BEHAVIORAL AND BRAIN SCIENCES (2003) **26,** 000–000
Printed in the United States of America

The Newell Test for a theory of cognition

John R. Anderson
Department of Psychology–BH345D, Carnegie Mellon University, Pittsburgh, PA 15213-3890.
ja+@cmu.edu **http://act.psy.cmu.edu/ACT/people/ja.html**

Christian Lebiere
Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15213-3890.
cl@cmu.edu **http://www.andrew.cmu.edu/~cl**

Abstract: Newell (1980; 1990) proposed that cognitive theories be developed in an effort to satisfy multiple criteria and to avoid theoretical myopia. He provided two overlapping lists of 13 criteria that the human cognitive architecture would have to satisfy in order to be functional. We have distilled these into 12 criteria: flexible behavior, real-time performance, adaptive behavior, vast knowledge base, dynamic behavior, knowledge integration, natural language, learning, development, evolution, and brain realization. There would be greater theoretical progress if we evaluated theories by a broad set of criteria such as these and attended to the weaknesses such evaluations revealed. To illustrate how theories can be evaluated we apply these criteria to both classical connectionism (McClelland & Rumelhart 1986; Rumelhart & McClelland 1986b) and the ACT-R theory (Anderson & Lebiere 1998). The strengths of classical connectionism on this test derive from its intense effort in addressing empirical phenomena in domains like language and cognitive development. Its weaknesses derive from its failure to acknowledge a symbolic level to thought. In contrast, ACT-R includes both symbolic and subsymbolic components. The strengths of the ACT-R theory derive from its tight integration of the symbolic component with the subsymbolic component. Its weaknesses largely derive from its failure, as yet, to adequately engage in intensive analyses of issues related to certain criteria on Newell's list.

Keywords: cognitive architecture; connectionism; hybrid systems; language; learning; symbolic systems

1. Introduction

Allen Newell, typically a cheery and optimistic man, often expressed frustration over the state of progress in cognitive science. He would point to such things as the "schools" of thought, the changes in fashion, the dominance of controversies, and the cyclical nature of theories. One of the problems he saw was that the field had become too focused on specific issues and had lost sight of the big picture needed to understand the human mind. He advocated a number of remedies for this problem. Twice, Newell (1980; 1990) offered slightly different sets of 13 criteria on the human mind, with the idea (more clearly stated in 1990) that the field would make progress if it tried to address all of these criteria. Table 1 gives the first 12 criteria from his 1980 list, which were basically restated in the 1990 list. Although the individual criteria may vary in their scope and in how compelling they are, none are trivial.

These criteria are functional constraints on the cognitive architecture. The first nine reflect things that the architecture must achieve to implement human intellectual capacity, and the last three reflect constraints on how these functions are to be achieved. As such, they do not reflect everything that one should ask of a cognitive theory. For example, it is imaginable that one could have a system that satisfied all of these criteria and still did not correspond to the human mind. Thus, foremost among the additional criteria that a cognitive theory must satisfy is that it has to correspond to the details of human cognition. In addition to behavioral adequacy, we emphasize that the theory be capable of practical applications in domains like education or therapy. Nonetheless, while the criteria on this list are not everything that one might ask of a full theory of human cognition, they certainly are enough to avoid theoretical myopia.

While Newell certainly was aware of the importance of having theories reproduce the critical nuances of particular experiments, he did express frustration that functionality did not get the attention it deserved in psychology. For instance, Newell (1992) complained about the lack of attention to this in theories of short-term memory (STM) – that it had not been shown that "with whatever limitation the particular STM theory posits, it is possible for the human to function intelligently." He asked, "why don't psychologists address it (functionality) or recognize that there might be a genuine scientific conundrum here, on which the conclusion could be that the existing models are not right?" A theory that predicts the correct serial position curve in a particular experiment, but also says that humans cannot keep track of the situation model implied by a text they are reading (Ericsson & Kintsch 1995), is simply wrong.

So, to repeat: we are not proposing that the criteria in Table 1 are the only ones by which a cognitive theory should be judged. However, such functional criteria need to be

 0140-525X/03 $12.50

given greater scientific prominence. To achieve this goal, we propose to evaluate theories by how well they do at meeting these functional criteria. We suggest calling the evaluation of a theory by this set of criteria "The Newell Test."

This target article reviews Newell's criteria and then considers how they apply to evaluating the various approaches to the study of human cognition. We focus on evaluating two approaches in detail. One is classical connectionism, as exemplified in publications like McClelland and Rumelhart (1986), Rumelhart and McClelland (1986b), and Elman et al. (1996). The other is our own ACT-R theory. To be concrete, we suggest a grading scheme and issue report cards for the two theoretical approaches.

2. Newell's criteria

When Newell first introduced these criteria in 1980, he devoted less than two pages to describing them, and he devoted no more space to them when he described them again in his 1990 book. He must have thought that the criteria were obvious, but the field of cognitive science has not found them all obvious. Therefore, we can be forgiven if we give a little more space to their consideration than did Newell. In this section, we will try to accomplish two things. The first is to make the case that each is a criterion by which all complete theories of cognition should be evaluated. The second is to try to state objective measures associated with the criteria so that their use in evaluation will not be hopelessly subjective. These measures are also summarized in Table 1. Our attempts to achieve objective measures vary in success. Perhaps others can suggest better measures.

John Anderson received his B.A. from the University of British Columbia in 1968 and his Ph.D. from Stanford University 1972. He has been at Carnegie Mellon University since 1978, where he is a professor of psychology and computer science. His current research is on three enterprises involved in the testing of various aspects of the ACT-R theory of cognitive architecture: (1) to model the acquisition of cognitive skills, particularly those involving dynamic problem solving; (2) the application of the architectures to developing intelligent tutoring systems and cognitive agents for training; and (3) research on brain imaging to identify the neural correlates of the cognitive architecture.

Christian Lebiere is a Principal Research Scientist at the Micro Analysis and Design Unit, School of Computer Science, Carnegie Mellon University. He received his B.S. in Computer Science from the University of Liege (Belgium), and his M.S. and Ph.D. from the School of Computer Science at Carnegie Mellon University. During his graduate career, he worked on the development of connectionist models, including the Cascade-Correlation neural network learning algorithm. Since 1990, he has worked on the development of the ACT-R hybrid architecture of cognition. His main research interest is cognitive architectures and their applications to psychology, artificial intelligence, human-computer interaction, decision-making, game theory, and computer-generated forces.

2.1. Flexible behavior

In his 1990 book, *Unified Theories of Cognition,* Newell restated his first criterion as "behave flexibly as a function of the environment," which makes it seem a rather vacuous criterion for human cognition. However, in 1980 he was quite clear that he meant this to be computational univer-

Table 1. *Newell's Functional Criteria for a Human Cognitive Architecture: Proposed Operationalizations and Gradings*

1. Behave as an (almost) arbitrary function of the environment
   –Is it computationally universal with failure?
   Classical Connectionism: Mixed; ACT-R: Better
2. Operate in real time
   –Given its timing assumptions, can it respond as fast as humans?
   Classical Connectionism: Worse; ACT-R: Best
3. Exhibit rational, i.e., effective adaptive behavior
   –Does the system yield functional behavior in the real world?
   Classical Connectionism: Better; ACT-R: Better
4. Use vast amounts of knowledge about the environment
   –How does the size of the knowledge base affect performance?
   Classical Connectionism: Worse; ACT-R: Mixed
5. Behave robustly in the face of error, the unexpected, and the unknown
   –Can it produce cognitive agents that successfully inhabit dynamic environments?
   Classical Connectionism: Mixed; ACT-R: Better
6. Integrate diverse knowledge
   –Is it capable of common examples of intellectual combination?
   Classical Connectionism: Worse; ACT-R: Mixed
7. Use (natural) language
   –Is it ready to take a test of language proficiency?
   Classical Connectionism: Better; ACT-R: Worse
8. Exhibit self-awareness and a sense of self
   –Can it produce functional accounts of phenomena that reflect consciousness
   Classical Connectionism: Worse; ACT-R: Worse
9. Learn from its environment
   –Can it produce the variety of human learning
   Classical Connectionism: Better; ACT-R: Better
10. Acquire capabilities through development
   –Can it account for developmental phenomena?
   Classical Connectionism: Better; ACT-R: Worse
11. Arise through evolution
   –Does the theory relate to evolutionary and comparative considerations?
   Classical Connectionism: Worst; ACT-R: Worst
12. Be realizable within the brain
   –Do the components of the theory exhaustively map onto brain processes?
   Classical Connectionism: Best; ACT-R: Worse

sality, and that it was the most important criterion. He devoted the major portion of the 1980 paper to proving that the symbol system he was describing satisfied this criterion. For Newell, the flexibility in human behavior implied computational universality. With modern fashion so emphasizing evolutionarily-prepared, specialized cognitive functions, it is worthwhile to remind ourselves that one of the most distinguishing human features is the ability to learn to perform almost arbitrary cognitive tasks to high degrees of expertise. Whether it is air-traffic control or computer programming, people are capable of performing with high facility cognitive activities that had no anticipation in human evolutionary history. Moreover, humans are the only species that show anything like this cognitive plasticity.

Newell recognized the difficulties he was creating in identifying this capability with formal notions of universal computability. For example, memory limitations prevent humans from being equivalent to Turing machines (with their infinite tapes), and their frequent slips prevent people from displaying perfect behavior. However, he recognized the true flexibility in human cognition that deserved this identification with computational universality, even as the modern computer is characterized as a Turing-equivalent device despite its physical limitations and occasional errors.

While computational universality is a fact of human cognition, it should not be seen in opposition to the idea of specialized facilities for performing various cognitive functions – even a computer can have specialized processors. Moreover, it should not be seen in opposition to the view that some things are much easier for people to learn and do than others. This has been stressed in the linguistic domain where it is argued that there are "natural languages" that are much easier to learn than nonnatural languages. However, this lesson is perhaps even clearer in the world of human artifacts, like air-traffic control systems or computer applications, where some systems are much easier to learn and to use than others. Although there are many complaints about how poorly designed some of these systems are, the artifacts that are in common use are only the tip of the iceberg with respect to unnatural systems. While humans may approach computational universality, it is only a tiny fraction of the computable functions that humans find feasible to acquire and perform.

Grading: If a theory is well specified, it should be relatively straightforward to determine whether it is computationally universal or not. As already noted, this is not to say that the theory should claim that people will find everything equally easy or that human performance will ever be error free.

2.2. Real-time performance

It is not enough for a theory of cognition to explain the great flexibility of human cognition, it must also explain how humans can do this in what Newell referred to as "real time," which means human time. As the understanding of the neural underpinnings of human cognition increases, the field faces increasing constraints on its proposals as to what can be done in a fixed period of time. Real time is a constraint on learning as well as performance. It is no good to be able to learn something in principle if it takes lifetimes to do that learning.

Grading: If a theory comes with well-specified constraints on how fast its processes can proceed, then it is relatively trivial to determine whether it can achieve real time for any specific case of human cognition. It is not possible to prove that the theory satisfies the real-time constraint for all cases of human cognition, so one must be content with looking at specific cases.

2.3. Adaptive behavior

Humans do not just perform marvelous intellectual computations. The computations that they choose to perform serve their needs. As Anderson (1991) argued, there are two levels at which one can address adaptivity. At one level, one can look at the basic processes of an architecture, such as association formation, and ask whether and how they serve a useful function. At another level, one can look at how the whole system is put together and ask whether its overall computation serves to meet human needs.

Grading: What protected the short-term memory models that Newell complained about from the conclusion that they were not adaptive was that they were not part of more completely specified systems. Consequently, one could not determine their implications beyond the laboratory experiments they addressed, where adaptivity was not an issue. However, if one has a more completely specified theory like Newell's Soar system (Newell 1990), one can explore whether the mechanism enables behavior that would be functional in the real world. Although such assessment is not trivial, it can be achieved as shown by analyses such as those exemplified in Oaksford and Chater (1998) or Gigerenzer (2000).

