Chapter 1

Questions and Answers

Ron Sun

In this chapter, I address a number of commonly raised questions about Clarion, as well as some other important questions that should, by all means. be addressed also.

Given the scopes of some of these questions, this chapter may simultaneously serve as a comparison chapter, a literature review chapter, as well as a discussion chapter.

1.1 The Architecture

Why is there the "top level" in Clarion?

Cognitively speaking, we need to capture the explicit knowledge that humans do exhibit (e.g., what they express when they verbalize). The existence of such knowledge is well established, as the distinction between implicit and explicit knowledge has been amply demonstrated (see, e.g., Reber 1989, Stadler and Frensch 1998, Sun et al 2005). Therefore it needs to be captured in some form in computational models.

Computationally speaking, the involvement of the top level (in addition to the bottom level) may lead to "synergy" (Sun 2002): better performance under various circumstances due to the interaction of the two levels.

Why is there the "bottom level" in Clarion?

The evidence for the existence of implicit knowledge through implicit learning, distinct from explicit knowledge that may be easily expressed verbally, is mounting. Although the issue is not uncontroversial, there are yet sufficient reasons to believe in the existence and the significance of implicit knowledge in various kinds of cognitive processes (as variously argued by Reber 1989, Seger 1994, Cleeremans et al 1998, Stadler and Frensch 1998, Sun 2002, Evans 2005). Early on, I have argued that implicit knowledge is best captured by neural networks with distributed representation (see also Sun 2002, Cleeremans 1997). Hence there is the bottom level in Clarion. In addition, the bottom level is also important in capturing bottom-up learning.

Why are there the two "levels" in Clarion after all?

First of all, there are plenty of philosophical arguments. Below are some evidence and arguments for the duality, or dichotomy, of the mind, which naturally leads to the *dual-representation hypothesis* (Sun 1992, 1994, 1995) and consequently the two-level models (Sun 1994, 1995, 1997, 2002).

First let us look into some of the early ideas concerning dichotomies or dualities of the mind that dated back before the inception of cognitive science. For instance, Heidegger's distinction — the preontological vs. the ontological — is a highly abstract version of such a dichotomy. As a first approximation, his view is that, since the essential way of being is existence in the world, an agent always embodies an understanding of its being through such existence. This embodied understanding is implicit and consists of skills, reactions, and know-hows, without an explicit "ontology", and is thus *preontological.* On that basis, an agent may also achieve an explicit understanding, an *ontological* understanding, especially through making explicit the implicitly held understanding; or in other words, the agent can turn preontological understanding into ontological understanding (Heidegger 1927, Dreyfus 1991). This dichotomy and progression from the concrete to the abstract are the basis of our model (to be explained later).

It is also worthwhile to mention William James's distinction of "empirical thinking" and "true reasoning". According to James, on the one hand, empirical thinking is associative, made up of sequences of "images" that are suggested by one another. It is "reproductive", because it is always replicating in some way past experience, instead of producing new or stand-alone ideas. Empirical thinking relies on overall comparisons and similarity among various concrete situations, and therefore may lose sight of some critical information. On the other hand, "true reasoning" can be arrived at by abstracting particular attributes (i.e., those attributes that are essential and critical) out of a situation. In so doing, we assume a particular way of conceiving things — we see things as a particular aspect of them. It is "productive", because it is capable of producing novel ideas through abstraction. An important function that "true reasoning" serves is to break up the direct link between thought and action, and to provide means for articulately and theoretically reasoning about consequences of an action without actually performing it.

Dreyfus and Dreyfus (1987) proposed the distinction of analytical and intuitive thinking, refining and revising Heidegger's distinction in a contemporary context. They claim that analytical thinking corresponds to what traditional (symbolic) AI models are aimed to capture: deliberate, sequential processing that follows rules and performs symbolic manipulation. According to them, when an agent first learns about a domain (for example, chess), an agent learns explicit rules and follows them one-by-one. After gaining some experience with the domain, one will start to develop certain overall understanding of a situation as a whole, without deliberate rule-following and analytical thinking. That is, one starts to use intuitive thinking, which has the characteristics of being situationally sensitive, "holographic", and non-rule-like. Contrasting the two types of thinking, there is clearly a dichotomy, although the focus here is the reverse progression from the abstract to the concrete.

There are a few arguments from within cognitive science, and from connectionists in particular, that are related to this dichotomy. Smolensky (1988) proposed the distinction of conceptual and subconceptual processing. Conceptual processing involves knowledge that possesses the following characteristics: (1) public access, (2) reliability, and (3) formality. In other words, they are what traditional symbolic AI tries to capture (as similarly identified by Dreyfus and Dreyfus 1987). On the other hand, there are other kinds of capacities, such as skill, intuition, and individual knowledge that are not expressible in linguistic forms and do not conform to the three criteria prescribed above. It has been futile so far to try to model such capacities with conceptual processing based models (traditional AI symbolic processing models). Some of the capacities should be viewed as an entirely different level in cognition and modeled as such, that is, at the subconceptual level. The subconceptual level may be better dealt with by the connectionist subsymbolic models, because the connectionist approach seems to be able to overcome some problems symbolic AI models encountered in modeling subconceptual processing.

We may also examine various arguments based on interpreting psychological data and/or based or computational modeling

For many decades up until very recently, in experimental studies of the human mind, various operationalized notions related to consciousness were proposed: concerning the explicit vs. the implicit, the controlled vs. the automatic, the intentional vs. the incidental, and so on. For instance, in cognitive psychology, there is the well established distinction of implicit memory vs. explicit memory (Schacter 1990, Roedeger 1990). Implicit memory refers to unconscious retrieval of memories, without explicit awareness. Based on the dissociation of explicit and implicit memory tests, it was suggested that implicit memory and explicit memory involved different memory systems (for example, the episodic memory system vs. the semantic memory system, or the declarative memory vs. the procedural memory; Bower 1996, Squire et al 1993, Schacter 1990). Related, but not identical to this, there is also the distinction of implicit learning and explicit learning. In the research on implicit learning, Berry and Broadbent (1988), Willingham et al (1989), and Reber (1989) expressly demonstrated a dissociation between explicit knowledge and performance in a variety of tasks. The notion of automaticity (Shiffrin and Schneider 1977, Logan 1988) is related to that of implicit learning. Automatic processing is assumed to be effortless and resource-wise (almost) unbounded, while its opposite, controlled processing, requires the use of limited cognitive resources (Navon and Gopher 1979). There is also the distinction of declarative and procedural knowledge from cognitive psychology. In Anderson (1983), based on psychological data on high-level cognitive skill learning (such as arithmetic and theorem proving), it was suggested that there were these two types of knowledge. As described by Anderson (1983) and many others, while the former type of knowledge is generic and easily accessible, the latter type is embodied and specific. In this theory, there is a clear dichotomy between these two types of knowledge. In all, the evidence for these dichotomies lies in experimental data that elucidate various dissociations and differences in performance under different conditions.

Also worth mentioning is the notions of two types of representations, used in the connectionist literature. Basically, in *localist* (or symbolic) representation, each representational unit (node) represents a distinct entity to be represented. There is a one-to-one correspondence between units of representation and entities to be represented. In *distributed* representation, each entity is represented by a pattern of activation among a pool of representational units (nodes). Although there is a one-to-one correspondence between an entity to be represented and a pattern of activation, there is no one-to-one correspondence between entities to be represented and representational units (nodes). This distinction is reminiscent of various other distinctions related to the conscious and the unconscious, and actually bears close relationships (Sun 1994, 1995, 1997) to these other distinctions, as extensively discussed in Sun (2002).

There seems to be a consensus with regard to the *qualitative* difference between different types of cognition (although there is no consensus regarding the actual details of the dichotomies). Moreover, many of the aforementioned authors believed in incorporating both sides of the dichotomies in cognitive models, because each side serves a unique cognitive function and is thus indispensable for a complete cognitive model. On this basis, it is natural to hypothesize the existence of two separate components, whereby each component is responsible for one side of a dichotomy, for example, as have been proposed in Anderson (1993), Hunt and Lansman (1986), Logan (1988), Reber (1989), Schacter (1990), and Sun (1994, 1995). The two components were variously described as production systems vs. semantic networks, as algorithmic processes vs. instances retrieval, or as localist representation vs. distributed representations.

Figure 1.1 summarizes the main characteristics associated with the two levels in relation to empirical data and to computational modeling.

