & DREYFUS, S. (1987). *Mind over machine: the power of human intuition*. New York: The Free Press.

DULANEY, D., CARLSON, R. & DEWEY, G. (1984). A case of syntactic learning and judgment: How conscious and how abstract. *Journal of Experimental Psychology: General*, *113*, 541–555.

FODOR, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

GAT, E. (1998). On three-layered architecture. In D. KORTENKAMP, R. BONASSO & R. MURPHY (Eds) *Artificial intelligence and mobile robots*. Menlo Park, CA: AAAI Press.

GIBSON, J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

GLUCK, M. & BOWER, G. (1988). From conditioning to category learning. *Journal of Experimental Psychology: General*, *117*, 227–247.

HEIDEGGER, M. (1927). *Being and time* [English translation New York: Harper and Row, 1962].

HIRSCHFIELD, L. & GELMAN, S. (Eds) (1994). *Mapping the mind: domain specificity in cognition and culture*. Cambridge: Cambridge University Press.

HUSSERL, E. (1970). *Logical investigation*. London: Routledge and Kegan Paul.

HUTCHINS, E. (1995). How a cockpit remembers its speeds. *Cognitive Science*, *19*, 265–288.

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

KEIL, F. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

LANGLEY, P. & LAIRD, J. (2002). Cognitive architectures: research issues and challenges. Manuscript.

LORENZ, K. (1950). The comparative method in studying innate behavior patterns. *Symposium of the Society of Experimental Biology*, *4*, 221–268.

MEYER, D. & KIERAS, D. (1997). A computational theory of executive cognitive processes and human multiple-task performance: part 1, basic mechanisms. *Psychological Review*, *104*, 3–65.

MANDLER, J. (1992). How to build a baby. *Psychological Review*, *99*, 587–604.

MATHEWS, R., BUSS, R., STANLEY, W., BLANCHARD-FIELDS, F., CHO, J. & DRUHAN, B. (1989). Role of implicit and explicit processes in learning from examples: a synergistic effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*, 1083–1100.

MERLEAU-PONTY, M. (1963). *The structure of behavior*. Boston: Beacon Press.

MINSKY, M. (1981). A framework for representing knowledge. In J. HAUGELAND (Ed.) *Mind design* (pp. 95–128). Cambridge, MA: MIT Press.

MINTON, S. (1990). Quantitative results concerning the utility of explanation-based learning. *Artificial Intelligence*, *42*, 363–391.

NEWELL, A. (1980). Physical symbol systems. *Cognitive Science*, *4*, 135–183.

NEWELL, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

OWEN, E. & SWELLER, J. (1985). What do students learn while solving mathematics problems? *Journal of Experimental Psychology*, *77*, 272–284.

PASHLER, H. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.

PEW, R.W. & MAVOR, A.S. (Eds) (1998). *Modeling human and organizational behavior: application to military simulations*. Washington, DC: National Academy Press.

PIAGET, J. (1971). *Biology and knowledge*. Edinburgh: Edinburgh University Press.

PINKER, S. (1994). *The language instinct*. New York: Morrow.

QUILLIAN, M.R. (1968). Semantic memory. In M. MINSKY (Ed.) *Semantic information processing*. Cambridge, MA: MIT Press.

RABINOWITZ, M. & GOLDBERG, N. (1995). Evaluating the structure-process hypothesis. In F. WEINERT & W. SCHNEIDER (Eds) *Memory performance and competencies*. Hilsdale, NJ: Erlbaum.

RATCLIFF, R. & MCKOON, G. (1998). Memory models. In E. TULVING (Ed.) *Oxford handbook of memory* (pp. 571–581). New York: Oxford University Press.

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

REBER, A. & ALLEN, R. (1978). Analogy and abstraction strategies in synthetic grammar learning: a functionalist interpretation. *Cognition*, *6*, 189–221.

REBER, A. & LEWIS, S. (1977). Implicit learning: an analysis of the form and structure of a body of tacit knowledge. *Cognition*, *5*, 333–361.

REBER, A., KASSIN, S., LEWIS, S. & CANTOR, G. (1980). On the relationship between implicit and explicit modes in the learning of a complex rule structure. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 492–502.

RESCORLA, R. & WAGNER, A. (1972). A theory of Pavlovian conditioning. In A. BLACK & W. PROKASY (Eds) *Classical conditioning II: current research and theory*, 64–99. New York: Appleton–Century–Crofts.

ROEDIGER, H. (1990). Implicit memory: retention without remembering. *American Psychologist*, *45*, 1043–1056.