2.4. Vast knowledge base

One key to human adaptivity is the vast amount of knowledge that can be called on. Probably what most distinguishes human cognition from various "expert systems" is the fact that humans have the knowledge necessary to act appropriately in so many situations. However, this vast knowledge base can create problems. Not all of the knowledge is equally reliable or equally relevant. What is relevant to the current situation can rapidly become irrelevant. There may be serious issues of successfully storing all the knowledge and retrieving the relevant knowledge in reasonable time.

Grading: To assess this criterion requires determining how performance changes with the scale of the knowledge base. Again, if the theory is well specified, this criterion is subject to formal analysis. Of course, one should not expect that size will have no effect on performance – as anyone knows who has tried to learn the names of students in a class of 200.

2.5. Dynamic behavior

Living in the real world is not like solving a puzzle like the Tower of Hanoi. The world can change in ways that we do not expect and do not control. Even human efforts to control the world by acting on it can have unexpected effects. People make mistakes and have to recover. The ability to deal with a dynamic and unpredictable environment is a precondition to survival for all organisms. Given the complexity of the environments that humans have created for themselves, the need for dynamic behavior is one of the major cognitive stressors that they face. Dealing with dynamic behavior requires a theory of perception and action as well as

a theory of cognition. The work on situated cognition (e.g., Greeno 1989; Lave 1988; Suchman 1987) has emphasized how cognition arises in response to the structure of the external world. Advocates of this position sometimes argue that all there is to cognition is reaction to the external world. This is the symmetric error to the earlier view that cognition could ignore the external world (Clark 1998; 1999).

Grading: How does one create a test of how well a system deals with the "unexpected"? Certainly, the typical laboratory experiment does a poor job of putting this to the test. An appropriate test requires inserting these systems into uncontrolled environments. In this regard, a promising class of tests looks at cognitive agents built in these systems and inserted into real or synthetic environments. For example, Newell's Soar system successfully simulated pilots in an Air Force mission simulation that involved 5,000 agents including human pilots (Jones et al. 1999).

2.6. Knowledge integration

We have chosen to retitle this criterion. Newell referred to it as "Symbols and Abstractions," and his only comment on this criterion appeared in his 1990 book: "[The] [m]ind is able to use symbols and abstractions. We know that just from observing ourselves" (p. 19). He never seemed to acknowledge just how contentious this issue is, although he certainly expressed frustration (Newell 1992) that people did not "get" what he meant by a symbol. Newell did not mean external symbols like words and equations, about whose existence there can be little controversy. Rather, he was thinking about symbols like those instantiated in list-processing languages. Many of these "symbols" do not have any direct meaning, unlike the sense of symbols that one finds in philosophical discussions or computational efforts, as in Harnad (1990; 1994). Using symbols in Newell's sense, as a grading criterion, seems impossibly loaded. However, if we look to his definition of what a physical symbol does, we see a way to make this criterion fair:

> Symbols provide distal access to knowledge-bearing structures that are located physically elsewhere within the system. The requirement for distal access is a constraint on computing systems that arises from action always being physically local, coupled with only a finite amount of knowledge being encodable within a finite volume of space, coupled with the human mind's containing vast amounts of knowledge. Hence encoded knowledge must be spread out in space, whence it must be continually transported from where it is stored to where processing requires it. Symbols are the means that accomplish the required distal access. (Newell 1990, p. 427)

Symbols provide the means of bringing knowledge together to make the inferences that are most intimately tied to the notion of human intellect. Fodor (2000) refers to this kind of intellectual combination as "abduction" and is so taken by its wonder that he doubts whether standard computational theories of cognition (or any other current theoretical ideas for that matter) can possibly account for it.

In our view, in his statement of this criterion Newell confused mechanism with functionality. The functionality he is describing in the preceding passage is a capacity for intellectual combination. Therefore, to make this criterion consistent with the others (and not biased), we propose to cast it as achieving this capability. In point of fact, we think that when we understand the mechanism that achieves this capacity, it will turn out to involve symbols more or less in the sense Newell intended. (However, we do think there will be some surprises when we discover how the brain achieves these symbols.) Nonetheless, not to prejudge these matters, we simply render the sixth criterion as the capacity for intellectual combination.

Grading: To grade on this criterion we suggest judging whether the theory can produce those intellectual activities which are hallmarks of daily human capacity for intellectual combination – things like inference, induction, metaphor, and analogy. As Fodor (2000) notes, it is always possible to rig a system to produce any particular inference; the real challenge is to produce them all out of one system that is not set up to anticipate any. It is important, however, that this criterion not become a test of some romantic notion of the wonders of human cognition that actually almost never happen. There are limits to the normal capacity for intellectual combination, or else great intellectual discoveries would not be so rare. The system should to be able to reproduce the intellectual combinations that people display on a day-to-day basis.

2.7. Natural language

While most of the criteria on Newell's list could be questioned by some, it is hard to imagine anyone arguing that a complete theory of cognition need not address natural language. Newell and others have wondered about the degree to which natural language is the basis of human symbol manipulation versus the degree to which symbol manipulation is the basis for natural language. Newell took the view that language depends on symbol manipulation.

Grading: It is not obvious how to characterize the full dimensions of that functionality. As a partial but significant test, we suggest looking at those tests that society has set up as measures of language processing – something like the task of reading a passage and answering questions on it. This would involve parsing, comprehension, inference, and relating current text to past knowledge. This is not to give theories a free pass on other aspects of language processing such as partaking in a conversation, but one needs to focus on something in specifying the grading for this criterion.

2.8. Consciousness

Newell acknowledged the importance of consciousness to a full account of human cognition, although he felt compelled to remark that "it is not evident what functional role self-awareness plays in the total scheme of mind." We too have tended to regard consciousness as epiphenomenal, and it has not been directly addressed in the ACT-R theory. However, Newell is calling us to consider all the criteria and not pick and choose the ones to consider.

Grading: Cohen and Schooler (1997) have edited a volume aptly titled *Scientific Approaches to Consciousness,* which contains sections on subliminal perception, implicit learning and memory, and metacognitive processes. We suggest that the measure of a theory on this criterion is its ability to produce these phenomena in a way that explains why they are functional aspects of human cognition.

2.9. Learning

Learning seems to be another uncontroversial criterion for a theory of human cognition. A satisfactory theory of cog-

nition must account for humans' ability to acquire their competences.

Grading: It seems insufficient to grade a theory simply by asking whether the theory is capable of learning because people must be capable of many different kinds of learning. We suggest taking Squire's (1992) classification as a way of measuring whether the theory can account for the range of human learning. The major categories in Squire's classification are semantic memory, episodic memory, skills, priming, and conditioning. They may not be distinct theoretical categories, and there may be more kinds of learning, but these do represent much of the range of human learning.

2.10. Development

Development is the first of the three constraints that Newell listed for a cognitive architecture. Although in some hypothetical world one might imagine the capabilities associated with cognition emerging full blown, human cognition in the real world is constrained to unfold in an organism as it grows and responds to experience.

Grading: There is a problem in grading the developmental criterion which is like that for the language criteria – there seems no good characterization of the full dimensions of human development. In contrast to language, because human development is not a capability but rather a constraint, there are no common tests for the development constraint per se, although the world abounds with tests of how well our children are developing. In grading his own Soar theory on this criterion, Newell was left with asking whether it could account for specific cases of developmental progression (for instance, he considered how Soar might apply to the balance scale). We are unable to suggest anything better.

2.11. Evolution

Human cognitive abilities must have arisen through some evolutionary history. Some have proposed that various content-specific abilities, such as the ability to detect cheaters (Cosmides & Tooby 2000b) or certain constraints on natural language (e.g., Pinker 1994; Pinker & Bloom 1990), evolved at particular times in human evolutionary history. A variation on the evolutionary constraint is the comparative constraint. How is the architecture of human cognition different from that of other mammals? We have identified cognitive plasticity as one of the defining features of human cognition, and others have identified language as a defining feature. What is it about the human cognitive system that underlies its distinct cognitive properties?

Grading: Newell expressed some puzzlement at how the evolutionary constraint should apply. Grading the evolutionary constraint is deeply problematical because of the paucity of the data on the evolution of human cognition. In contrast to judging how adaptive human cognition is in an environment (Criterion 3), reconstruction of a history of selectional pressures seems vulnerable to becoming the construction of a just-so story (Fodor 2000; Gould & Lewontin 1979). The best we can do is ask loosely how the theory relates to evolutionary and comparative considerations.

2.12. Brain

The last constraint collapses two similar criteria in Newell (1980) and corresponds to one of the criteria in Newell (1990). Newell took seriously the idea of the neural implementation of cognition. The timing of his Soar system was determined by his understanding of how it might be neurally implemented. The last decade has seen a major increase in the degree to which data about the functioning of specific brain areas are used to constrain theories of cognition.

Grading: Establishing that a theory is adequate here seems to require both an enumeration and a proof. The enumeration would be a mapping of the components of the cognitive architecture onto brain structures, and the proof would be that the computation of the brain structures match the computation of the assigned components of the architecture. There is possibly an exhaustive requirement as well – that no brain structure is left unaccounted for. Unfortunately, knowledge of brain function has not advanced to the point where one can fully implement either the enumeration or the proof of a computational match. However, there is enough knowledge to partially implement such a test, and even as a partial test, it is quite demanding.

2.13. Conclusions

It might seem reckless to open any theory to an evaluation on such a broad set of criteria as those in Table 1. However, if one is going to propose a cognitive architecture, it is impossible to avoid such an evaluation as Newell (1992) discovered with respect to Soar. As Vere (1992) described it, because a cognitive architecture aspires to give an integrated account of cognition, it will be subjected to the "attack of the killer bees" – each subfield to which the architecture is applied is "resolutely defended against intruders with improper pheromones." Vere proposed creating a "Cognitive Decathlon"

> to create a sociological environment in which work on integrated cognitive systems can prosper. Systems entering the Cognitive Decathlon are judged, perhaps figuratively, based on a cumulative score of their performance in each cognitive "event." The contestants do not have to beat all of the narrower systems in their one specialty event, but compete against other well-rounded cognitive systems. (Vere 1992, p. 460)

This target article could be viewed as a proposal for the events in the decathlon and an initial calibration of the scoring for the events by providing an evaluation of two current theories, classical connectionism and ACT-R.

While classical connectionism and ACT-R offer some interesting contrasts when graded by Newell's criteria, both of these two theories are ones that have done rather well when measured by the traditional standard in psychology of correspondence to the data of particular laboratory experiments. Thus, we are not bringing to this grading what are sometimes called *artificial intelligence* theories. It is not as if we were testing "Deep Blue" as a theory of human chess, but it is as if we were asking of a theory of human chess that it be capable of playing chess – at least in principle, if not in practice.

3. Classical connectionism

Classical connectionism is the cognitively and computationally modern heir to behaviorism. Both behaviorism and connectionism have been very explicit about what they accept and what they reject. Both focus heavily on learning

and emphasize how behavior (or cognition) arises as an adaptive response to the structure of experience (Criteria 3 and 9 in Newell's list). Both reject any abstractions (Newell's original Criterion 6, which we have revamped for evaluation) except as a matter of verbal behavior (Criterion 8). Being cognitively modern, connectionism, however, is quite comfortable in addressing issues of consciousness (Criterion 8), whereas behaviorism often explicitly rejected consciousness. The most devastating criticisms of behaviorism focused on its computational adequacy, and it is here that the distinction between connectionism and behaviorism is clearest. Modern connectionism established that it did not have the inadequacies that had been shown for the earlier Perceptrons (Minsky & Papert 1969). Connectionists developed a system that can be shown to be computationally equivalent to a Turing machine (Hartley 2000; Hartley & Szu 1987; Hornik et al. 1989; Siegelman & Sontag 1992) and endowed it with learning algorithms that could be shown to be universal function approximaters (Clark 1998; 1999).