Dimensions	bottom	top
Cognitive phenomena	implicit learning	explicit learning
	implicit memory	explicit memory
	automatic processing	controlled processing
	intuition	explicit reasoning
Source of knowledge	trial-and-error	external sources
	assimilation of explicit knowledge	extraction from implicit knowledge
Representation	distributed (micro)features	localist conceptual units
Operation	similarity-based	explicit symbol manip.
Characteristics	more context sensitive, fuzzy	more crisp, precise
	less selective	more selective
	more complex	simpler

Figure 1.1. Comparisons of the two levels of the Clarion architecture

There is also neurobiological evidence. For example, Posner et al (1997) reviewed evidence from brain imaging research that indicated the possibility of different brain regions being associated with implicit and explicit learning and memory. See also Aizenstein (2000) and Poldrack et al (2001). Keele et al (2003) reviewed evidence from brain imaging studies that further delineated the nature of two separate learning processes. Mishkin et al (1984) studied the contrast between (explicit) one-shot learning and (implicit) repeated habit formation training through using brain lesioned primates, which showed physiological separation of these two types of processes. The studies involving brain damaged amnesic patients (such as by Nissen and Bullemer 1987) indicated also that such patients were able to learn as well as normals in task settings where implicit learning was dominant, but not in settings where explicit learning was required, which also lent support for the physiological separation of the two types of processes. However, the results of Boyd and Weistein (2001) and Muente et al (2001) seemed different.

See also LeDoux (1996).

See also Milner and Goodale (1995).

Clarion provides unified re-interpretations of these (partial) notions and comes up with a unified computational model encompassing many of these ideas, in a two-level framework.

Solely from the cognitive modeling and simulation perspective, why are there two levels?

Simply put, the presence of the two levels in Clarion provides a unified and succinct account of a variety of data and phenomena, ranging from serial reaction time tasks and artificial grammar learning tasks to Tower of Hanoi. For example, various synergy effects have been accounted for by Clarion, which include, for example, improved performance through explicit search or verbalization (Sun et al 2001, Sun et al 2005).

Furthermore, the difference between the two levels accounts for a variety of psychological constructs concerning psychological differences as exhibited in the existing data, which has been mentioned in the answer to the previous question, and extensively discussed in chapter 5 of Sun (2002).

What are the fundamental differences between the two levels—the implicit and the explicit?

There are several different kinds of differences between the two levels in the architecture.

- Phenomenological differences: the distinction between the conscious and the unconscious in a phenomenological sense.
- Psychological differences: the distinction revealed by experiments in psychology (e.g., implicit versus explicit learning, implicit versus explicit memory, unconscious versus conscious perception, and so on).
- Implementational differences: for example, the representational difference (symbolic versus distributed representation) in the two levels of Clarion.

In Clarion, the implementational differences lead to accounting for the phenomenological and the psychological differences. So in this sense, the implementational differences are fundamental to the architecture.

What are the fundamental differences between action-centered and non-action-centered knowledge?

Most importantly, these are two different kinds of knowledge for two different kinds of purposes: action-centered knowledge is (mostly) for directly controlling actions, while non-action-centered knowledge is used for reasoning etc.

Consequently, action-centered knowledge is (mostly) used in one direction —from condition to action. Non-action-centered knowledge tend to be more flexible in terms of its use and manipulation.

Do you realize that the idea that both implicit and explicit processes contribute to task performance is not a new idea?

There indeed have been a few (rather lonely) voices arguing that point (a very important point, in my view) early on. Mishins et al (1984), Reber (1989), Mathews et al (1989), and so on have been cited regarding their contribution to the idea that both implicit and explicit processes contribute to skill learning and skilled performance.

However, the *novelty* of CLARION in this regard is also evident. The main novel point in CLARION is the focus on the interaction of implicit and explicit processes, in the context of research on implicit learning and other related topics, (1) in terms of adjustable amounts of contributions from these two types during skilled performance, and hence synergy effects (depending on contextual factors), and (2) more importantly, in terms of their interaction during skill learning (that is, bottom-up and top-down learning and so on). Another novel point concerns computational modeling of learning: (3) while some other models involving some implicit/explicit interaction did not have a detailed and implemented process of learning, CLARION has. And it has been demonstrated computationally that it works.

This work provides some evidence that the interaction between these two parallel types of processes is important, and a framework that addresses this issue can account for a wide variety of data in a unified way, both quantitatively and qualitatively. The interaction of implicit/explicit processes has been used to account for: the verbalization effect, the explicit search effect, the explicit instruction effect, the salience difference effect, the dual tasks effect, and so on.

The cognitive architecture succeeds in interpreting many findings in skill learning that have not yet been adequately captured in computational modeling (such as bottom-up learning and synergy effects) and points to a way of incorporating such findings in a unified model, and thus it has significant theoretical implications.

What is the advantage of bottom-up learning?

The main computational advantage is that it enables learning in complex domains where there is no or very little a priori domain-specific knowledge. This is because implicit learning may be capable of dealing with more complex situations (see the answers to the relevant question), and the bottom-up process makes learning explicit knowledge easier by learning implicit knowledge first (through reducing search by utilizing implicit statistical information to guide the search for explicit knowledge; see the answers to the related questions).

Why can an agent not learn explicit knowledge directly? Why is bottom-up learning emphasized?

Of course an agent can learn explicit knowledge directly. The reason why bottom-up learning has been emphasized in this work is because it has not been emphasized enough in the past in the literature.

Having said that, I should add that there are certain *computational* advantages that come with bottom-up learning, as opposed to directly learning explicit knowledge. For one thing, employing this two-stage approach may be a more efficient way of learning explicit knowledge, because, guided by implicit knowledge, the search space for explicit rules is narrowed down and an on-line, incremental search for explicit rules is then performed. This fact may explain why evolution has chosen this approach (as I believe it).

In addition, there have been some human data reported in the literature that indicate that it is likely that humans do engage in bottom-up learning (e.g., Stanley et al 1989, Karmiloff-Smith 1986, Sun et al 2001, Sun 2002). So, bottom-up learning may have the advantage of being cognitively realistic.

Are symbolic representations in CLARION arbitrary and/or arbitrarily created?

No. First of all, whatever symbolic representations acquired through bottom-up learning are closely tied to, and grounded in, subsymbolic representations and processes (as well as in the ongoing interactions between the agent and the world; Sun 2000, 2002). Second, such symbolic representations (mostly) pick out those aspects of the world that are relevant to the essential activities of the agent, and help to focus cognitive processing on those aspects. Those aspects picked out by such symbolic representations, by necessity, reflect the intrinsic needs/purposes/goals of the agent in its everyday activities (i.e., the intrinsic teleology forged by the evolutionary process).

Is it true that top-down learning is more practical and more useful?

Practically speaking, maybe. Given the culturally created systems of schooling, apprenticeship, and other forms of guided (or instructional) learning, top-down learning is quite prevalent in society. However, I insist that bottom-up learning is more fundamental. It is more fundamental in two senses: the ontological sense and the ontogenetic sense.

Ontologically, explicit conceptual knowledge needs to be obtained in the first place before it can be imparted to people to enable top-down learning. Therefore, bottom-up learning, which creates new explicit conceptual knowledge, is more fundamental. Only after bottom-up learning (and other types of learning) created explicit conceptual knowledge, can top-down learning be possible.

Ontogenetically, there seem to be some indications that children learn sensory-motor skills (as well as knowledge concerning concepts) implicitly first, and then acquire explicit knowledge on that basis. See, for example, Karmiloff-Smith (1986), Mandler (1992), Keil (1989), and so on. Therefore, bottom-up learning is also more important ontogenetically (or developmentally).

In addition, the study of bottom-up learning may lead to practically important advances in relation to the goal of artificial intelligence: for example, (1) self-learning autonomous systems (e.g., self-learning autonomous robots), (2) automated knowledge extraction in place of knowledge engineering, (3) systems for creativity that lead to the creation of new knowledge, and so on.

Is that true that there is a continuum from the explicit to the implicit, as opposed to a simple dichotomy of the explicit and the implicit? How can Clarion capture such a continuum?

Indeed, there appears on the surface to be a continuum from the completely explicit to the completely implicit, with many shades of gray in between. However, the two-level structure of Clarion, despite its two-level dichotomy, may fully account for such a “continuum”.

For example, to account for completely inaccessible (completely implicit) processes (such as visceral processes), a module in the ACS with a well developed bottom level but without a corresponding top level may be posited. This module will have a well-developed implicit process running, but it will not have any explicit representation available. Thus it can never become explicit.

For another example, to account for completely explicit processes, a module may be used in which there is a top level but no corresponding bottom level. Thus, the module will have explicit processes, but not implicit processes.

In between, there can be modules with both a top level and a bottom level, thus involving both explicit and implicit processes. Some such modules may have a well developed top level (with rich representational structures and contents there) and thus are more explicit, while some other modules may have very little structure and content available within their top levels and thus are less explicit (relatively speaking).

Different degrees of explicitness among different modules may also be (partially) determined by the availability and applicability of algorithms for acquiring explicit representations (such as RER) and for applying explicit knowledge: When such algorithms are more available or more applicable to a particular module, the module becomes more explicit.