ROSENBLOOM, P., LAIRD, J. & NEWELL, A. (1993). *The SOAR papers: research on integrated intelligence*. Cambridge, MA: MIT Press.

SAVAGE, T. (2003). The grounding of motivational constructs in artificial animals: indices of motivational behavior. *Cognitive Systems Research*, *4*, 23–55.

SCHACTER, D. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *13*, 501–518.

SEGER, C. (1994). Implicit learning. *Psychological Bulletin*, *115*, 163–196.

SHANKS, D. (1993). Human instrumental learning: a critical review of data and theory. *British Journal of Psychology*, *84*, 319–354.

SIEGLER, R. & STERN, E. (1998). Conscious and unconscious strategy discovery: a microgenetic analysis. *Journal of Experimental Psychology: General*, *127*, 377–397.

SMITH, E. & DECOSTER, J. (2000). Dual process models in social and cognitive psychology: conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*, 108–131.

SMOLENSKY, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*, 1–74.

STANLEY, W., MATHEWS, R., BUSS, R. & KOTLER-COPE, S. (1989). Insight without awareness: on the interaction of verbalization, instruction and practice in a simulated process control task. *Quarterly Journal of Experimental Psychology*, *41A*, 553–577.

SUN, R. (1994). *Integrating rules and connectionism for robust commonsense reasoning*. New York: Wiley.

SUN, R. (1999). Accounting for the computational basis of consciousness: a connectionist approach. *Consciousness and Cognition*, *8*, 529–565.

SUN, R. (2000). Symbol grounding: a new look at an old issue. *Philosophical Psychology*, *13*, 403–418.

SUN, R. (2002). *Duality of the mind*. Mahwah, NJ: Erlbaum.

SUN, R. & PETERSON, T. (1998). Autonomous learning of sequential tasks: experiments and analyses. *IEEE Transactions on Neural Networks*, *9*, 1217–1234.

SUN, R. & SESSIONS, C. (2000). Self-segmentation of sequences: automatic formation of hierarchies of sequential behaviors. *IEEE Transactions on Systems, Man, and Cybernetics: Part B, Cybernetics*, *30*, 403–418.

SUN, R., MERRILL, E. & PETERSON, T. (2001). From implicit skills to explicit knowledge: a bottom–up model of skill learning. *Cognitive Science 25*, 203–244.

SUTTON, R. & BARTO, A. (1981). Towards a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, *88*, 135–170.

THORNDIKE, E. (1911). *Animal intelligence*. Darien, CT: Hafner.

TIMBERLAKE, W. & LUCAS, G. (1989). Behavior systems and learning: from misbehavior to general principles. In S.B. KLEIN & R.R. MOWRER (Eds) *Contemporary learning theories: instrumental conditioning theory and the impact of biological constraints on learning* (pp. 237–275). Hillsdale, NJ: Erlbaum.

TINBERGEN, N. (1951). *The study of instinct*. London: Oxford University Press.

TONONI, G. & EDELMAN, G. (1998). Consciousness and complexity. *Science*, *282*, 1846–1851.

TULVING, E. (1972). Episodic and semantic memory. In E. TULVING & W. DONALDSON (Eds) *Organization of memory* (pp. 381–403). New York: Academic Press.

TULVING, E. (1983). *Elements of episodic memory*. Oxford: Clarendon Press.

TULVING, E. (1985). How many memory systems are there? *American Psychologist*, *40*, 385–398.

TYRELL, T. (1993). Computational mechanisms for action selection. PhD thesis, Oxford University.

VERE, S.A. (1992). A cognitive process shell. *Behavioral and Brain Sciences*, *15*, 460–461.

WASSERMAN, E., ELEK, S., CHARTLOSH, D. & BAKER, A. (1993). Rating causal relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 174–188.

WILLINGHAM, D., NISSEN, M. & BULLEMER, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1047–1060.

WILSON, E. (1975). *Sociobiology*. Cambridge, MA: Harvard University Press.

WITTGENSTEIN, L. (1953). *Logical investigation*. New York: Macmillan.

ZACHARY, W., LE MENTEC, J. & RYDER, J. (1996). Interface agents in complex systems. In C. NITUEN & E. PARK (Eds) *Human interaction with complex systems: conceptual principles and design practice*. Needham, MA: Kluwer.