However, as history would have it, connectionism did not replace behaviorism. Rather, there was an intervening era in which an abstract information-processing conception of mind dominated. This manifested itself perhaps most strongly in the linguistic ideas surrounding Chomsky (e.g., 1965) and the information-processing models surrounding Newell and Simon (e.g., 1972). These were two rather different paradigms, with the Chomskian approach emphasizing innate knowledge only indirectly affecting behavior, and the Newell and Simon approach emphasizing the mental steps directly underlying the performance of a cognitive task. However, both approaches deemphasized learning (Criterion 9) and emphasized cognitive abstractions (Original Criterion 6). Thus, when modern connectionism arose, the targets of its criticisms were the *symbols* and *rules* of these theories. It chose to focus largely on linguistic tasks emphasized by the Chomskian approach and was relatively silent on the problem-solving tasks emphasized by the Newell and Simon approach. Connectionism effectively challenged three of the most prized claims of the Chomskian approach – that linguistic overgeneralizations were evidence for abstract rules (Brown 1973), that initial syntactic parsing was performed by an encapsulated syntactic parser (Fodor 1983), and that it was impossible to acquire language without the help of an innate language-acquisition device (Chomsky 1965). We will briefly review each of these points, but at the outset we want to emphasize that these connectionist demonstrations were significant because they established that a theory without language-specific features had functionalities which some claimed it could not have. Thus, the issues were very much a matter of functionality in the spirit of the Newell test.

Rumelhart and McClelland's (1986b) past-tense model has become one of the most famous of the connectionist models of language processing. They showed that by learning associations between the phonological representations of stems and past tense, it was possible to produce a model that made overgeneralizations without building any rules into it. This attracted a great many critiques, and, while the fundamental demonstration of generalization without rules stands, it is acknowledged by all to be seriously flawed as a model of the process of past-tense generation by children. Many more recent and more adequate connectionist models (some reviewed in Elman et al. 1996) have been proposed, and many of these have tried to use the backpropogation learning algorithm.

While early research suggested that syntax was in some way separate from general knowledge and experience (Ferreira & Clifton 1986), further research has suggested that syntax is quite penetrable by all sorts of semantic considerations and in particular the statistics of various constructions. Models like those of MacDonald et al. (1994) are quite successful in predicting the parses of ambiguous sentences. There is also ample evidence now for syntactic priming (e.g., Bock 1986; Bock & Griffin 2000) – that people tend to use the syntactic constructions they have recently heard. There are also now sociolinguistic data (reviewed in Matessa 2001) showing that the social reinforcement contingencies shape the constructions that one will use. Statistical approaches to natural-language processing have been quite successful (Collins 1999; Magerman 1995). While these approaches are only sometimes connectionist models, they establish that the statistics of language can be valuable in untangling the meaning of language.

While one might imagine these statistical demonstrations being shrugged off as mere performance factors, the more fundamental challenges have concerned whether the syntax of natural language actually is beyond the power of connectionist networks to learn. "Proofs" of the inadequacy of behaviorism have concerned their inability to handle the computational complexity of the syntax of natural language (e.g., Bever et al. 1968). Elman (1995) used a recurrent network to predict plausible continuations for sentence fragments like *boys who chase dogs see girls* that contain multiple embeddings. This was achieved by essentially having hidden units that encoded states reflecting the past words in the sentence.

The preceding discussion has focused on connectionism's account of natural language, because that is where the issue of the capability of connectionist accounts has received the most attention. However, connectionist approaches have their most natural applications to tasks that are more directly a matter of perceptual classification or continuous tuning of motor output. Some of the most successful connectionist models have involved things like letter recognition (McClelland & Rumelhart 1981). Pattern classification and motor tuning underlie some of the more successful "performance" applications of connectionism including NETtalk (Sejnowski & Rosenberg 1987), which converts orthographic representation of words into a code suitable for use with a speech synthesizer; TD-Gammon (Tesauro 2002), a world-champion backgammon program; and ALVINN (Autonomous Land Vehicle In a Neural Network) (Pomerleau 1991), which was able to drive a vehicle on real roads.

So far we have used the term *connectionism* loosely, and it is used in the field to refer to a wide variety of often incompatible theoretical perspectives. Nonetheless, there is a consistency in the connectionist systems behind the successes just reviewed. To provide a roughly coherent framework for evaluation, we will focus on what has been called *classical connectionism.* Classical connectionism is the class of neural network models that satisfy the following requirements: feedforward or recurrent network topology, simple unit activation functions such as sigmoid or radial basis functions, and local weight-tuning rules such as backpropagation or Boltzmann learning algorithms. This defini-

tion reflects both the core and the bulk of existing neural network models while presenting a coherent computational specification. It is a restriction with consequence. For instance, the proofs of Turing equivalence include assumptions not in the spirit of classical connectionism and often involving nonstandard constructs.

4. ACT-R

4.1. ACT-R's history of development

While ACT-R is a theory of cognition rather than a framework of allied efforts like connectionism, it has a family-resemblance aspect too, in that it is just the current manifestation of a sequence of theories stretching back to Anderson (1976), when we first proposed how a subsymbolic activation-based memory could interact with a symbolic system of production rules. The early years of that project were concerned with developing a neurally plausible theory of the activation processes and an adequate theory of production rule learning, resulting in the ACT* theory (Anderson 1983). The next ten years saw numerous applications of the theory, a development of a technology for effective computer simulations, and an understanding of how the subsymbolic level served the adaptive function of tuning the system to the statistical structure of the environment (Anderson 1990). This resulted in the ACT-R version of the system (Anderson 1993), where the "R" denotes rational analysis.

Since the publication of ACT-R in 1993, a community of researchers has evolved around the theory. One major impact of this community has been to help prepare ACT-R to take the Newell Test by applying it to a broad range of issues. ACT had traditionally been a theory of "higher-level" cognition and largely ignored perception and action. However, as members of the ACT-R research community became increasingly concerned with timing and dynamic behavior (Newell's Criteria 2 and 5), it was necessary to address attentional issues about how the perceptual and motor systems interact with the cognitive system. This has led to the development of ACT-R/PM (PM for perceptual-motor) (Byrne & Anderson 1998), based in considerable part on the perceptual-motor components of EPIC (Meyer & Kieras 1997). This target article focuses on ACT-R 5.0, which is an integration of the ACT-R 4.0 described in Anderson and Lebiere (1998) and ACT-R/PM.

4.2. General description of ACT-R

Since it is a reasonable assumption that ACT-R is less well known than classical connectionism, we will give it a fuller description, although the reader should refer to Anderson and Lebiere (1998) for more formal specifications and the basic equations. Figure 1 displays the current architecture of ACT-R. The flow of cognition in the system is in response to the current goal, currently active information from declarative memory, information attended to in perceptual modules (vision and audition are implemented), and the current state of motor modules (hand and speech are implemented). These components (goal, declarative memory, perceptual, and motor modules) hold the information that the productions can access in *buffers,* and these buffers serve much the same function as the subsystems of Baddeley's (1986) working-memory theory. In response to the current state of these buffers, a production is selected and executed. The central box in Figure 1 reflects the processes that determine which production to fire. There are two distinct subprocesses – pattern matching to decide which productions are applicable, and conflict resolution to select among these applicable productions. While all productions are compared in parallel, a single production is selected to fire. The selected production can cause changes in the current goal, make a retrieval request of declarative memory, shift attention, or call for new motor actions. Unlike EPIC, ACT-R is a serial-bottleneck theory of cognition (Pashler 1998) in which parallel cognitive, perceptual, and motor modules must interact through a serial process of production execution.

The architecture in Figure 1 is an abstraction from the neural level, but nonetheless it is possible to give tentative neural correlates. The motor and perceptual modules correspond to associated cortical areas; the current goal, to frontal areas; and declarative memory, to posterior cortical and hippocampal areas. There is evidence (Wise et al. 1996) that the striatum receives activation from the full cortex and recognizes patterns of cortical activation. These recognized patterns are gated by other structures in the basal ganglia (particularly the internal segment of the globus pallidus and the substantia nigra pars reticulata) (Frank et al. 2000) and the frontal cortex to select an appropriate action. Thus, one might associate the striatum with the pattern-recognition component of the production selection and the basal ganglia structures and the frontal cortex with the conflict resolution.

ACT-R is a hybrid architecture in the sense that it has both symbolic and subsymbolic aspects. The symbolic aspects involve declarative chunks and procedural production rules. The declarative chunks are the knowledge-representation units that reside in declarative memory, and the production rules are responsible for the control of cognition. Access to these symbolic structures is determined by a subsymbolic level of neural-like activation quantities. Part of the insight of the rational analysis is that the declarative and procedural structures, by their nature, need to be guided by two different quantities. Access to declarative chunks is controlled by an activation quantity that reflects the probability that the chunk will need to be retrieved. In the case of production rules, choice among competing rules is controlled by their utilities, which are estimates of the rule's probability of success and cost in leading to the goal. These estimates are based on the past reinforcement history of the production rule.

The activation of a chunk is critical in determining its retrieval from declarative memory. A number of factors determine the level of activation of a chunk in declarative memory:

1. The recency and frequency of usage of a chunk will determine its base-level activation. This base-level activation represents the probability (actually, the log odds) that a chunk is needed, and the estimates provided for by ACT-R's learning equations represent the probabilities in the environment (see Anderson 1993, Ch. 4, for examples).

2. Added to this base-level activation is an associative component that reflects the priming that the chunk might receive from elements currently in the focus of attention. The associations among chunks are learned on the basis of past patterns of retrieval according to a Bayesian framework.

ED:
Quality of art okay through out?
Please advise.
–Comp.

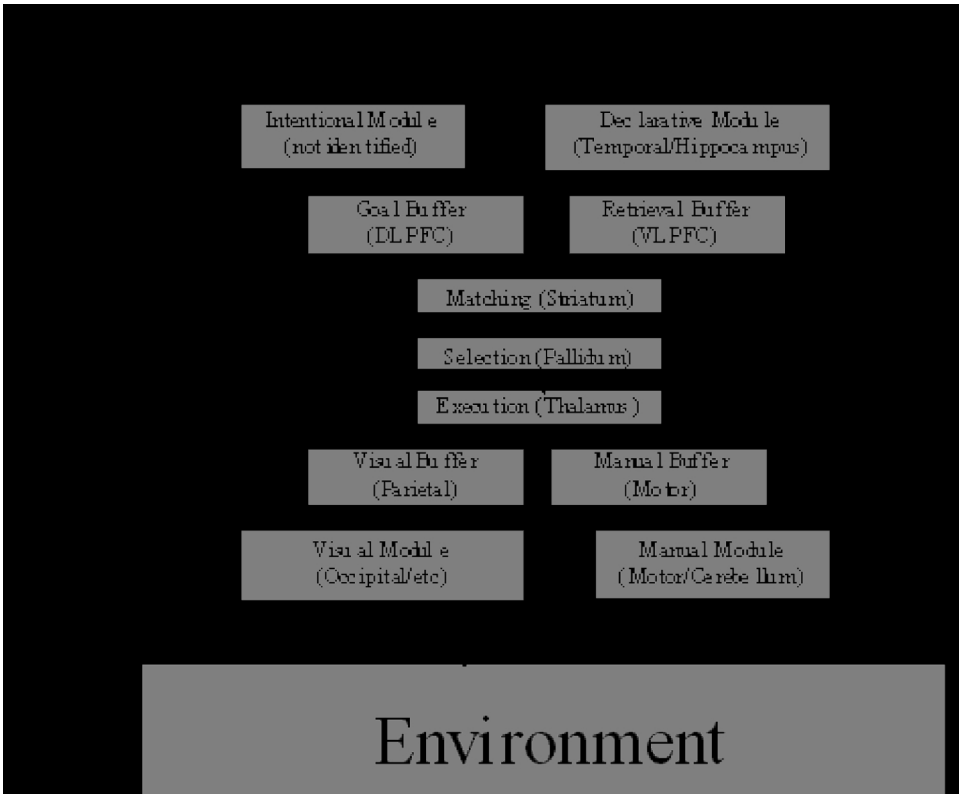

Figure 1. ACT-R Architecture

3. The activation controlled by factors 1 and 2 is modulated by the degree to which the chunk matches current retrieval specifications. Thus, for example, a chunk that encodes a similar situation to the current one will receive some activation. This partially matching component in ACT-R allows it to produce the soft, graceful behavior characteristic of human cognition. Similarities among chunks serve a similar purpose to distributed representations in connectionist networks.