In addition, the integration parameters (that regulates the integration of outcomes from the two levels) may be adjusted to involve different proportions of explicit and implicit processes for any specific task, which change the explicitness of a module during task performance.

Is implicit learning controversial? If so, how can you base your model on implicit learning?

Implicit learning is admittedly a somewhat controversial topic. But the existence of implicit processes in skill learning is generally not in question—what is in question is their extent and importance (see Stadler and Frensch 1998, Cleeremans et al 1998). We allow for the possibility that both types of processes and both types of knowledge coexist and interact with each other to shape learning and performance, so we manage to go beyond the controversies that focused mostly on the minute details of implicit learning.

For example, some criticisms of implicit learning focused on the alleged inability to isolate processes of implicit learning (e.g. Perruchet and Pacteau 1990, Knowlton and Squire 1994, Shanks and St.John 1994). Such methodological problems are not relevant to our approach, because in our approach, we recognize that both implicit and explicit learning are present and that they are likely to influence each other in a variety of ways. Criticisms of implicit learning also focused on the degree of cognitive involvement in implicit learning tasks (e.g. Shanks and St.John 1994). These criticisms are not relevant to our approach either, because our approach makes no claim in this regard. Yet another strand of criticisms concerned the fact that implicit learning was not completely autonomous and was susceptible to the influence of explicit cues, attention, and intention to learn (e.g., Berry, 1991, Curran and Keele 1993, Stadler 1995). These findings are in fact consistent with our view of two interacting systems. Therefore, the controversies concerning implicit learning are generally not relevant here.

Is there an executive control function in your architecture?

First, note that this notion is somewhat loaded and controversial, and its theoretical status is less than completely clear (Logan 2003). Having said that, there is indeed some "executive control" in Clarion. However, this notion is somewhat generalized in Clarion and linked up to some other control functions. Executive control also becomes more distributed in Clarion.

For example, what the action-centered subsystem (the ACS) controls includes not only the control of some cognitive processes (in another part of the architecture—namely, the NACS) but also the control of primary actions that affect the external world. That is, in Clarion, the control of internal processes and the control of external processes may be one and the same. They both result from the decisions of the action-centered subsystem (the ACS). The decision made by the ACS may be concerned with external actions or may be concerned with internal actions (i.e., "executive control" of memory and reasoning).

On top of that, there is meta-cognitive monitoring and meta-cognitive control, in the meta-cognitive subsystem (the MCS) of Clarion. The MCS may alter the functioning of other subsystems. Some researchers may regard some of these functions performed by the MCS as also belonging to "executive control". In that sense, executive control is distributed between the ACS and the MCS in Clarion.

The justification for this approach may rely on two design principles for cognitive architectures: (1) completeness of functionality (to include as many functionalities as possible), but (2) parsimony of mechanism (to reduce the number of distinct mechanisms as much as possible). Completeness of functionality requires the inclusion of meta-cognitive mechanisms in the architecture, the distinctiveness of which leads to the separate MCS, and parsimony requires that the ACS be used for the control of internal processes as well as external processes, when the control processes are similar.

So, the upshot is that executive control in Clarion is layered and distributed. And there is no homunculus in the model either.

What is working memory in Clarion?

First, the specific component named "working memory" in Clarion is narrower in scope than the general notion of working memory as has been used in the psychological literature. This re-definition is necessary for precisely specifying a cognitive architecture, because the notion of working memory has been vague in the literature and it has been used to denote a number of quite different phenomena.

Second, despite the narrower definition of working memory in Clarion, the cognitive architecture can account for many "working memory" phenomena, either through using the working memory component as defined in Clarion or through using other components or mechanisms in Clarion. (For example, it accounts for the limited capacity of working memory, the need for refreshing working memory, the limited number of explicit hypotheses that can be entertained at the same time, the limited ability to deal explicitly with long-range temporal dependencies, etc.)

Third, the working memory as defined in Clarion is neither solely implicit nor solely explicit. It is both at the same time.

Why is the motivational subsystem necessary?

The motivational subsystem is necessary, because first and foremost it represents the intrinsic motivations of human behavior, which have been forged through a very long evolutionary process (and it also represents other, derived motivations that add to the human behavioral complexity and flexibility). In essence, it represents the innate inclinations and capacities of the humankind — the crystallized history of the struggle to survival by the humankind.

Computationally speaking, with the motivational subsystem, the whole Clarion system together carries out so-called unsupervised learning, as opposed to supervised learning or reinforcement learning, because the supervision or reinforcement needed for the various learning algorithms used within Clarion is hence generated entirely internally, within Clarion itself. This feature makes Clarion much more cognitively realistic.

Why are there "goals" in addition to "drives" in the Clarion architecture ?

The drives provide the context in which goals are set and executed. Note that "drives" here refer to the desire to act in accordance with some *perceived* deficits or needs, which may or may not be physiological and may or may not reduce the perceived deficits/needs (cf. Hull 1951). It is a very generalized notion that provides essential underlying motivations for actions in an implicit and embodied fashion. On top of such implicit and embodied motivations, explicit goals are set, which are unique, explicit, and specific. For example, (1) there may be multiple drives being activated at the same time (e.g., being hungry and being thirsty at the same time). However, there is normally only one goal being pursued at a time (Anderson and Lebiere 1998; although a goal may encode multiple action objectives, that is, having multiple dimensions; see Sun 2003 for details). (2) Drives are more diffused in terms of focus, while goals are more specific (McFarland 1989, Anderson and Lebiere 1998). (3) Drives are more implicit (in terms of accessibility and representation), while goals are more explicit (Hull 1951). (4) Drives often tend to be hardwired, while goals are more flexibly set and executed (Hull 1951, Sun 2003). (5) Drives may be transient, while goals are longer-lasting (Sun 2003).

Does the meta-cognitive subsystem cover all meta-cognitive processes? Should meta-cognition be explicit by definition?

First, the MCS covers only some most basic meta-cognitive processes. Other "meta-cognitive" functions may be carried out by other subsystems such as the ACS. The normal use of the term "meta-cognition" is much broader than what we intended here.

Second, we do not believe that meta-cognition should be exclusively explicit or exclusively implicit. As other cognition functionalities, it should be a combination of explicit and implicit processes in general. On some occasions, it may be rather explicit, while in some others implicit. This variability can be seen in prior experimental and theoretical work (such as Reder 1996 and Mazzoni and Nelson 1998). In this regard, we disagree with, for example, A.Sloman, who described meta-cognition as an explicit process.

Can the meta-cognitive subsystem be considered a part of the action-centered subsystem, since their processes are similar?

It is true that the functionalities of the two subsystems are somewhat similar: Both involve making (internal or external) action decisions based on input information. However, content-wise, they are quite different: the meta-cognitive subsystem is solely concerned with a limited range of meta-cognitive control actions (as described before), while the action-centered subsystem is concerned with other types of actions. If we ignore that content difference, then indeed the meta-cognitive subsystem may be considered (standalone) modules of the action-centered subsystem. However, for conceptual clarity, it might be better off to view it as a separate subsystem. Either way of seeing the meta-cognitive subsystem is fine and does not change the essential structure and the essential operations of CLARION.

What about emotion? Should there be a place in CLARION **for emotion?**

There are reasons to believe that emotion is the collective result of operations throughout a cognitive system. It should not be viewed as a unitary thing. It may involve action readiness, physiological reactions, physical (external) actions, motivational processes and evaluations, as well as reasoning (e.g., related to attribution) and decision making.

What determines the explicitness/implicitness of processing when we face a particular task? How and why is a task "assigned" to one level or the other (or both)?

Although I touched upon this issue before, appealing to a generic notion of complexity (Sun 2002), we need to explicate this notion and thus draw a more detailed picture of the division of labor between the two levels. In general, when the task to be learned is complex, explicit learning mechanisms may not work well in learning the task, and therefore implicit learning becomes prominent. In contrast, when the task is simple, explicit learning mechanisms may work well, and therefore explicit learning becomes prominent instead.

We can speculate on the following factors determining complexity:

- Number of input dimensions. The higher the number, the more prominent implicit learning is (i.e., the more "holistic" the processing is). For example, see the review by Seger (1994). Computationally, the same effect has been demonstrated: for example, comparing Sun and Peterson (1998a) with Sun and Peterson (1998b) shows that a simpler (maze running) task with fewer input dimensions led to more explicit knowledge than a more complex (minefield navigation) task with more input dimensions.
- Number of output dimensions. The higher this number, the more prominent implicit learning is, by analogy to the number of input dimensions.
- Stochasticity. The more stochastic a task is, the more likely that implicit learning will be prominent. See, for example, DeShon and Alexander (1996) and Cleeremans and McClelland (1991) for some human data alluding to this issue.
- Sequentiality, that is, whether a task is sequential, how long a typical sequence is, and how distant the dependency relations are. The more sequential a task is, the more prominent implicit learning is. For example, we can compare the data of Lewicki et al (1987) and Willingham et al (1989) in this regard: the sequences were longer and more complex in the task of Lewicki et al and there was less explicit learning and more implicit learning. The same goes for Mathews et al (1989), in which using complex finite state grammars led to more implicit learning while using simple correspondence rules led to more explicit learning. See also Reber (1989) and Berry and Broadbent (1988) for similar results.
- Complexity and saliency of (correct) input/output mappings. [1]

[1] Complexity may be measured by, for example, (1) the minimum encoding length of explicit knowledge that is necessary for performing a task explicitly, and (2) the learnability of of such knowledge. Both are formal mathematical measures of complexity. The latter measures the complexity of learning and the former the complexity of the outcomes of learning, for example, the size of the minimum rule set needed to perform a task and the complexity of learning it. See Mitchell (1998).