Appendix

Below are some relevant details concerning CLARION from Sun *et al.* (2001) and Sun (2002). Overall,

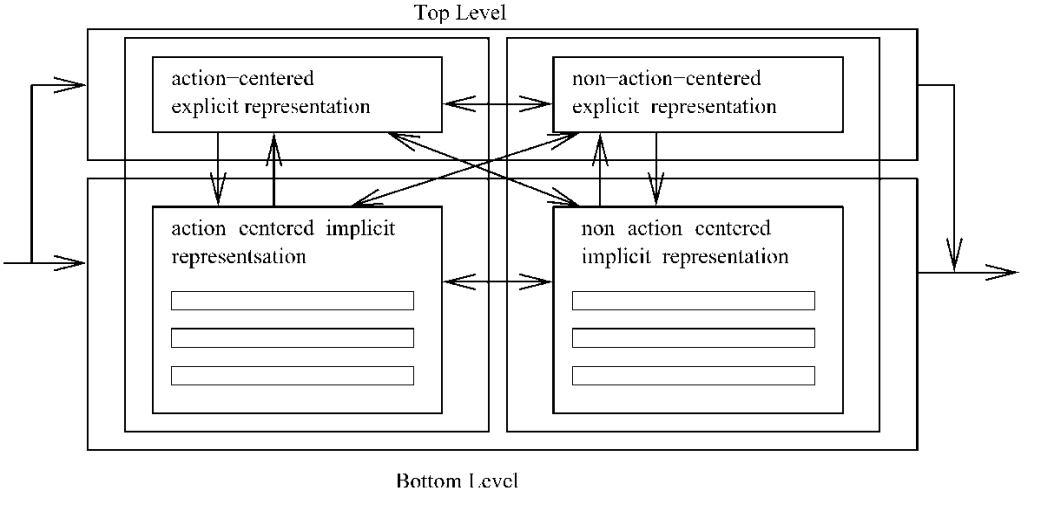

FIG. 1. The CLARION architecture.

CLARION has a dual-representational structure. It consists of two levels: the top level captures explicit processes and the bottom level implicit processes (see Figure 1). In so doing, as mentioned before, CLARION provides a concrete instantiation of the notions of a *fundamental* reactive coping mechanism and a *derivative* conceptual reasoning capability.

The *Rule-Extraction-Refinement* algorithm (RER) learns explicit knowledge at the top level (in the form of rules) using information from the bottom level, to capture the bottom–up learning process (Karmiloff-Smith, 1996; Stanley *et al.*, 1989). The basic idea of this algorithm is as follows: if an action decided by the bottom level is successful (i.e. if it satisfies a certain criterion), then the agent extracts an explicit rule (with its action corresponding to that selected by the bottom level and with its condition specifying the current input state), and adds the rule to the top level. Then, in subsequent interactions with the world, the agent refines the constructed rule at the top level by considering the outcome of applying the rule: if the outcome is successful, the agent may try to generalize the condition of the rule to make it more universal; if the outcome is not successful, then the condition of the rule may be made more specific and exclusive of the current state.

The details of the operations used in the above algorithm (including rule extraction, generalization, and specialization) and the criteria measuring whether a result is successful or not (used in deciding whether or not to apply some of these operators) are described in Sun *et al.* (2001) and Sun and Peterson (1998). Essentially, successfulness is measured by an information gain measure, which indicates whether a rule provides useful information or not. The information gain measure is computed from data generated by the bottom level. The current states, and therefore the rule conditions, are described based on dimension-value representation. Generalization amounts to adding an additional value to one input dimension in the condition of a rule, so that the rule will have more opportunities of matching inputs. Specialization amounts to removing one value from one input dimension in the condition of a rule, so that it will have less opportunities of matching inputs. Iterative processes of rule generalization and specialization, under the guidance of the information gain measure (and thus the bottom level), lead to useful explicit rules at the top level for a particular task. Conditions of these learned rules constitute concepts in the conceptual representation of an agent (at the top level), which are geared toward specific prior experience (the experienced tasks). It is clear that this whole process of bottom–up learning is under the guidance of bottom-level reactive routines, which are trained by reinforcement learning algorithms (Sutton & Barton, 1981).

In the bottom level, *Q-learning* is a reinforcement learning algorithm. In the algorithm, each Q value estimates the maximum total reinforcement that can be received from the current state and the currently chosen action on. A Q value is an evaluation of the "quality" of an action in a given state. Thus, actions are selected based on Q values. Specifically, at each step, given the current state, we compute the Q values of all the possible actions. We then use the Q values to decide on an action to be performed (e.g. by choosing the action with the highest Q value). Q values are gradually tuned, online, through successive updating during interaction with the world (i.e. through the Q-learning algorithm), to enable reactive sequential behavior to emerge in the bottom level (Sun & Peterson, 1998).