4. The activation quantities are fundamentally noisy, so there is some variability in which chunk is most active, producing a stochasticity in behavior.

The activation of a chunk determines the time to retrieve it. Also, when multiple chunks can be retrieved, the most active one is selected. This principle, combined with variability in activation, produces predictions for the probability of recall according to the softmax Boltzmann distribution (Ackley et al. 1985; Hinton & Sejnowski 1986). These latency and probability functions in conjunction with the activation processes have led to a wide variety of successful models of verbal learning (e.g., Anderson et al. 1998a; Anderson & Reder 1999a).

Each production rule has a real-valued utility that is calculated from estimates of the cost and probability of reaching the goal if that production rule is chosen. ACT-R's learning mechanisms constantly update these estimates based on experience. If multiple production rules are applicable to a certain goal, the production rule with the highest utility is selected. This selection process is noisy, so the production with the highest utility has the greatest probability of being selected, but other productions get opportunities as well. This may produce errors or suboptimal behavior, but also allows the system to explore knowledge and strategies that are still evolving. The ACT-R theory of utility learning has been tested in numerous studies of strategy selection and strategy learning (e.g., Lovett 1998).

In addition to the learning mechanisms that update activation and expected outcome, ACT-R can also learn new chunks and production rules. New chunks are learned automatically: Each time a goal is completed or a new percept is encountered, it is added to declarative memory. New production rules are learned by combining existing production rules. The circumstance for learning a new production rule is that two rules fire one after another, with the first rule retrieving a chunk from memory. A new production rule is formed that combines the two into a macro-rule but eliminates the retrieval. Therefore, everything in an ACT-R model (chunks, productions, activations, and utilities) is learnable.

The symbolic level is not merely a poor approximation to the subsymbolic level as claimed by Rumelhart and McClelland (1986b) and Smolensky (1988); rather, it provides the essential structure of cognition. It might seem strange that neural computation should just so happen to satisfy the well-formedness constraints required to correspond to the symbolic level of a system like ACT-R. This would indeed be miraculous if the brain started out as an unstructured net that had to organize itself just in response to experience. However, as illustrated in the tentative brain correspondences for ACT-R components and in the following description of ACT-RN, the symbolic structure emerges out of the structure of the brain. For example, just as the two eyes converge in adjacent columns in the visual cortex to enable stereopsis, a similar convergence of information

(perhaps in the basal ganglia) would permit the condition of a production rule to be learned.

4.3. ACT-RN

ACT-R is not in opposition to classical connectionism except in connectionism's rejection of a symbolic level. Although strategically ACT-R models tend to be developed at a larger grain size than connectionist models, we do think these models could be realized by the kinds of computation proposed by connectionism. Lebiere and Anderson (1993) instantiated this belief in a system called ACT-RN that attempted to implement ACT-R using standard connectionist concepts. We will briefly review ACT-RN here because it shows how production system constructs can be compatible with neural computation.

ACT-R consists of two key memories – a declarative memory and a procedural memory. Figure 2 illustrates how ACT-RN implements declarative chunks. The system has separate memories for each different type of chunk – for example, addition facts are represented by one type memory, whereas integers are represented by a separate type memory. Each type memory is implemented as a special version of Hopfield nets (Hopfield 1982). A chunk in ACT-R consists of a unique identifier called the header, together with a number of slots, each containing a value, which can be the identifier of another chunk. Each slot, as well as the chunk identifier itself, is represented by a separate pool of units, thereby achieving a distributed representation. A chunk is represented in the pattern of connections between these pools of units. Instead of having complete connectivity among all pools, the slots are only connected to the header and vice versa. Retrieval involves activating patterns in some of the pools and trying to fill in the remaining patterns corresponding to the retrieved chunk. If some slot patterns are activated, they are mapped to the header units to retrieve the chunk identifier that most closely matches these contents (path 1 in Fig. 2). Then, the header is mapped back to the slots to fill the remaining values (path 5). If the header pattern is specified, then the step corresponding to path 1 is omitted.

To ensure optimal retrieval, it is necessary to "clean" the header. This can be achieved in a number of ways. One is to implement the header itself as an associative memory. We chose instead to connect the header to a pool of units called the chunk layer in which each unit represented a chunk, achieving a localist representation (path 2). The header units are connected to all the units in the chunk layer. The pattern of weights leading to a particular localist unit in the chunk layer corresponds to the representation of that chunk in the header. By assembling these chunk-layer units in a winner-take-all network (path 3), the chunk with the representation closest to the retrieved header ultimately wins. That chunk's representation is then reinforced in the header (path 4). A similar mechanism is described in Dolan and Smolensky (1989). The initial activation level of the winning chunk is related to the number of iterations in the chunk-layer needed to find a clear winner. This maps onto retrieval time in ACT-R, as derived in Anderson and Lebiere (1998, Ch. 3 Appendix).

ACT-RN provides a different view of the symbolic side of ACT-R. As is apparent in Figure 2, a chunk is nothing more or less than a pattern of connections between the chunk identifier and its slots.

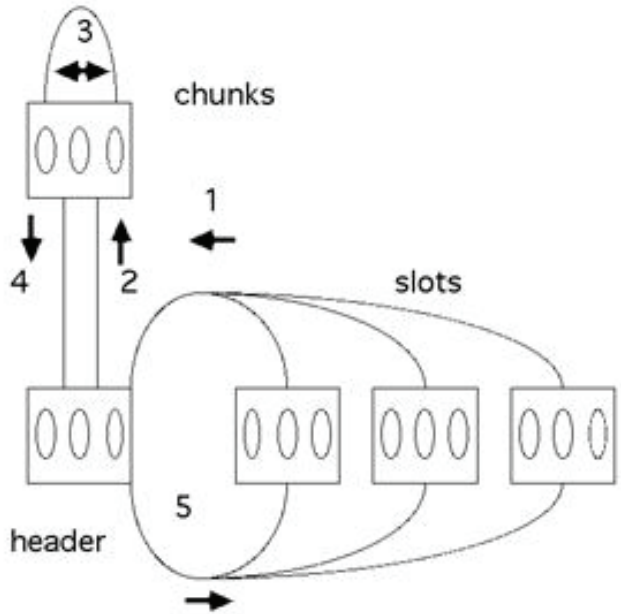

Figure 2. Declarative Memory in ACT-RN

ACT-R is a goal-oriented system. To implement this, ACT-RN has a central memory (which probably should be identified with dorsolateral prefrontal cortex), which at all times contains the current goal chunk (Fig. 3) with connections to and from each type memory. Central memory consists of pools of units, where each pool encodes a slot value of the goal. There was an optional goal stack (represented in Fig. 3), but we do not use a goal stack in ACT-R anymore. Productions in ACT-RN retrieve information from a type memory and deposit it in central memory. Such a production might retrieve from an addition memory the sum of two digits held in central memory. For example, given the goal of adding 2 and 3, a production would copy to the addition-fact memory the chunks 2 and 3 in the proper slots by enabling (gating) the proper connections between central memory and that type memory, let the memory retrieve the sum 5, and then transfer that chunk to the appropriate goal slot.

To provide control over production firing, ACT-RN needs a way to decide not only what is to be transferred where but also under what conditions. In ACT-RN, that task is achieved by gating units (which might be identified with gating functions associated with basal ganglia). Each gating unit implements a particular production and has incoming connections from central memory that reflect the goal constraints on the left-hand side of that production. For example, suppose goal slot S is required to have as value chunk C in production P. To implement this, the connections between S and the gating unit for P will be the representation for C, with an appropriate threshold. At each production cycle, all the gating units are activated by the current state of central memory, and a winner-take-all competition selects the production to fire.

Note that production rules in ACT-RN are basically rules for enabling pathways back and forth between a central goal memory and the various declarative memory modules. Thus, production rules are not really structures that are stored in particular locations but are rather specifications of information transfer. ACT-RN also offers an interesting

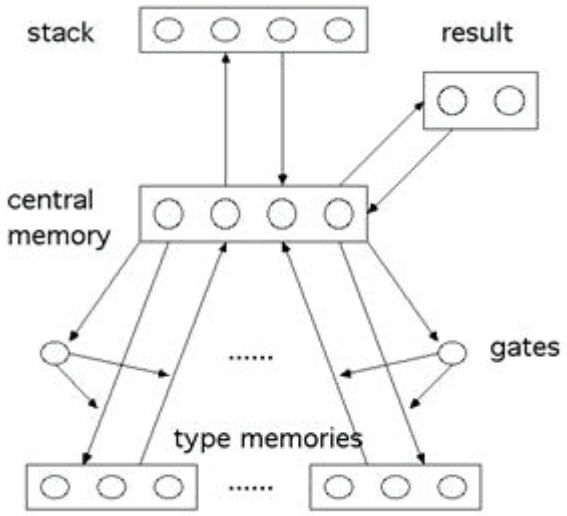

Figure 3. Procedural Memory in ACT-RN

perspective on the variables (see Marcus [2001] for a discussion of variables in connectionist models) that appear in production rules and their bindings. The effect of such bindings is basically to copy values from the goal to declarative memory and back again. This is achieved in ACT-RN without having any explicit variables or an explicit process of variable binding. Thus, while the computational power that is represented by variables is critical, one can have this without the commitment to explicit variables or a process of variable binding.

4.4. Learning past tense in ACT

Recently, Taatgen (2001; Taatgen & Anderson 2002) has developed a successful ACT-R model for learning the past-tense in English, which provides an interesting comparison point with the connectionist models. Unlike many past-tense models, it learns based on the actual frequency of words in natural language, learns without feedback, and makes the appropriate set of generalizations. While the reader should go to the original papers for details, we will briefly describe this model because the past tense has been critical in the connectionist-symbolic debate. It also serves to illustrate all of the ACT-R learning mechanisms working at once.

The model posits that children initially approach the task of past-tense generation with two strategies. Given a particular word like "give," they can either try to retrieve the past tense for that word or try to retrieve some other example of a past tense (e.g., "live"–"lived") and try to apply this by analogy to the current case. Eventually, through the production-rule learning mechanisms in ACT-R, the analogy process will be converted into a production rule that generatively applies the past-tense rule. Once the past-tense rule is learned, the generation of past tenses will largely be determined by a competition between the general rule and retrieval of specific cases. Thus, ACT-R is basically a dual-route model of past-tense generation, where both routes are implemented by production rules. The rule-based approach depends on general production rules, whereas the exemplar approach depends on the retrieval of declarative chunks by production rules that implement an instance-based strategy. This choice between retrieval and rule-based computation is a general theme in ACT-R models and is closely related to Logan's model of skill acquisition (Logan 1988). It has been used in a model of cognitive arithmetic (Lebiere 1998) and in models for a number of laboratory tasks (Anderson & Betz 2001; Lerch et al. 1999; Wallach & Lebiere, in press).