Obviously, the more complex or the less salient the input/output mapping of a task is, the more prominent implicit learning will be (e.g., Halford et al 1998). We can again, for example, compare Lewicki et al (1987) and Willingham et al (1989) in this regard: There were a lot more input/output rules in the task of Lewicki et al, and there were more implicit learning there. See also Stanley et al (1989), Berry and Broadbent (1988), and Lee (1995) for demonstrations of this factor. Note that complexity of input/output mappings is certainly highly correlated with other factors listed above.

- Amount and type of instructions (given prior to or during task performance). Generally speaking, the more explicit instructions are given about a task, or the more explicitly focused the instructions are, the more prominent explicit learning will be. See, for example, Stanley et al (1989) and Sun et al (2001) for human data demonstrating this factor. Also consider conditions of verbalization, explicit search instructions, and explicit how-to instructions in various human experiments, as explored in Sun (2002).
- Multiplicity of tasks. Generally, under dual-tasks conditions, implicit learning becomes more prominent, compared with single-task conditions. See, for example, Sun et al (2001), Stadler (1995), Nissen and Bullemer (1987), and Szymanski and MacLeod (1996) for human data demonstrating this factor.

Our experimental findings with Clarion were highly consistent with the above conjectures. In addition to the simulations reported here, see also more systematic demonstrations in Sun and Peterson (1998a, b) (concerning the factors of numbers of input/output dimensions, complexity, and sequentiality) and Sun et al (2001) (concerning the factors of task multiplicity and instructions), as well as Sun (2002).

When does bottom-up learning happen and when does top-down learning happen?

Bottom-up learning happens when it is easier to learn implicit knowledge than explicit knowledge (see the discussion of factors determining that earlier), and it is possible to learn explicit knowledge on the basis of implicit knowledge.

For example, in a complex (or non-salient) process control task, one often develops implicit knowledge first, given that the situation makes directly learning explicit knowledge difficult. However, after a substantial amount of implicit knowledge accumulates, explicit knowledge often emerges on that basis (Stanley et al 1989, Sun et al 2001, 2005).

Top-down learning usually occurs when explicit conceptual knowledge is available from external sources, or when it is relatively easy to learn such knowledge (compared with learning corresponding implicit knowledge). Such knowledge, learned or directly received from external sources, is then assimilated into an implicit form.

For example, learning to play chess would be a good illustration. One often first learns the basic rules of chess, and some essential guidelines as to what to do in prototypical situations. One may then develop more complex and more nuanced knowledge that is largely implicit.

Of course, learning directions may vary from individual to individual and real-life learning scenarios are often much more complex than the examples above (Dreyfus and Dreyfus 1987).

How is the synergy between the two separate, interacting components of the mind generated?

Clarion may shed some light on this issue, by allowing systematic experimentations with the two corresponding levels in the model.

For example, Sun and Peterson (1998 b) did a thorough computational analysis of the source of the synergy between the two levels of Clarion in learning and in performance. Their conclusion, based on the systematic analysis, was that the explanation of the synergy between the two levels rests on the following factors: (1) the complementary representations of the two levels (discrete vs. continuous); (2) the complementary learning processes (one-shot rule learning vs. gradual Q-value/weight tuning); and (3) the bottom-up rule learning criterion used in Clarion. (See Sun and Peterson 1998b for details.)

It is very likely, in view of the match between the model and the human performance, that the corresponding synergy in human performance results also from these same factors (in the main). The analysis in Sun et al (2005) (in sections 5.3–5.6) identified distinct characteristics of the two levels similar to the above three factors. It is conceivable that these same characteristics contribute to the generation of synergy in human performance.

From the point of view of machine learning theories, the combination of a set of diversified processes may lead to better performance overall. Mathematical proofs have been provided in a number of cases in the machine learning and statistics literatures. However, this idea may be more generally applicable than these special cases, and is certainly consistent with our view here.

Can communication be accounted for in Clarion?

Communication is just a type of action from the perspective of Clarion. Therefore, Clarion should be able to account for it through the use of the ACS (along with the NACS possibly). However, currently, there is no built-in mechanism in Clarion for addressing some language-specific issues such as syntactic and semantic processing (although they may be added).

Can language comprehension and generation be added to Clarion?

Yes, although they are not currently included.

Can sensory-motor processes be added to Clarion?

Currently, these processes are not available in Clarion. However, they can certainly be added into the architecture, at a certain level of abstraction, for example, at the level of ACT-R/PM or at a more detailed level.

In fact, there has been some rudimentary implementation of sensory-motor processes in Clarion based on EPIC.

Why does your description of the details of the cognitive architecture seem vague or convoluted, or otherwise unclear?

The question regarding the clarity of technical details is a difficult one. Computational models are inherently complex and detail-rich. Every algorithmic steps and every data structure may require much effort to explain to laypersons. Some details may require a significant background in computer science to understand. It is, therefore, not possible to explain every computational detail of a cognitive architecture, completely, in a tutorial fashion, because of (1) the highly technical nature of some of the algorithms used, (2) the diversity of the background of readers, and (3) practical issues such as length limitations.

Why is a unified implementation of Clarion not adopted?

It is certainly desirable to have a unified implementation of a model, a theory, or a cognitive architecture. But the question is: Is connectionism (or symbolicism) the best medium for computational modeling of all parts of the mind, rather than just some parts?

Naturally, we can expect to be able to implement all symbolic processes in connectionist models or vice versa. But what does that kind of implementation buy us? I would prefer to use the most suitable tool for each part of my job. Similarly, I would prefer to use the best medium for implementing each part of my model.

At first glance, it may not seem "elegant" to use hybrid models involving both symbolic and connectionist techniques. However, notice that, looking through the vast literature on this, most connectionist implementations of sufficiently complex symbolic reasoning (or some other types of symbolic processing) are not "elegant" in any sense, and vice versa. Therefore, the use of hybrid models did not actually introduce any additional "inelegance", but only what is minimally necessary for capturing the complexity of the human mind.

Can CLARION be disproved?

I very much like Kuhn's and Lakatos' notions of scientific work. Consequently, I do not believe in the simplistic notion of proving or disproving of broad scientific frameworks.

According to Kuhn, on the assumption that a current theory is consistent and correct, observations are collected and fitted within the current theory. New and unexpected phenomena may be uncovered in the process, and they may lead to revision and refinement of the existing theory. In cognitive modeling, architectural assumptions and other commitments constitute an initial theory, which undergoes testing and validation through matching with data. Revision and refinement are undertaken when inconsistencies and incorrect predictions are discovered, or when the model is incapable of predicting something. However, when given a sufficiently high degree of mismatch between the data and the current architecture, that is, when revision and refinement are no longer able to accommodate problems that arise, a crisis may develop, which leads to a new "paradigm", that is, new architectures or even new approaches towards building cognitive architectures.

According to Lakatos, scientific growth should be assessed in terms of progressive and degenerating research programs. A research program consists of methodological rules: Some tell us what paths of research to avoid (negative heuristics), and others what paths to pursue (positive heuristics). An important heuristic is to devise conjectures that have more empirical content than their predecessors. The best programs in the growth of science are characterized by this kind of continuity. A research program is degenerating only if it does not generate new hypotheses that have more empirical content. So, as long as a research program (such as a cognitive architecture) is making progress toward theories of more and more empirical coverage, it is likely to be on the right track.

Therefore, the original question above is rather ill-posed in this context, and thus (almost) irrelevant.

Is it true that a more uniform and more constrained model provides deeper explanations rather than those with a large pool of ad hoc, specialized mechanisms?

It is true that in many cases, a more uniform and/or more constrained model provides deeper explanations than those with a large pool of ad hoc, specialized mechanisms, given that the empirical coverages (as well as other relevant aspects) of these models are comparable.

However, in the case of existing cognitive architectures, this argument failed to take account of the fact that there are severe limitations in those existing cognitive architectures that are more uniform and more constrained in terms of the range of cognitive (including decision making, learning, reasoning, memory, motivational, and meta-cognitive) phenomena that those cognitive architectures can capture.

If a cognitive architecture fails to capture the breadth of cognitive phenomena, then there is very little to be gained in being "uniform" or "constrained", because no "deep" explanations come out of it when it cannot explain many of the phenomena.

1.2 Comparisons with Other Models

How is CLARION different from ACT-R?

There are quite a lot of differences between CLARION and ACT-R. However, some of the differences stem from the difference in emphasis, while others are more substantial.

To enumerate just a few major differences:

- In ACT-R, there is no principled distinction and separation between implicit and explicit knowledge. And there is no deep and principled explanation of the distinction between implicit and explicit knowledge (e.g., based on a representational distinction, say, between symbolic/localist and distributed representations), and additional, ad hoc assumptions have to be made regarding which particular component is explicit or implicit. As a result, ACT-R does not capture well the psychological process of the interaction between implicit and explicit processes. It provides no direct explanation of *synergy* effects between the two types of knowledge.
- ACT-R is not meant for autonomous learning, without a lot of a priori knowledge to begin with. It does not directly capture the psychological process of bottom-up learning either.
- Also as a result of the dual representational structure, CLARION is capable of "automatic" and effortless similarity-based reasoning, while ACT-R may have to use costly and cumbersome pairwise similarity relations to enable similarity-based reasoning.
- CLARION uses distributed representation (in the hidden layers of the bottom level) and consequently has a general functional approximation capability (in its bottom level), as has been shown mathematically before.

- In ACT-R, there is no sufficiently psychologically realistic, built-in modeling of motivational processes. As a result, goals, in a sense, are always externally set and directly hand-coded. It does not reflect the diversity and flexibility of human motivations and behaviors.
- In ACT-R, there is no well developed, built-in metacognitive process.
- ACT-R has some detailed sensory-motor modules that CLARION currently does not include.
- CLARION and ACT-R often account for different tasks, although there have been some overlaps also.

How is CLARION different from Soar?

In Soar, based on the framework of a state space and operators for searching the space, decisions are made by different productions proposing different operators, when there is a goal on a goal stack. When a sequence of productions leads to achieving a goal, chunking occurs, which creates a single production that summarizes the process (using explanation-based learning). A large amount of initial (a priori) knowledge about states and operators is required. Hence the learning process can be characterized as being top-down, not bottom-up.

Soar is different from CLARION, also because Soar makes no distinction between explicit and implicit learning, and its learning is based on specialization (top-down learning of a sort), using only symbolic representations. There is no (built-in) modeling of the psychological process of bottom-up learning as a result. There is no (built-in) modeling of the psychological process of the interaction and synergy between the two types of processes either.

In Soar, there is no distinction between symbolic/localist and distributed representations (as discussed earlier). Therefore, it does not embody similarity-based reasoning processes resulting from their interaction. Also, due to the absence of distributed representation, there is no sufficient function approximation capability either (see the previous answer for more discussions of this point).

Finally, in Soar, there is no sufficiently complex and psychologically realistic (built-in) motivational process. Nor is there sufficiently complex and psychologically realistic (built-in) meta-cognitive process.

How is CLARION different from Hunt and Lansman (1986)?

The implementation of the two types of knowledge (implicit and explicit) is similar between CLARION and Hunt and Lansman's model. The production system in Hunt and Lansman's model clearly resembles the top level in CLARION, in that explicit rules are used in much the same way. Likewise, the spreading activation in the semantic network in Hunt and Lansman's model resembles spreading activation in the bottom level of Clarion. However, learning directions are different between these two models. While learning in Clarion is mostly bottom-up but capable of being top-down, learning in Hunt and Lansman's model is completely top-down: That is, the working of the production system is assimilated into the semantic network, but the opposite process is not available. This makes their model more limited than Clarion (which is capable of both directions). Schneider and Oliver (1991) employed essentially the same idea for capturing automatization data. Logan (1988), which was also meant to capture automatization data, was also somewhat similar in this regard.

How is Clarion different from Schneider and Oliver (1991)?

Schneider and Oliver (1991) was concerned with automatization in skill learning. Automatization refers to skill learning that goes from explicit processes to implicit processes, that is, the exact opposite of bottom-up learning. A deliberate (explicit) calculation was performed first but later, through repeated trials, an automatized (implicit) process took over. Both processes were implemented in neural networks. Their model thus implemented the psychological process of top-down learning, in which explicit knowledge was assimilated into implicit skills. While Clarion can handle both top-down learning and bottom-up learning, their model is limited to top-down learning.

1.3 Simulations

Is there any ecological validity to simulating these artificial laboratory tasks?

Yes. Although some of these tasks may seem artificial, we believe that they tap into some essential cognitive processes that are ecologically relevant or important. That is why simulating data from these tasks may indeed shed light on essential cognitive processes.

The "artificiality" of these tasks may sometimes be the results of the attempts to design psychological tasks that tease apart some essential cognitive processes in an effort to better understand them. "Real world" tasks may seem more ecologically valid, but they may not serve as well the purpose of teasing apart and understanding essential cognitive processes.

In addition, the "artificiality" of some of these tasks may sometimes be the result of the deliberate effort at avoiding the contamination of a priori knowledge on performance or at exploring bottom-up learning without a priori knowledge to begin with.

Can you compare your architecture with other existing

models quantitatively and numerically through simulation?

Comparing, quantitatively and numerically, the simulations by Clarion with the simulations by other existing models is often difficult for the following reasons: (1) existing models often dealt with different tasks and therefore not directly comparable to Clarion (because each of them is often eager to demonstrate its capabilities to capture and explain new data and new phenomena), (2) many existing models focused on different issues, not directly related to the focuses of Clarion, (3) some existing models were highly specialized, dealing only with one task or even one data set, and therefore it is difficult to compare them with Clarion, which is a generic cognitive architecture that is coarser by necessity.

Nevertheless, in our prior and current work, we often did include numerical comparisons of our simulations with the simulations conducted by others using other models, when there were indeed overlapping coverages in a few cases (see Sun 2002, Sun et al 2005).

Why did you not discuss mismatches between your model and human data in various tasks?

I did mention some mismatched aspects of the Clarion simulations and the corresponding human data. However, it is not possible to get into all the subtlety of all the tasks involved. We have mostly focused on a few important issues (such as implicit versus explicit learning). As such, we have to center our discussions on these issues, and avoided being sidetracked. Thus, we have avoided discussing details of many related but tangential issues. Some mismatches may be understood in this light.

Also, it is important to avoid capturing noise along with meaningful data in computational (and mathematical) modeling. Hence, too fine-grained matching may not be that desirable from such a perspective.

In addition, Clarion is a broad-based model, and it accounts for a broad range of data. Thus it is justified to be more complex and at the same time coarser in matching some human data than more specialized, more narrowly-scoped models. See also the answer to the next question.

Why did you not capture finer details of some of the simulated tasks?

It should first be noted that our initial goal regarding Clarion was to show the effect of the interaction between the two types of learning (implicit versus explicit). Thus, for example, capturing finer details of learning curves and other characteristics was not our focus.

Having said that, another of the main reasons why finer details of task performance are not captured in some instances is that details of human data of these tasks are often not available. Thus, a detailed statistical analysis of matching between CLARION and human data is impossible, which makes finer matching unattainable. As a comparison, many other projects took a similar approach; for example, see McClelland, McNaughton and O'Reilly (1995).

Yet another reason is that, as a generic cognitive architecture, CLARION is a broad-based model. The architecture is broader, and it accounts for a broader range of data, compared with many specialized models. Thus it is justified to be more complex and at the same time coarser in matching some human data.

Besides, we are more interested in general phenomena than idiosyncratic patterns in individual data sets. So we avoid getting trapped by idiosyncratic details of individual tasks or data sets (and noises inevitably embedded in them), and focus instead on multiple tasks and/or multiple data sets in order to extract general principles or general phenomena that are applicable to a broad range of tasks (through generalizing over a set of tasks or data sets). The reader may find the arguments for such generic cognitive architectures in, for example, Newell (1990) and Anderson and Lebiere (1998). Generic cognitive architectures have their inherent advantages and shortcomings.

What accounts for the match you did obtain between the simulations and the human data? Is it the architecture itself or is it the parameters?

What accounts for the match between the model and the empirical data (and thus what constitutes the focal point of the explanation of the data) varies from task to task. Different tasks (or different types of tasks) may require different amounts of details in order to obtain a match. Some tasks are easier to model than others — maybe only a minimum number of mechanisms would be required, while others are more difficult to model and require a large set of mechanisms. Some tasks are highly sensitive to parameter settings, while others are not. And so on.

Are there any useful interpretations generated by the simulations? What are they if any?

Yes. There have many useful explanations or interpretations generated by the simulations using CLARION. The computational mechanisms, processes, and parameters in CLARION provide detailed, process-based (i.e., computational) explanations of why human data are what they are, the source of performance variations, or effects of manipulations.