The general past-tense rule, once discovered by analogy, gradually enters the competition as the system learns that this new rule is widely applicable. This gradual entry, which depends on ACT-R's subsymbolic utility-learning mechanisms, is responsible for the onset of overgeneralization. Although this onset is not all-or-none in either the model or the data, it is a relatively rapid transition in both model and data and corresponds to the first turn in the U-shaped function. However, as this is happening, the ACT-R model is encountering and strengthening the declarative representations of exceptions to the general rule. Retrieval of the exceptions comes to counteract the overgeneralizations. Retrieval of exceptions is preferred because they tend to be shorter and phonetically more regular (Burzio 1999) than regular past tenses. Growth in this retrieval process corresponds to the second turn in the U-shaped function and is much more gradual – again, both in model and data. Note that the Taatgen model, unlike many other past-tense models, does not make artificial assumptions about frequency of exposure but learns, given a presentation schedule of words (both from the environment and its own generations) like that actually encountered by children. Its ability to reproduce the relatively rapid onset of overgeneralization and slow extinction depends critically on both its symbolic and subsymbolic learning mechanisms. Symbolically, it is learning general production rules and declarative representations of exceptions. Subsymbolically, it is learning the utility of these production rules and the activation strengths of the declarative chunks.

Beyond just reproducing the U-shaped function, the ACT-R model explains why exceptions should be high-frequency words. There are two aspects to this explanation. First, only high-frequency words develop enough base-level activation to be retrieved. Indeed, the theory predicts how frequent a word has to be in order to maintain an exception. Less obviously, the model explains why so many high-frequency words actually end up as exceptions. This is because the greater efficiency of the irregular form promotes its adoption according to the utility calculations of ACT-R. In another model that basically invents its own past-tense grammar without input from the environment, Taatgen showed that it develops one or more past-tense rules for low-frequency words but tends to adopt more efficient irregular forms for high-frequency words. In the ACT-R economy the greater phonological efficiency of the irregular form justifies its maintenance in declarative memory if it is of sufficiently high frequency.

Note that the model receives no feedback on the past tenses it generates, unlike most other models but in apparent correspondence with the facts about child language learning. However, it receives input from the environment in the form of the past tenses it hears, and this input influences the base-level activation of the past-tense forms in declarative memory. The model also uses its own past-tense generations as input to declarative memory and can learn its own errors (a phenomenon also noted in cognitive arithmetic – Siegler 1988). The amount of overgeneralization

displayed by the model is sensitive to the ratio of input it receives from the environment to its own past-tense generations.

While the model fully depends on the existence of rules and symbols, it also critically depends on the subsymbolic properties of ACT-R to produce the graded effects. This eclectic position enables the model to achieve a number of other features not achieved by many other models:

1. It does not have to rely on artificial assumptions about presentation frequency.
2. It does not need corrective feedback on its own generations.
3. It explains why irregular forms tend to be of high frequency and why high-frequency words tend to be irregular.
4. It correctly predicts that novel words will receive regular past tenses.
5. It predicts the gradual onset of overgeneralization and its much more gradual extinction.

4.5. What ACT-R doesn't do

Sometimes the suspicion is stated that ACT-R is a general computational system that can be programmed to do anything. To address this issue, we would like to specify four senses in which the system falls short of that.

First of all, ACT-R is also a system with strong limitations. Because of prior constraints on its timing, there are limits on how fast it can process material. The perceptual and motor components of the system take fixed time – for instance, it would be impossible for the system to press a button in response to a visual stimulus in less than 100 msec. At a cognitive level, it has limits on the rate of production selection and retrieval of declarative memory. This has been a major challenge in our theories of natural-language processing (Anderson et al. 2001; Budiu & Anderson, submitted), and it remains an open issue whether the general architecture can process language at the speed with which humans process it. The serial bottleneck in production selection causes all sorts of limitations – for example, the theory cannot perform mental addition and multiplication together as fast as it can perform either singly (Byrne & Anderson 2001). Limitations in memory mean that the system cannot remember a long list of digits presented at a 1-second rate (at least without having acquired a large repertoire of mnemonic skills (Chase & Ericsson 1982)). The limitations actually are successes of ACT-R as a theory of human cognition, since humans appear to display these limitations (with the issue about language open). However, their existence means that we cannot just "program" arbitrary models in ACT-R.

Second, there are also numerous mechanisms of cognition not yet incorporated into ACT-R, although there may be no in-principle reason why they cannot be incorporated. For example, ACT-R lacks any theory of the processes of speech perception or speech production. This is not without consequence for the claims of the theory. For instance, the just reviewed past-tense model made critical claims about the phonological costs of various past-tense inflections but these were just assertions not derived from the model. The absence of a phonological component makes it difficult to extend the model to making predictions about other inflectional constructions. Among other domains for which ACT-R seems to be lacking adequate mechanisms are perceptual recognition, mental imagery, emotion, and motivation. We do not think these absences reflect anything fundamentally incompatible between what the theory claims and what people can do, but that possibility always exists until it is shown how such mechanisms could be added in a consistent way to ACT-R.

Third, there are also numerous domains of great interest to cognitive science that have yet to be addressed by ACT-R. Many of these are concerned with perceptual recognition where the mechanisms of the theory are weak or lacking (the perceptual modules in ACT-R are really theories of perceptual attention) but others just reflect the failure of ACT-R researchers to take up the topic. For example, there are no ACT-R models of deductive reasoning tasks. Also, within domains that ACT-R has addressed, there are important phenomena left unaddressed. For example, although there is an ACT-R model of recognition memory (Anderson et al. 1998a), it has not addressed the *remember-know* distinction (Reder et al. 2000) or data on latency distributions (Ratcliff et al. 1999). It is not clear whether these open issues reflect simply things that ACT-R researchers have not addressed, or whether they are fundamental failings of the theory. For example, Reder (personal communication) has argued that the failure to address the *remember-know* distinction reflects the fact that ACT-R cannot deal with a whole class of metacognitive judgments because it does not have conscious access to its own subsymbolic quantities.

Finally, there is a set of implementation issues rife among researchers in the ACT-R community. We do not want to belabor them, as they have an esoteric flavor, but just to acknowledge that such things exist, we name a few (and ACT-R researchers will recognize them): avoiding repeatedly retrieving a chunk because retrievals strengthen the chunk, creating new chunk types, producing a latency function that adequately reflects competition among similar memories, and setting the temporal bounds for utility learning.

5. Grading classical connectionism and ACT-R according to the Newell Test

Having described Newell's criteria and the two theories, it is now time to apply these criteria to grading the theories. Regrettably, we were not able to state the Newell criteria in such a way that their satisfaction would be entirely a matter of objective fact. The problems are perhaps most grievous in the cases of the developmental and evolutionary criteria, where it is hard to name anything that would be a satisfactory measure, and one is largely left with subjective judgment. Even with hard criteria like computational universality, there is uncertainty about what approaches are really in keeping with the spirit of an architecture and how complete an answer particular solutions yield.

We had originally proposed a letter-grading scheme for the criteria that we applied to ACT-R. However, we were persuaded in the review process to apply the criteria to classical connectionism by the argument that the criteria became more meaningful when one sees how they apply to two rather different theories. It did not make sense to be competitively grading one's own theory alongside another one, and therefore we decided to change the grading into a rough within-theory rank ordering of how well that theory did on those criteria. That is, we will be rating how well that theory has done on a particular criterion, relative to how well it has done on other criteria (not relative to the other theory). Therefore, we will be using the following grading:

Best: The criteria on which that theory has done the best
Better: Four criteria on which that theory has done better
Mixed: Two criteria on which that theory has the most mixed record
Worse: Four criteria on which that theory has done worse
Worst: The criteria on which that theory has done the worst

This is actually more in keeping with our intentions for the Newell Test than the original letter grading because it focuses on directions for improving a given theory rather than declaring a winner. Of course, the reader is free to apply an absolute grading scheme to these two theories or any other.

5.1. Flexible behavior

Grading: *Connectionism: Mixed*
ACT-R: Better

To do well on this criterion requires that the theory achieve an interesting balance: It must be capable of computing any function, but have breakdowns in doing so, and find some functions easier to compute than others. It has been shown possible to implement a Turing machine in connectionism, but not in the spirit of classical connectionism. Breakdowns in the execution of a sequence of actions would be quite common (Botvinick & Plaut, submitted). There is a balance between capability and limitation in classical connectionism, but we and some others (e.g., Marcus 2001) believe that this is an uneven balance in favor of limitations. It is not clear that complex, sequentially organized, hierarchical behavior can be adequately produced in classical connectionistic systems, and there seems to a paucity of demonstrations. Indeed, a number of the high-performance connectionist systems have been explicitly augmented with handcrafted representations (Tesauro 2002) and symbolic capabilities (Pomerleau et al. 1991). Moreover, the connectionist models that do exist tend to be single-task models. However, the essence of computational universality is that one system can give rise to an unbounded set of very different behaviors.

ACT-R does well on this criterion in no small part because it was exactly this criterion that has most driven the design of this model. ACT-R, except for its subsymbolic limitations, is Turing equivalent, as are most production systems (proof for an early version of ACT appears in Anderson 1976). However, because of variability and memory errors, ACT-R frequently deviates from the prescribed course of its symbolic processing. This shows up, for example, in ACT-R models for the Tower of Hanoi (Anderson & Douglass 2001; Altmann & Trafton 2002), where it is shown that memory failures produce deviations from well-learned algorithms at just those points where a number of goals have to be recalled. (These are also the points where humans produce such deviations.) Nonetheless, ACT-R has also been shown to be capable of producing complex sequential behavior such as operation of an air-traffic control system (Taatgen 2002). The functions that it finds easy to compute are those with enough support from the environment to enable behavior to be corrected when it deviates from the main course.

5.2. Real-time performance

Grading: *Connectionism: Worse*
ACT-R: Best

Connectionist processing often has a poorly defined (or just poor) relationship to the demands of real-time processing. The mapping of processing to reaction time is inconsistent and often quite arbitrary; for example, some relatively arbitrary function of the unit activation is often proposed (e.g., Rumelhart & McClelland 1982). For feedforward models that depend on synchronous updates across the various levels of units, it is fundamentally inconsistent to assume that the time for a unit to reach full activation is a function of that activation. The natural factor would seem to be the number of cycles, but even when this is adopted, it is often arbitrarily scaled (e.g., a linear function of number of cycles with a negative intercept; see Plaut & Booth 2000). Another problem is that connectionist systems typically only model a single step of the full task (the main mapping) and do not account for the timing effects produced by other aspects of the task such as perceptual or motor. Finally, with respect to learning time, the number of epochs that it takes, to acquire an ability, maps poorly to the learning of humans (Schneider & Oliver 1991). This last fact is one of the major motivations for the development of hybrid models.

One of the great strengths of ACT-R is that every processing step comes with a commitment to the time it will take. It is not possible to produce an ACT-R model without timing predictions. Of course, it is no small matter that ACT-R not only makes predictions about processing time, but that these happen to be correct over a wide range of phenomena. As knowledge accumulates in the ACT-R community, these timing predictions are becoming a priori predictions. As one sign of this, in recent classes that we have taught, undergraduates at CMU were producing models that predicted absolute, as well as relative times, with no parameter estimation. In addition to performance time, ACT-R makes predictions about learning time. In a number of simulations, ACT-R was able to learn competences in human time (i.e., given as many training experiences as humans). This includes cognitive arithmetic (Lebiere 1998), past-tense formations (Taatgen & Anderson 2002), and backgammon (Sanner et al. 2000). ACT-R's treatment of time provides one answer to Roberts and Pashler's (2000) critique of model-fitting efforts. These researchers view it as so easy to fit a model to data that it is at best an uninformative activity. Their claim that it is easy or uninformative can be challenged on many grounds, but the ACT-R effort highlights the fact that one need not be fitting one experiment or paradigm in isolation.