For example, in Sun et al (2005), we showed that the verbalization effect, the explicit how-to instruction effect, the explicit search effect, and so on could all be explained based on the explicit-implicit interaction.

Can an one-level model capture all of these human data you simulated?

Although it is conceivable that a one-level model may be designed so as to capture all the data we simulated, we failed in our experiments to do so. However, the human data indeed does not unambiguously point to the Clarion architecture. It is still possible that some one-level models may work.

One may argue that if a one-level model can account for the data, then there is no need for the second level. However, note that it is seldom, if ever, the case that human data can be used to demonstrate the *unique* validity of a cognitive architecture. We need to rely on converging evidence from various sources, including, for example, philosophical arguments (such as those outlined earlier), to justify a model. By such a standard, this cognitive architecture fares well.

Is there any evidence that some of these high-level tasks (such as Tower of Hanoi) involve implicit processes at all?

In general, it is well known that even high-level cognitive tasks may involve implicit processes. In fact there has been some evidence that Tower of Hanoi, categorical learning, reasoning, and so on may indeed involve implicit processes. For example, Gagne and Smith (1962) showed specifically that verbalization improved performance in learning Tower of Hanoi. Bower and King (1967) showed the same effect of verbalization in classification rule learning. Gick and Holyoak (1980) found that good problem solvers in high-level problem solving domains could better state rules that described their actions in problem solving. In all of these cases, it could be the explication of implicit knowledge that helped the performance.

Some more direct evidence may be found in Dreyfus and Dreyfus (1987). Dreyfus and Dreyfus argued that learning to play chess involves turning analytic thinking into intuitive (implicit) thinking through extensive practice.

Evans (2005) showed some evidence and arguments that even deductive reasoning might be partially implicit.

It is well known that even mathematical theorem proving involves intuitive, implicit thinking to a very significant extent. For example, in mathematical theorem proving, it is extremely important to develop good intuition in order to narrow down search spaces in constructing proofs. The search space of different possibilities of constructing a mathematical proof is huge, and explicit exploration of the space is prohibitive in terms of cost. Therefore, intuition is crucial in guiding the search implicitly and efficiently. Intuition is often implicit, as shown by, for example, Lewicki (1986), Hasher and Zacks (1979), and so on. In a sense, good (implicit) intuition is what separates a good mathematician and a poor one.

Given the length of this document, it would be unwise to add lengthy discussions of these points, which would be needed if we were to fully justify the implicit nature of these tasks.

If simulation using a cognitive architecture can generate almost any behavior, including humanly implausible behaviors, then what is the point of simulating psychological data?

It is not true that simulation can produce almost any behavior. A cognitive architecture is *designed* to generate those and only those behaviors that are humanly plausible (although this lofty goal may not have been achieved one hundred percent). Once the general framework is in place, parameters may be further tuned to correspond even better with observed human data, by narrowing down the ranges of the parameters. Parameter tuning can be accomplished through a variety of means, such as theoretical derivation, as well as empirical estimation. See my answers to later questions regarding finding the right structures for simulating particular tasks for further details.

Are there too many free parameters in your architecture?

At first glance, the Clarion cognitive architecture may seem to have too many parameters. However, upon closer examination, the number of its parameters (e.g., in a subsystem such as the action-centered subsystem) is not significantly higher than usual computational models such as backpropagation networks.

We may look specifically into the action-centered subsystem for example, which we used to simulate a large variety of tasks. In addition to parameters of backpropagation networks (as in the bottom level of the ACS), at the top level of the ACS, there are only three important parameters concerning rule extraction and revision (when using RER). That is to say the architecture is approximately comparable to backpropagation networks in complexity.

Furthermore, although the values of all of these parameters affect performance, most of them were not changed throughout the simulations of various conditions of a particular task, and thus they should be treated as part of the fixed model specification. In this sense, they are not free parameters. More specifically, they do not contribute to the degree of freedom that we have to match the change of performance across different conditions in a particular task by human subjects (see, e.g., Sun et al 2001, 2005).

There are in fact three different types of parameters in Clarion: (1) domain-independent parameters, (2) domain-specific parameters, and (3) free parameters. The first two types are usually viewed as part of the fixed model specification for any particular task. Free parameters are those that are *changed* for capturing different experimental conditions of a task. In the past simulations involving Clarion, the actual number of free parameters was usually only one or two (usually a rule learning threshold or two at the top level). See, for example, Sun et al (2001, 2005). Also see the following paper for a similar perspective on this issue: McClelland, McNaughton and O'Reilly (1995).

How do we find the right structure (e.g., a right set of rules) for simulating a task, since different structures may simulate a task equally well on some measures (such as behavioral outcomes) but lead to quite different conclusions on some other measures (such as response time)?

First, recall that Clarion is capable of autonomous and bottom-up learning, often without domain-specific a priori knowledge to begin with. Therefore, one may begin with a minimum initial structure and use the learning capabilities of the architecture to acquire knowledge (if the task to be simulated allows such an approach). This approach, when it is applicable, eliminates most of the ambiguities in constructing a simulation model.

Second, when a priori domain-specific structures (such as explicit rules at the top level of the ACS) are necessary, some rules of thumb may be followed. One general approach is to make a priori domain-specific structures (e.g., rules) as parsimonious as possible while still keeping them psychologically plausible. (Some cognitive architectures may also have additional modeling policies that partially determine how rules should be constructed.)

Another general approach for coming up with cognitively plausible a priori domain-specific structures (e.g., rules) is to write two sets of rules that "bracket" the observed data, where one set represents a slowest but yet reasonable strategy, while the other set represents a fastest and reasonable strategy. The observed data should fall in between, on an averaged basis or even for individual subjects. Sometimes, one may analyze the data to determine which aspects of the two strategies were used by the subjects (on an averaged or even on an individual basis). See Gray and Boehme-Davis (2000) for a discussion of this idea.

In general, different resulting models represent different hypotheses, which may be adjudicated by empirical and/or theoretical means.

Third, one may even use machine learning techniques for automatic tuning of structures and parameters. For example, Slusarz (2001) used some quite sophisticated search algorithms for finding parameters that fit data well. However, so far, most such efforts focused on tuning parameters only, although structural tuning is also possible (see, e.g., Weng et al 2001).

In relation to simulating social phenomena, is it true that the social is more than just the sum of the cognitive? If so, how can Clarion **be very useful in simulating social phenomena?**

It is true that social dynamics is often “more” than individual cognition, in the sense that some social dynamics sometimes may not be easily predicted from the understanding of the details of individual cognition. However, it should be emphasized that

- social phenomena are manifested through individual actions (if nobody takes any action, then there will not be any social phenomena), and thus through individual cognition (which necessarily underlies individual actions);
- rather than just leaving the issue at being “emergent”, we want to investigate how social phenomena “emerge” from individual actions and cognition, thus producing deeper explanations—cognitively-based explanations;
- we also would like to investigate how different cognitive characteristics lead to different “emergent” social phenomena.

Therefore, it is not mistaken to try to understand, capture, and model social phenomena from the ground up—from the details of individual cognition.

Some may claim that the social is not reducible to the individual. But there is little reason to believe this claim: For example, thermodynamics can be reduced to Newtonian physics; chemistry can be reduced to physics; connectionist models can be reduced to nodes and links between nodes; ..... So, why can social phenomena not be reduced to individual interactions? If by “not reducible” it was meant that the reduction of the social to the individual is not a trivial matter and may require a significant amount of intellectual effort, or that additional laws may be usefully posited at the social level, then the answer is yes.

1.4 Theoretical Implications

Is Clarion **a theory, an implementation of a theory, or something else?**

Clarion in fact contains several different types of things.

First of all, it contains a core theory of the mind. For instance, it posits some essential theoretical distinctions such as implicit versus explicit, action-centered versus non-action-centered, and so on. With these distinctions, it posits a core theory of the essential structures and processes of the mind (as described before).

Second, it contains a more detailed (but generic) computational model implementing the theory. This implementation constitutes what is usually referred to as a cognitive architecture, that is, a generic computational cognitive model describing the architecture of the mind (which, by the way, also constitutes a theory of the mind, albeit more detailed, as will be argued later).

Third, with the cognitive architecture, one may construct specific simulation models of specific cognitive phenomena or processes. That is, one may "derive" specific computational cognitive models from the generic computational cognitive model.

Clarion contains all of the above simultaneously.

Does your cognitive architecture constitute a theory of the mind?

Definitely. There has been a well argued view that every computational model provides a theory (e.g., Newell 1990). A computational cognitive model is a formal description of relevant cognitive phenomena. The language of a model is, by itself, a distinct symbol system for formulating a theory (Newell 1990). No verbal-conceptual theory completely specifies the computational mechanism, let alone the dynamic process that may emerge. Thus, computational cognitive modeling is needed to describe these complex aspects in order to produce a computational simulation, which at the same time also produces a more precise and detailed theory. The language for computational modeling is, in essence, just another language for presenting a theory. Like verbal-conceptual theories or equation-based mathematical theories, computational models can be used to generate predictions. In fact, they can generate more precise predictions that can be more precisely tested. This position has been advocated by many in the cognitive modeling community (e.g., Anderson and Lebiere 1998, Sun 2002).