5.3. Adaptive behavior

Grading: *Connectionism: Better*
ACT-R: Better

The positions of connectionism and ACT-R on this criterion are quite similar. Both have made efforts, often Bayesian in character (McClelland & Chappell1998), to have their underlying learning rules tune the system to the statistical structure of the environment. This is quite central to ACT-R because its subsymbolic level derives from the earlier rational analysis of cognition (Anderson 1990). However, adaptivity is not a direct function of these subsymbolic equations but rather is a function of the overall behavior of the system. ACT-R lacks an overall analysis of adaptivity, including an analysis of how the goals selected by ACT-R are biologically significant. An overall analysis is similarly lacking in classical connectionism.

The reader will recall that Newell raised the issue of the adaptivity of limitations like short-term memory. In ACT-

R, short-term memory effects are produced by decay of base-level activations. ACT-R's use of base-level activations delivers a computational embodiment of the rational analysis of Anderson (1990), which claimed that such loss of information with time reflected an adaptive response to the statistics of the environment where information loses its relevance with time. Thus, ACT-R has implemented this rational analysis in its activation computations and has shown that the resulting system satisfies Newell's requirement that it be functional.

5.4. Vast knowledge base

Grading: *Connectionism: Worse*
ACT-R: Mixed

Just because a system works well on small problems, one has no guarantee that it will do so on large problems. There have been numerous analyses of the scaling properties of neural networks. In models like NETtalk, it has shown how a great deal of knowledge can be captured in the connections among units, but that this depends on a similarity in the input-output mappings. One of the notorious problems with connectionism is the phenomenon of catastrophic interference whereby new knowledge overwrites old knowledge (McCloskey & Cohen 1989; Ratcliff 1990). Connectionists are much aware of this problem and numerous research efforts (e.g., McClelland et al. 1995) address it.

In ACT-R, the function of the subsymbolic computations is to identify the right chunks and productions out of a large data base, and the rational analysis provides a "proof" of the performance of these computations. The success of these computations has been demonstrated in "life-time" learning of cognitive arithmetic (Lebiere 1998) and past-tense learning (Taatgen 2001). However, they have been models of limited domains, and the knowledge base has been relatively small. There have been no ACT-R models of performance with large knowledge bases approaching human size. The subsymbolic mechanisms are motivated to work well with large knowledge bases, but that is no guarantee that they will. The one case of dealing with a large knowledge base in ACT-R is the effort (Emond, in preparation) to implement WordNet (Fellbaum 1998) in ACT-R, which involves more than 400,000 chunks, but this implementation awaits more analysis.

5.5. Dynamic behavior

Grading: *Connectionism: Mixed*
ACT-R: Better

Connectionism has some notable models of interaction with the environment such as ALVINN and its successors, which were able to drive a vehicle, although it was primarily used to drive in fairly safe predictable conditions (e.g., straight highway driving) and was disabled in challenging conditions (interchanges, perhaps even lane changes). However, as exemplified in this model, connectionism's conception of the connections among perception, cognition, and action is pretty ad hoc, and most connectionist models of perception, cognition, and action are isolated, without the architectural structure to close the loop, especially in timing specifications. McClelland's (1979) Cascade model offers an interesting conception of how behavior might progress from perception to action, but this conception has not actually been carried through in models that operate in dynamic environments.

Many ACT-R models have closed the loop, particularly in dealing with dynamic environments like driving, air traffic control, simulation of warfare activities, collaborative problem solving with humans, control of dynamic systems like power plants, and game playing. These are all domains where the behavior of the external system is unpredictable. These simulations take advantage of both ACT-R's ability to learn and the perceptual-motor modules that provide a model of human attention. However, ACT-R is only beginning to deal with tasks that stress its ability to respond to task interruption. Most ACT-R models have been largely focused on single goals.

5.6. Knowledge integration

Grading: *Connectionism: Worse*
ACT-R: Mixed

We operationalized Newell's symbolic criterion as achieving the intellectual combination that he thought physical symbols were needed for. Although ACT-R does use physical symbols more or less in Newell's sense, this does not guarantee that it has the necessary capacity for intellectual combination. There are demonstrations of it making inference (Anderson et al. 2001), performing induction (Haverty et al. 2000), metaphor (Budiu 2001), and analogy (Salvucci & Anderson 2001), and these all do depend on its symbol manipulation. However, these are all small-scale, circumscribed demonstrations, and we would not be surprised if Fodor found them less than convincing.

Such models have not been as forthcoming from classical connectionism (Browne & Sun 2001). A relatively well-known connectionist model of analogy (Hummel & Holyoak 1998) goes beyond classical connectionist methods to achieve variable binding by means of temporal synchrony. The Marcus demonstration of infants' learning rules has become something of a challenge for connectionist networks. It is a relatively modest example of intellectual combination – recognizing that elements occurring in different positions need to be identical to fit a rule and representing that as a constraint on novel input. The intellectual elements being combined are simply sounds in the same string. Still, it remains a challenge to classical connectionism, and some classical connectionists (e.g., McClelland & Plaut 1999) have chosen instead to question whether the phenomenon is real.

5.7. Natural language

Grading: *Connectionism: Better*
ACT-R: Worse

Connectionism has a well-articulated conception of how natural language is achieved, and many notable models that instantiate this conception. However, despite efforts like Elman's, it is a long way from providing an adequate account of human command of the complex syntactic structure of natural language. Connectionist models are hardly ready to take the SAT. ACT-R's treatment of natural language is fragmentary. It has provided models for a number of natural-language phenomena including parsing (Lewis 1999), use of syntactic cues (Matessa & Anderson 2000), learning of inflections (Taatgen 2001), and metaphor (Budiu 2001).

ACT-R and connectionism take opposite sides on the chicken-and-egg question about the relationship between symbols and natural language that Newell and others wondered about: Natural-language processing depends in part on ACT-R's symbolic capabilities, and it is not the case that natural-language processing forms the basis of the symbolic capabilities, nor is it equivalent to symbolic processing. However, classical connectionists are quite explicit that whatever might appear to be symbolic reasoning really depends on linguistic symbols like words or other formal symbols like equations.

5.8. Consciousness

Grading: *Connectionism: Worse*
ACT-R: Worse

The stances of connectionism and ACT-R on consciousness are rather similar. They both have models (e.g., Cleeremans 1993; Wallach & Lebiere 2000; in press) that treat one of the core phenomena – implicit memory – in the discussion of consciousness. However, neither have offered an analysis of subliminal perception or metacognition. With respect to functionality of the implicit/explicit distinction, ACT-R holds that implicit memory represents the subsymbolic information that controls the access to explicit declarative knowledge. To require that this also be explicit, would be inefficient and invite infinite regress.

ACT-R does imply an interpretation of consciousness. Essentially, what people are potentially conscious of is contained in ACT-R's set of buffers in Figure 1 – the current goal, the current information retrieved from long-term memory, the current information attended in the various sensory modalities, and the state of various motor modules. There are probably other buffers not yet represented in ACT-R to encode internal states like pain, hunger, and various pleasures. The activity of consciousness is the processing of these buffer contents by production rules. There is no Cartesian Theater (Dennett 1991; Dennett & Kinsbourne 1995) in ACT-R. ACT-R is aware of the contents of the buffers only as they are used by the production rules.

5.9. Learning

Grading: *Connectionism: Better*
ACT-R: Better

A great deal of effort has gone into thinking about and modeling learning in both connectionist models and ACT-R. However, learning is such a key issue and so enormous a problem that both have much more to do. They display complementary strengths and weaknesses. While connectionism has accounts to offer of phenomena in semantic memory like semantic dementia (Rogers & McClelland 2003), ACT-R has been able to provide detailed accounts of the kind of discrete learning characteristic of episodic memory such as the learning of lists or associations (Anderson et al. 1998a; Anderson & Reder 1999a). Whereas there are connectionist accounts of phenomena in perceptual and motor learning, ACT-R offers accounts of the learning of cognitive skills like mathematical problem solving. Whereas there are connectionist accounts of perceptual priming, there are ACT-R accounts of associative priming. The situation with respect to conditioning is interesting. On the one hand, the basic connectionist learning rules have a clear relationship to some of the basic learning rules proposed in the conditioning literature, such as the Rescorla-Wagner rule (see Anderson [2000] for a discussion). On the other hand, known deficits in such learning rules have been used to argue that at least in the case of humans, these inferences are better understood as more complex causal reasoning (Schoppek 2001).

5.10. Development

Grading: *Connectionism: Better*
ACT-R: Worse

As with language, development is an area that has seen a major coherent connectionist treatment but only spotty efforts from ACT-R. Connectionism treats development as basically a learning process, but one that is constrained by the architecture of the brain and the timing of brain development. The connectionist treatment of development is in some ways less problematic than its treatment of learning because connectionist learning naturally produces the slow changes characteristic of human development. Classical connectionism takes a clear stand on the empiricist–nativist debate, rejecting what it calls representational nativism.

In contrast, there is not a well-developed ACT-R position on how cognition develops. Some aspects of a theory of cognitive development are starting to emerge in the guise of cognitive models of a number of developmental tasks and phenomena (Emond & Ferres 2001; Jones et al. 2000; Simon 1998; submitted; Taatgen & Anderson 2002; van Rijn et al. 2000). The emerging theory is one that models child cognition in the same architecture as adult cognition and that sees development as just a matter of regular learning. Related to this is an emerging model of individual differences (Jongman & Taatgen 1999; Lovett et al. 2000) that relates them to a parameter in ACT-R that controls the ability of associative activation to modulate behavior by context. Anderson et al. (1998b) argue that development might be accompanied by an increase in this parameter.

5.11. Evolution

Grading: *Connectionism: Worst*
ACT-R: Worst

Both theories, by virtue of their analysis of the Bayesian basis of the mechanisms of cognition, have something to say about the adaptive function of cognition (as they were credited with under Criterion 3), but neither has much to say about how the evolution of the human mind occurred. Both theories basically instantiate the puzzlement expressed by Newell as to how to approach this topic.

We noted earlier that cognitive plasticity seems a distinguishing feature of the human species. What enables this plasticity in the architecture? More than anything else, ACT-R's goal memory enables it to abstract and retain the critical state information needed to execute complex cognitive procedures. In principle, such state maintenance could be achieved using other buffers – speaking to oneself, storing and retrieving state information from declarative memory, writing things down, and so forth. However, this would be almost as awkward as getting computational universality from a single-tape Turing machine, besides being very error-prone and time-consuming. A large expansion of the frontal cortex, which is associated with goal manipulations, occurred in humans. Of course, the frontal cortex is somewhat expanded in other primates, and it would probably be

unwise to claim that human cognitive plasticity is totally discontinuous from that of other species.

5.12. Brain

Grading: Connectionism: Best
ACT-R: Worse

Classical connectionism, as advertised, presents a strong position on how the mind is implemented in the brain. Of course, there is the frequently expressed question of whether the brain that classical connectionism assumes happens to correspond to the human brain. Assumptions of equipotentiality and the backprop algorithm are frequent targets for such criticisms, and many nonclassical connectionist approaches take these problems as starting points for their efforts.

There is a partial theory about how ACT-R is instantiated in the brain. ACT-RN has established the neural plausibility of the ACT-R computations, and we have indicated rough neural correlates for the architectural components. Recently completed neural imaging studies (Anderson et al. 2003; Fincham et al. 2002; Sohn et al. 2000) have confirmed the mapping of ACT-R processes onto specific brain regions (e.g., goal manipulations onto the dorsolateral prefrontal cortex). There is also an ACT-R model of frontal patient deficits (Kimberg & Farah 1993). However, there is not the systematic development that is characteristic of classical connectionism. While we are optimistic that further effort will improve ACT-R's performance on this criteria, it is not there yet.