In relation to the question of the need to validate all the minute details of a computational model ("how do you justify or validate all these computational details?"), it is worth noting that there is a well-argued position in the philosophy of science known as constructive empiricism, which argues (roughly) that not all details of a scientific theory need to be strictly derived from empirical data (which is impossible anyway), but only the observable parts need to be mapped to the empirical data. It may make a more sensible philosophical foundation for computational cognitive modeling than the naive empiricist accounts and/or the Popperian methodology that some cognitive scientists seem to subscribe to. See, for example, van Fraassen (1980) for a detailed account of this position.

How does your cognitive architecture relate to mathematical theories on the one hand and verbal-conceptual theories on the other?

The difference between this theory and a mathematical theory or between this theory and a verbal-conceptual theory is a matter of descriptive medium, descriptive complexity, and descriptive style (see also the answer to the previous question).

Mathematical equations and computational cognitive models are both instances of the class of formal models. In this sense, they are not fundamentally different. But they are certainly different in some (less fundamental) ways. One difference is that of the languages they are based on: mathematical equations versus computer algorithms. [2] Another difference is that, due to the difference in language, mathematical models are simple to specify (in terms of length of description) while computational models often take longer descriptions to express. Yet another difference is that mathematical models are often in the closed form (i.e., with the relationship between input and output variables apparent) while computational models are often in the open form. However, issues of validation, matching, and prediction are common to all formal models, whether mathematical or computational.

Another perspective unifying various forms of scientific theories is centered on the notion of the descriptive complexity. A theory should represent our best knowledge regarding the nature of a class of phenomena. However, depending on domains, our best knowledge varies in terms of explanatory succinctness. In some cases, a small and rigorous set of equations are able to express the regularity of a domain to a sufficient extent, approximating it with an acceptable level of accuracy. For example, in physics, Newtonian classical mechanics is such a case. However, in some other cases, a succinct set of equations are not found that can express domain regularities to a satisfactory extent. In that case, a more complex form of theory may be required. Computational models are but a possible class of complex theory for such domains. Understanding the human mind is one such domain in which notably no simple form of theory is available.

Kolmogorov complexity is a theoretical/mathematical measure of complexity based on how many binary bits are needed on a theoretical model of computation (i.e., the Turing machine) to capture (i.e., to encode) a computational process (that is, to express an algorithm); or more loosely,

[2] Algorithms and program codes are viewed as being equivalent here.

Kolmogorov complexity measures the minimum length of the description of an algorithm. See Li and Vitanyi (1997) for technical details. It is a solid, though often neglected, foundation upon which we may compare different scientific theories. A key difference between different types of scientific theories (verbal-conceptual, mathematical, or computational) may be captured in terms of the descriptive length (Kolmogorov complexity) of a theory and, by extension, the numbers of individual entities and causal relationships required by the theory (Sun et al 2005b).

Finally, different types of theorizing may have different roles to play. For constructing a computational model, one often takes some specific contents of verbal-conceptual theories and tries to formalize them into a set of equations or algorithms. Moreover, a computational simulation model likely combines various theories, or various aspects of a theory, regardless of whether they are verbal-conceptual, mathematical, or computational. It therefore enables us to integrate different perspectives, for example, as "subroutines". That is, a computational model may specify when a subroutine with a particular set of equations or algorithms representing a theory or a particular aspect of a theory is called (Sun et al 2001, Sun 2002, 2003b). Thus, fragmentary (verbal-conceptual, mathematical, or computational) theories compete and cooperate with each other in the simulation, and they also compete and cooperate with each other in explaining simulation results.

How does your theory relate to the notion of declarative/procedural knowledge?

Let us explicate the relationship between the implicit/explicit distinction that is being emphasize here and the procedural/declarative distinction emphasized in some other theories. In Anderson (1983, 1993), procedural knowledge is represented in an action-oriented way (using production rules that can only be used in one direction — from conditions to actions), and declarative knowledge in a non-action-oriented way (with knowledge chunks that can be used in any possible direction). The difference in action-orientedness seems to be the main factor in distinguishing the two types, while explicit accessibility seems a secondary factor. [3]

This view of declarative knowledge unnecessarily confounds two issues: action-orientedness and accessibility, and can be made clearer by separating the two issues. In Clarion, there are both explicit and implicit action-centered (procedural) knowledge, and both explicit and implicit non-

[3] A common interpretation is that while procedural knowledge is inaccessible, declarative knowledge consists of both accessible symbolic representations and inaccessible subsymbolic representations.

action-centered (declarative) knowledge (Sun 2003). As demonstrated in Clarion, action-orientedness does not necessarily go with inaccessibility (see, e.g., Sun et al 2001), and non-action-orientedness does not necessarily go with accessibility either (e.g., in the case of priming and implicit memory; see, e.g., Schacter 1987).

Note that our perspective on this issue is closer to Hunt and Lansman's (1986).

How does your theory relate to the notion of automaticity?

The notion of automaticity has been variously associated with (1) the absence of competition for limited resources (attention) and thus the lack of performance degradation in multi-task settings (Navon and Gopher 1979), (2) the absence of conscious control/intervention/intention in processes (J.Cohen et al 1990), (3) the general inaccessibility of processes (Logan 1988), (4) the general speedup of skilled performance (Hunt and Lansman 1986). Although not focused on these issues, Clarion is compatible with them. The top level of Clarion can account for controlled processes (the opposite of these above properties), and the bottom level has the potential of accounting for all the afore-mentioned properties of automatic processes. In the simulations, we have in fact separately covered these issues (see Sun 2002, Sun et al 2005): the speedup of skilled performance, the direct inaccessibility of processes at the bottom level, including their ability of running without conscious intervention, and the lack of resource competition (due to the existence of multiple bottom-level modules that can run in parallel). Thus, in Clarion, automaticity serves as an umbrella term that describes a set of phenomena occurring in implicit processes at the bottom level.

On a related note, what is commonly referred to as automatic processing in the literature is often the result of top-down learning (Shiffrin and Schneider 1977), while implicit processes, in reality, are often (though not always) the beginning of bottom-up learning (Sun 2002, Sun et al 2005).

How does your theory relate to the notion of consciousness?

The implicit/explicit distinction bears clear relationships to the study of consciousness, because this distinction involves, in its core, the issue of awareness, which is also the key to consciousness. The study of the implicit/explicit distinction may help us to better understand issues concerning consciousness, by identifying physical or computational mechanisms and processes correlated with consciousness (Schacter 1987, Reber 1989, Sun 1997, Sun 1999, Dienes and Perner 1999). In this regard, Clarion may shed light on the question of what constitutes consciousness. Our central thesis has been that direct accessibility (i.e., explicit representation), along with explicit manipulability (on directly accessible, i.e., explicit, representation), constitutes the essence of consciousness (see Sun 1997, 1999; cf. Dienes and Perner 1999). Clarion naturally embodies the difference between accessibility and inaccessibility (i.e., explicit and implicit processes or representations) through the use of symbolic and distributed representations in the two different levels respectively, and provides a plausible grounding for the notion of accessibility and hence the notion of awareness. Although there are a variety of views concerning consciousness, each based on a different physical substrate, [4] Sun (1997, 1999) argued that the distinction between localist/symbolic and distributed representations provided a far superior alternative. All things considered, Clarion has significant bearings on theorizing on consciousness.

How does your theory relate to the notion of instance/exemplar in instance-based theories?

Logan (1988) showed that skill learning (automatization) could be captured by the acquisition of a domain-specific knowledge base that was composed of experienced instances represented in individuated forms (Hintzman 1986). Shanks and St.John (1994) developed a theoretical perspective in which implicit learning was viewed as nothing more than learning instances (however, this perspective has been criticized for various failings). Stanley et al (1989) also described implicit learning/performance as mainly the result of relying on memory of past instances, which were utilized by being compared to a current situation and being transformed into a response to the current situation (through similarity-based analogical processes). At first glance, these models may seem at odds with Clarion. However, upon a closer examination, it is clear that the connectionist networks used in the bottom level of Clarion can be either exemplar-based (essentially storing instances; Kruschke 1992) or prototype-based (summarizing instances; Rumelhart et al 1986), often depending on the parameters and structures of the connectionist networks. The similarity-based processes in these models can also be performed in connectionist networks, which are known to excel in such processes. Instance-based models, however, generally do not account for the learning of generic explicit knowledge, nor bottom-up learning.

How does your theory relate to the notion of situated/embodied cognition?