6. Conclusion

Probably others will question the grading and argue that certain criteria need to be re-ranked for one or both of the theoretical positions. Many of the arguments will be legitimate complaints, and we are likely to respond by either defending the grading, or conceding an adjustment in it. However, the main point of this target article is that the theories should be evaluated on all 12 criteria, and the grades point to where the theories need more work.

Speaking for ACT-R, where will an attempt to improve lead? In the case of some areas like language and development, it appears that improving the score simply comes down to adopting the connectionist strategy of applying ACT-R in depth to more empirical targets of opportunity. We could be surprised, but so far these applications have not fundamentally impacted the architecture. The efforts to extend ACT-R to account for dynamic behavior through perception and action yielded a quite different outcome. At first, ACT-R/PM was just an importation, largely from EPIC (Meyer & Kieras 1997) to provide input and output to ACT-R's cognitive engine. However, it became clear that ACT-R's cognitive components (the retrieval and goal buffers in Fig. 1) should be redesigned to be more like the sensory and motor buffers. This led to a system that more successfully met the dynamic behavior criterion and has much future promise in this regard. Thus, incorporating the perceptual and motor modules fundamentally changed the architecture. We suspect that similar fundamental changes will occur as ACT-R is extended to deal further with the brain criterion.

Where would attention to these criteria take classical connectionism? First, we should acknowledge that it is not clear that classical connectionists will pay attention to these criteria or even acknowledge that the criteria are reasonable. However, if they were to try to achieve the criteria, we suspect that it would move connectionism to a concern with more complex tasks and symbolic processing. We would not be surprised if it took them in a direction of a theory more like ACT-R, even as ACT-R has moved in a direction that is more compatible with connectionism. Indeed, many attempts have been made recently to integrate connectionist and symbolic mechanisms into hybrid systems (Sun 1994; 2002). More generally, if researchers of all theoretical persuasions did try to pursue a broad range of criteria, we believe that distinctions among theoretical positions would dissolve and psychology will finally provide "the kind of encompassing of its subject matter – the behavior of man – that we all posit as a characteristic of a mature science" (Newell 1973, p. 288).

NOTE

1. The complete list of published ACT-R models between 1997 and 2002 is available from the ACT-R home page at: **act.psy.cmu.edu**

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by ONR grant N00014–96–1-C491. We would like to thank Gary Marcus and Alex Petrov for their comments on this manuscript. We would also like to thank Jay McClelland and David Plaut for many relevant and helpful discussions, although we note they explicitly chose to absent themselves from any participation that could be taken as adopting a stance on anything in this paper.

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Newell's list

Joseph Agassi

Department of Philosophy, Tel-Aviv University, Tel-Aviv 69978, Israel.
agass@post.tau.ac.il http://www.tau.ac.il/~agass/

Abstract: Newell wanted a theory of cognition to abide by some explicit criteria, here called the Newell Test. The test differs from the Turing Test because it is explicit. The Newell Test will include the Turing Test if its characterization of cognition is complete. It is not. Its use here is open-ended: A system that does not pass it well invites improvement.

Alan Newell asserted that an adequate theory of a functioning system of human cognition should abide by some explicit criteria, and he offered a list of such criteria. The list includes characteristics such as flexible, adaptive behavior; possession of a vast knowledge base; and the ability to integrate knowledge, use a natural language, and learn. The target article authors say that, although this list is not complete, it certainly is "enough to avoid theoretical myopia" (sect. 1, para. 2). Hardly: Myopia is the outcome of the claim for knowledge of natural languages and learning sufficient to per-

mit decision as to whether a given theory of cognition captures them adequately. We just do not know that much as yet.

The authors say that the criteria deserve "greater scientific prominence." They therefore try to "evaluate theories by how well they do at meeting" the criteria (sect. 1, para. 4). This may be premature. Whether it is, depends on the merit of Newell's idea more than on its applications. So, it requires examination. What the authors call the Newell Test is a test not of Newell's idea but of the theories that should agree with it – provided that it is valid. Is it? How are we to judge this?

Anderson & Lebiere (A&L) apply Newell's Test to two new ideas that are controversial, so the application cannot be such a test. Hence, their work is begging the question: Some test of it is required to show that it deserves a "greater scientific prominence."

"Newell is calling us to consider all the criteria and not pick and choose the ones to consider" (sect. 2.8). This remark renders the whole venture too questionable. The authors make it apropos discussion of the criterion of consciousness.

> Newell acknowledged the importance of consciousness to a full account of human cognition, although he felt compelled to remark that "it is not evident what functional role self-awareness plays in the total scheme of mind." We too have tended to regard consciousness as epiphenomenal . . . (sect. 2.8)

This is very shaky. Whether consciousness is or is not epiphenomenal is a red herring: It is an empirical fact that in many cases cognitive conduct differs depending on whether it is accompanied with consciousness or not, and the question may arise, should a system emulating human consciousness reflect this fact? Importantly, Turing's celebrated idea of the Turing Test is designed to avoid this question altogether.

The authors examine two sets, classical connectionism and ACT-R. Classical connectionism is a computerized version of behaviorism. ACT-R is "a theory of higher-level cognition," "a subsymbolic activation-based memory" able "to interact with a symbolic system of production rules"; the R in ACT-R "denotes rational analysis" (sect. 4.1, first paragraph). The two sets, then, are artificial intelligence or expert-systems programs. The authors report a claim that classical connectionism passes the Turing Test. Presumably they disagree. The same holds for ACT-R. "ACT-R, but for its subsymbolic limitations, is Turing equivalent, as are most production systems" and "(proof for an early version of ACT [is due to] Anderson . . .)" (sect. 5.1, para. 2). This is a bit cryptic; I will explain the difference between the Turing and the Newell Tests in the following paragraph.

The Turing Test was meant to render the mind-body problem empirically decidable. Were there a computer program that could fool an expert, Turing suggested, then it would be empirically indistinguishable from humans, and so the attribution to humans of a metaphysical soul would be redundant. Because Newell's criteria depict human characteristics, any interlocutor who can pass the Turing Test should certainly possess them, because the inability to exhibit any human characteristic the like of which Newell mentions would expose the impostor. And yet, the Turing Test is implicit and Newell's Test is explicit. This permits finding a partial success in passing the Newell Test. But, to be an explicit version of the Turing Test, the Newell Test must refer to a complete list of characteristics. We do not have this, and the Turing Test may be preferred just because it leaves this task to the experts who wish to test the humanity of their enigmatic interlocutor. Consequently, a Turing Test can never be decisive: Both expert and programmer can improve on prior situations and thus deem failure a merely temporary setback. True, the Turing Test is generally deemed possibly decisive, and, being a thought-experiment, actually decisive. Some writers, notably Daniel Dennett, claim that only exorbitant costs prevent the construction of a machine that will pass the Turing Test. That machine, then, should certainly be able to pass the Newell Test with flying colors. It is a pity that A&L do not refer to this claim and expose it as a sham. If they are any close to being right, they should be able to do so with ease.

The interesting aspect of the target article is that it is open-ended: Whenever the system A&L advocate, which is ACT-R, does not pass the examination as well as they wish, they recommend trying an improvement, leading to a retest. They should observe that such a move may be two-pronged. They refer to the improvement of the ability of a program to abide by the theory of flexibility; adaptive behavior; and the ability to integrate knowledge, use a natural language, and learn. They should not ignore the need to improve on these theories. When they refer to natural languages or to learning, they view the connectionist idea of them as more satisfactory than that of ACT-R, because it is more complete. Yet, whatever completeness is exactly, it is not enough: We seek explanations, and so to accept axiomatically what we want to understand is not good enough. We still do not know what a natural language is and how we learn; and we do not begin to understand these. Let me end with an insight of David Marr that should not be forgotten. Emulation is helpful for the understanding but is no substitute for it; sometimes, the very success of emulation, Marr (1982) observed, renders it less useful as a problematic one. We want understanding, not mere emulation.

Think globally, ask functionally

Erik M. Altmann

Department of Psychology, Michigan State University, East Lansing, MI 48824. **ema@msu.edu** **http://www.msu.edu/~ema**

Abstract: The notion of functionality is appropriately central to the Newell Test but is also critical at a lower level, in development of cognitive sub-theories. I illustrate, on one hand, how far this principle is from general acceptance among verbal theoreticians, and, on the other hand, how simulation models (here implemented within ACT-R) seem to drive the functional question automatically.

Anderson & Newell (A&L) have been carrying integrative cognitive theory, in shifts, for the past 30 years or so (if one goes back to Newell 1973). We are fortunate that Anderson is young; formulating dichotomous questions – seeing the trees but not the forest – may be the modal tenure procedure in psychology departments today, but perhaps in another generation it will be acceptable not to conduct new experiments at all but simply to integrate old data into increasingly complete computational models.

In the meantime, how can we avoid theoretical myopia in our daily research? Applying the Newell Test is well and good once a decade or so, with that many years' interim progress available to assess it. In terms of the next chunk of publishable research, however, it's useful to have more immediate guidance.

Central to the Newell Test is the idea of functionality: A theory has to explain how the cognitive system accomplishes some particular function. Among the Newell Test criteria, this function is high level, related in some relatively direct way to the fitness of the organism. However, as one develops micro-theories within a larger theory, functionality is still a relevant question; one can ask, for each process within a model, what it is for. Its outputs could, for example, be necessary inputs for another process in a chain that leads ultimately to accomplishing the task at hand. Or, one could ask whether each behavioral measure reflects a distinct process at all; perhaps it reflects a side effect of some other functionally necessary process. In both cases, it is difficult if not impossible to address the functional question without a precise representation of the processes one is talking about. In practice, this implies a computational simulation.

How does functionality play out at the level of the micro-theory that is the next chunk of publishable research? Curiously, even at this level, functionality seems to be regarded as optional, if not actually vulgar. A&L raise the example of short-term memory constructs (and Newell's frustration over them), but let's have a newer one, if only to see what might have changed. In the domain of ex-

ecutive control, there is a burgeoning literature on "switch cost" – the time cost associated with switching to a different task, as compared to performing the same task over again. One regularity to have emerged is that switch cost is difficult to erase; even with time and motivation to prepare for the other task, people are slower on the first trial under that task than on the second. The dominant theoretical account of this residual switch cost is arguably the "stimulus cued completion" hypothesis of Rogers and Monsell (1995, p. 224):

> This hypothesis proposes that an endogenous act of control deployed before onset of the stimulus can achieve only part of the process of task-set reconfiguration. Completion of the reconfiguration is triggered only by, and must wait upon, the presentation of a task-associated stimulus.

In terms of functionality, this hypothesis is vacuous. It need not be; one could ask how the system might benefit from stimulus-cued completion. For example, one could propose a benefit to the system hedging its bets and waiting to complete the reconfiguration process until there is evidence (in the form of the trial stimulus) that the new task set will be needed. One could then try to formulate scenarios in which this benefit would actually be realized and evaluate them for plausibility, or perhaps even against existing data. None of this was attempted by Rogers and Monsell, or by authors since who have invoked stimulus-cued completion as an explanatory construct. Call this a working definition of theoretical myopia: a "hypothesis" that merely relabels an empirical phenomenon.

In a subsequent ACT-R model, Sohn and Anderson (2001) explain residual switch cost in terms of stochasticity. Their model contains a "switching" production that retrieves the new task from memory and installs it in the system's focus of attention. Selection of productions is, like most other cognitive processes, subject to noise, which explains why this production is not always selected in advance of stimulus onset. Functionally, it can be selected after stimulus onset, though must be selected before response selection. This account is an improvement; it makes predictions (in terms of response-time variability), and it explains residual switch cost as a side-effect of the noise that accompanies any communication channel.