[4] There are of course also dualistic views that rely on the assumption of nonphysical properties, which we will not deal with here.

The situated/embodied cognition (including reactive planning) view claims that cognition (and sophisticated artificial intelligence for that) does not rely on a system that contains a single all-purpose general symbol processor, with overly complex (and cognitively unrealistic) symbolic representations. Instead, it contains a large number of more specialized systems that work together to achieve various types of functionalities, including functionalities that could emerge to produce operations equivalent to an all-purpose general symbol processor, in response to situations as perceived by agents. (Related views might be found in some neurally-inspired approaches to computational modeling, not only in connectionism that is somewhat divorced from neuroscience, but also in some neural networks models inspired by neuroscience.)

Such situated/embodied views are fully compatible with Clarion. In fact, it is the foundation of Clarion, as has been extensively argued in the book "Duality of the mind" (see Sun 2002). Briefly, Clarion does contain a set of specialized modules (subsystems and so on) interacting with each other. There is no central symbol processor in the traditional sense (Newell and Simon 1976). The operation of Clarion is the result of the interaction of various modules, which together give rise to complex cognitive phenomena. Clarion is closely coupled to, and acts in response to, situations as perceived. It avoids unnecessarily complex symbolic representations of belief, desire, and intention, and so on.

Moreover, in Clarion, perception itself is shaped by the interaction between the agent and the world, that is, by its actions and reactions in the world. Similarly, cognition in general is also shaped by the agent/world interaction and by the actions/reactions of the agent in the world.

However, Clarion goes beyond situated/embodied cognition, in the following ways: (1) Clarion addresses the existence and the importance of symbolic processes in human cognition (Sun et al 2005, Sun 2002; although no general-purpose, central symbolic processor is posited). (2) Clarion also addresses the emergence of symbolic processes from ongoing subsymbolic processes in interacting with the world (Sun et al 2001). (3) Clarion furthermore addresses the grounding of symbolic representations in subsymbolic processes and in ongoing interactions with the world (Sun 2000, Sun 2002).

How is Clarion related to and different from the enactive AI approach?

Clarion agrees with the following views of the enactive AI approach: (1) the agent is situated in the world and interacting with the world in a rather direct way, and this way of interaction is the basis of the agent's cognition, (2) it learns and adapts in the process of interacting with the world, (3) it is embodied physically, and this physical embodiment has significant ramifications for its cognition, (4) the agent is self-contained, self-sustained, and self-reproduced ("autopioesis"), (5) the agent and the world are co-determined by each other, in that the world is, in some sense, the projection of the agent, and the agent consists, largely, of the patterns of interactions with the world, (6) the behavior of the agent necessarily reflects the intrinsic teleology of the agent, which has been forged by the long evolutionary process.

However, Clarion goes beyond those views above, in that it hypothesizes and argues for the following points: (1) the innate dichotomy of implicit and explicit cognitive processes, (2) the dual-systems approach in realizing this dichotomy, (3) the importance of symbolic processes (in the dichotomy and in the resulting dual systems), (4) the importance of bottom-up learning (in the dichotomy and in the resulting dual systems), that is, the emergence of symbolic processes and representations from subsymbolic processes and representations, in the interaction between the agent and the world. See Sun (2000) and Sun (2002) for further details of these points.

How does your approach relate to the dynamic systems approach? Does the dynamic systems approach suggest that the approach of building cognitive architectures is wrongheaded?

It has been claimed by some proponents of the dynamic systems approach that there is no good example to point to as a success of understanding cognition based on cognitive architectures. For example, Clarion emphasizes the distinction between implicit vs. explicit processes, which, as it has been claimed, has a questionable empirical pedigree. Such two-kind distinctions show up periodically at conferences such as Psychonomics. They generate a lot of initial interest but after two or three annual conferences they are brought into question and are no longer generally trusted as distinct empirical categories, except by the "cult following" they inevitably attract.

However, to me, this phenomenon illustrates exactly why cognitive architectures are needed in cognitive science. This need arises because of the complexity and variability of psychological experiments and data. There are too many contextual factors and too many minute variations. Therefore, it is futile to try to understand the human mind purely through empirical means (e.g., see Sun et al 2005 in Philosophical Psychology for more extensive discussions of this point). Cognitive architectures are needed to provide some frameworks and to instill clarity.

Regarding the implicit-explicit distinction specifically, Clarion has indeed provided some clarity to the issues involved. See, for example, Sun et al (2005) in Psychological Review for detailed discussions of how experimental subtleties and variabilities, and even apparently contradictory empirical evidence, can be accounted for by CLARION. See also my earlier answers regarding the whole spectrum between the purely implicit and the purely explicit. Sun (2002) also provides extensive discussions in clarifying this distinction.

Moreover, some proponents of the dynamic systems approach have even claimed that, looking across the literature on cognition, the patterns of cognitive effects do not divide neatly among mental functions, processes, or representations. The patterns of cognitive effects can only be captured through interactions among histories of participants, stimuli, and cognitive factors, and the special circumstances of task demands, culture, language, and so on. They claim that apparently the essence of cognition is such context sensitivity.

On the contrary, in this regard, I would argue that the various distinctions in CLARION can account for complex and sometimes seemingly contradictory empirical findings from its complex internal dynamics. To the enthusiasts of the dynamic systems approach, CLARION is indeed a dynamic system, with many interacting components. Perception, categorization, representation (in a broad sense), memory of all sorts, decision making, reasoning, planning (in a broad sense), problem solving, meta-cognition, communication, action control and execution, and learning of all sorts, motivational processes, goal representation, and meta-cognitive processes all interact with each other in CLARION, and furthermore, their patterns of interaction change with changing task demands. In CLARION, all effects of all cognitive factors are in motion, so to speak, with respect to each other and with respect to the task contexts in which they are observed.

One may take this vast catalog of interactions at face value and make the claim that there is no "absolute, fixed frame of reference", and cognitive scientists may be better served by giving up the "misleading pursuit" of an absolute frame of reference. One may claim that researchers should instead pursue "context sensitive, statistical structures" of the mind, and avoid the "reductive logic" of cognitive science.

I would disagree with such claims. Even dynamic systems can be attributed to its constituting elements—otherwise the field of dynamic systems does not need to exist. For one thing, cognitive architectures do not have to represent "reductive logic", any more than any other possible implementation of dynamic systems. For example, neural networks and other learning algorithms in CLARION capture "context sensitive, statistical structures" very well in fact. Therefore, the pursuit of cognitive architectures is by no means mistaken. The exploration of cognitive architectures should indeed be integrated with the dynamic systems approach, but not replaced by it.

How does your theory relate to the notion of task-specific cognitive architecture?

Some have claimed that cognitive architectures need not be totally generic. There can be a cognitive architecture for a particular specifiable range of knowledge domains and applications, but not all domains. An architecture only needs to be sufficiently generic for some class of domains. For example, connectionist models may be appropriate architectures for perceptual processes in many modalities and low-level execution of actions, but symbolic production system architectures are natural for simple heuristics for problem solving or reasoning. In principle, architectures can range in generality.

I disagree with the view above. If a model is not meant to be totally generic, I would not call it cognitive architecture. One may refer to such models as "generic models for perceptual modeling", or something along that line. Of course, no model has so far achieved complete domain-generality, but it remains the ultimate goal nevertheless.

What distinguishes humans from, say, monkeys according to the CLARION framework?

The following distinguishing features may be hypothesized in this regard within the CLARION framework: (1) in humans, there are much more extensive explicit representation in various subsystems, including the action-centered subsystem (the ACS), the non-action-centered subsystem (the NACS), the motivational subsystem (the MS), and the meta-cognitive subsystem (the MCS); (2) in humans, there are much more extensive explicit reasoning abilities in the non-action-centered subsystem; (3) in humans, there are much better developed meta-cognitive abilities in the metacognitive subsystem; (4) in humans, there are more complex motivational representations and dynamics. .... These differences, however, are generally quantitative rather than qualitative.

Do the explanations provided above regarding various prior theories seem forced? That is, are they being twisted to fit into the framework of CLARION?

I believe that CLARION provides a comprehensive and conceptually clearer (and thus more convincing) theoretical framework. As such, it encompasses many existing theoretical frameworks and provides clarifications to and explanations for them.

I do not believe that prior theoretical frameworks have been forced to fit the CLARION framework. Rather, CLARION provides conceptual and theoretical clarifications, theoretical interpretations, and sometimes grounding, for them. Thus, some conceptual re-interpretations necessarily take place regarding these prior frameworks.

For people deeply engrossed in some of these prior theoretical frameworks, this process may appear to be forced at first glance. However, this is nothing unusual. Many extremely successful scientific theories in the history of science appeared that way to many when they were first proposed. Examples from the past include Ptolemaic theory of planetary movements, Darwinian theory of evolution, and so on. It often takes time to get used to new theoretical frameworks.