One could go further and ask, does residual switch cost reflect a process that directly contributes in some way to task performance? In another ACT-R model, Gray and I proposed that residual switch cost reflects a redundant task-encoding process that affects quality control (Altmann & Gray 2000). (Initial task encoding activates a memory trace for the current task, but noisily; redundant task encoding catches and properly strengthens memory traces that were initially weakly encoded.) The proof of functionality lay in Monte Carlo simulations showing that overall performance accuracy was higher with this redundant phase than without.

Are Sohn and Anderson right, or are Altmann and Gray? We have not found a behavioral test; perhaps neuroimaging will someday afford a diagnostic. I would predict, however, that the stimulus-cued completion hypothesis will not find its way into a precisely formulated cognitive theory, micro or otherwise, unless relevant functional questions are posed first.

ACKNOWLEDGMENT

This work was supported by an Office of Naval Research Grant N00014–03–1–0063.

The Newell Test should commit to diagnosing dysfunctions

William J. Clancey

Computational Sciences Division, MS 269–3, NASA Ames Research Center, Moffett Field, CA 94035. **william.j.clancey@nasa.gov**
http://bill.clancey.name

Abstract: "Conceptual coordination" analysis bridges connectionism and symbolic approaches by positing a "process memory" by which categories are physically coordinated (as neural networks) in time. Focusing on dysfunctions and odd behaviors, like slips, reveals the function of consciousness, especially constructive processes that are often taken for granted, which are different from conventional programming constructs. Newell strongly endorsed identifying architectural limits; the heuristic of "diagnose unusual behaviors" will provide targets of opportunity that greatly strengthens the Newell Test.

Anderson & Lebiere's (A&L's) article evaluates cognitive theories by relating them to the criteria of functionality derived from Newell. Suppose that the Newell Test (NT) has all the right categories, but still requires a significant architectural change for theoretical progress. I claim that "conceptual coordination" (CC) (Clancey 1999a) provides a better theory of memory, and that, without committing to explaining cognitive dysfunctions, NT would not provide sufficient heuristic guidance for leading in this direction.

Conceptual coordination (CC) hypothesizes that the store, retrieve, and copy memory mechanism is not how the brain works. Instead, all neural categorizations are activated, composed, and sequenced "in place," with the assumption that sufficient (latent) physical connections exist to enable necessary links to be formed (physically constructed) at run time (i.e., when a behavior or experience occurs). For example, if comprehending a natural language sentence requires that a noun phrase be incorporated in different ways, it is not moved or copied but is physically connected by activation of (perhaps heretofore unused) neural links. Effectively, Newell's "distal access" is accomplished by a capability to hold a categorization active and encapsulate it (like a pointer) so that it can be incorporated in different ways in a single construction. The no-copying constraint turns out to be extremely powerful for explaining a wide variety of odd behaviors, including speaking and typing slips, perceptual aspects of analogy formation, developmental "felt paths," multimodal discontinuity in dreams, and language comprehension limitations. CC thus specifies a cognitive architecture that bridges connectionist and symbolic concerns; and it relates well to the NT criteria for which ACT-R scores weakest – development, consciousness, language, and the brain. To illustrate, I provide a diagnostic analysis of an autistic phenomenon and then relate this back to how NT can be improved.

In CC analysis, a diagram notation is used to represent a behavior sequence, which corresponds in natural language to the conceptualization of a sentence. For example, according to Baron-Cohen (1996), an autistic child can conceptualize "I stroke the cat that drinks the milk." In one form of the CC notation, a slanting line to the right represents categorizations activated sequentially in time (e.g., "I – stroke" in Figure 1). Another sequence may qualify a categorization (e.g., "the cat – drinks" qualifies "stroke"). This pattern of sequences with qualifying details forming compositions of sequences occurs throughout CC analysis. The essential idea in CC is to understand how categories (both perceptual and higher-order categorizations of sequences and compositions of them) are related in time to constitute conscious experience (Clancey1999a).

The challenge is to understand why an autistic child finds it problematic to conceptualize "I see the cat that sees the mouse." A traditional view is that the child lacks social understanding. But CC analysis suggests a mechanistic limitation in the child's ability to physically sequence and compose categories. Relating to other agents requires being able to construct a second-order conceptualization that relates the child's activity to the other agent's activity. Figure 2 shows the CC notation for the required construction.

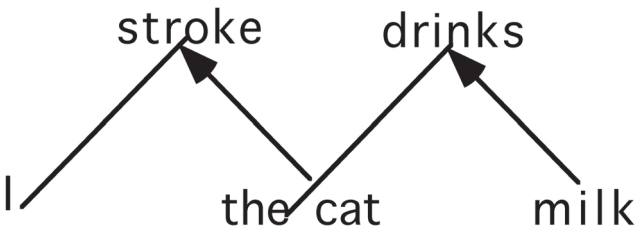

Figure 1 (Clancey). Unproblematic: "I stroke the cat that drinks the milk."

The statement (the conceptualization being constructed) involves a triadic relation: I see the cat, the cat sees the mouse, and I see the mouse. There is one mouse that we are both seeing. Two "see" constructions are unified by identifying a detail (the mouse) as common to both. In effect, the child must conceive of a problem space (Clancey 1999a): A common categorization of an operand (mouse) enables categorization of multiple actions as being one action (seeing), an operator. Because the two actions are by different agents, accomplishing this identification integrates perspectives of *self* (what I am doing now) and *other* (what that object is doing now). Indeed, the conceptualization of agent appears to be inherent in this construction.

Put another way, two sequentially occurring conceptualizations (I see the cat; the cat sees the mouse) are held active and related: "I see the cat that sees the mouse" and "I see the mouse" become "I see that the cat sees the mouse" (i.e., the mouse that I am seeing). (The second-order relation is represented in Figure 2 by the solid arrow below "I see"). Conceiving this relation is tantamount to conceiving what joint action is. Barresi and Moore (1996) characterize this as "integrating third and first person information" (p. 148), and contrast it with (Figure 1) "embedding one third person representation in a separate first person frame" (p. 148). Related to Langacker's (1986) analysis, logical relations are not extra capabilities or meta "inference" capabilities, but generalizations of concrete accomplishments that arise through the capability to physically coordinate categories through identification, sequence, and composition in time. Mental operations are physical, subconscious processes, constrained by physical limits on how inclusion in new sequences can occur. The ability to hold two sequences active and relate them constitutes a certain kind of consciousness (e.g., not present in dreaming; Clancey 2000).

To summarize, the example requires relating sequential categorizations of seeing so that they become simultaneous; it exemplifies a second-order conceptualization of intentionality (my seeing is about your seeing; Clancey 1999b); and suggests that joint action requires being able to conceive the ideas we call operator and agent.

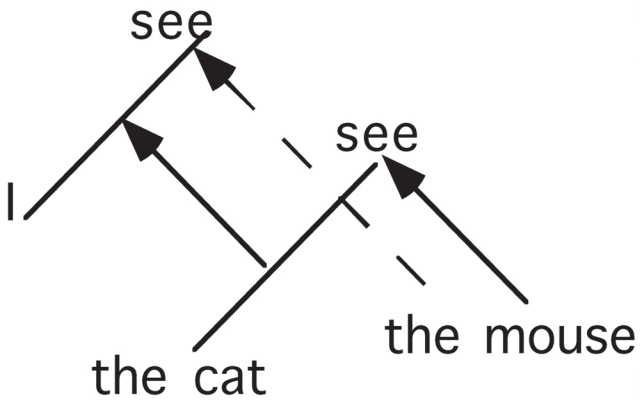

Figure 2 (Clancey). Problematic: "I see the cat that sees the mouse."

The pivotal heuristic in CC analysis is addressing *unusual* behaviors and experiences. These "targets of opportunity" appear to be de-emphasized by A&L's focus on normal behaviors "that people display on a day-to-day basis." For NT to provide heuristic guidance for discovering a theory like CC, grading for each criteria should include diagnosing unusual phenomena that everyone experiences (e.g., slips) and dysfunctions. For example, for the criteria of consciousness, we should direct theorization at explaining the phenomenology of dreaming, autism, compulsive-obsessive disorders, and the like. For natural language, include comprehension difficulties (e.g., subject relatives with center-embedded noun phrases; Clancey 1999a, Ch. 10). For development, explain how "felt paths" are constructed in children's learning (Ch. 5). For knowledge integration, explain slips (Ch. 6) and "seeing as" in analogy formation (Ch. 7). In this manner, learning in well-known architectures (e.g., MOPS, EPAM, SOAR) can be evaluated and the nature of problem spaces reformulated (Ch. 12).

The evolution criterion highlights the limitations of NT as stated. Rather than focusing on human evolution, this criterion should be about the evolution of cognition broadly construed, and hence should be inherently comparative across species (Clancey 1999b). Viewed this way, there is no "paucity of data," but rather a largely unexploited potential to make the study of animal cognition an integrated discipline with human problem solving. By including the heuristic "explain odd behavior" in the grading, we will naturally be guided to characterize and relate cognition in other primates, ravens, and the like. This is essential for relating "instinctive" mechanisms (e.g., weaving spider webs) to brain mechanisms, development, and learned higher-order categorizations (e.g., conceptualization of intentionality). A&L mention comparative considerations, but we should view this as a diagnostic problem, much as cognitive theories like ACT* have been used to explain students' different capabilities (Anderson et al. 1990). Furthermore, the research community should collect behaviors that have been heretofore ignored or poorly explained by computational theories and include them in the grading criteria.

Applying the Newell Test in this way – moving from the routine behaviors already handled more or less well, to diagnostic theories that relate aberrations to architectural variations – might bring symbolic and connectionist theories together and make the study of cognition a more mature science.

A complete theory of tests for a theory of mind must consider hierarchical complexity and stage

Michael Lamport Commons and Myra Sturgeon White
Department of Psychiatry, Harvard Medical School, Massachusetts Mental Health Center, Boston, MA 02115-6113. **Commons@tiac.net**
mswhite@fas.harvard.edu **http://www.tiac.net/~commons/**

Abstract: We distinguish traditional cognition theories from hierarchically complex stacked neural networks that meet many of Newell's criteria. The latter are flexible and can learn anything that a person can learn, by using their mistakes and successes the same way humans do. Shortcomings are due largely to limitations of current technology.

Anderson & Lebiere (A&L) raise important issues concerning criteria for evaluating the cognitive theories on which computational systems designed to simulate human intellectual abilities are based. Typically, cognitive theories are indirectly evaluated based on a theory's capacity to be translated into a computational system that produces correct answers or workable rules. The Newell 12-Criteria Test (1992; Newell & Simon 1963/1995) that A&L propose to measure theories with, makes an important move towards

measuring a theory's capacity to exhibit underlying behaviors supporting the expression of human cognitive processes.

We suggest a further dimension. Most cognitive theories are, like Athena, born fully formed, modeling the highest stages of development. However, human cognition is a product of developmental process. Humans learn to act by building one stage's actions on actions from previous stages, creating the capacity to perform ever more complex behaviors. Thus, to fully explain or model human intellectual capacity, hierarchical complexity must be factored into a theory. The *Model of Hierarchical Complexity* (MHC) (Commons et al. 1998) delineates these developmental changes (see Dawson 2002 for validity and reliability).

MHC identifies both sequences of development and reasons why development occurs from processes producing stage transition. It may be used to define complex human thought processes and computer systems simulating those processes. With this model, performed tasks are classified in terms of their order of hierarchical complexity using the following three main axioms (Commons et. al 1998). Actions at a higher order of hierarchical complexity

1. Are defined in terms of lower order actions;
2. Organize and transform lower stage actions;
3. Solve more complex problems through the nonarbitrary organization of actions.

The order of the hierarchical complexity of a task is determined by the number of its concatenation operations. An order-three task action has three concatenation operations and operates on output from order-two actions, which by definition has two concatenation operations and operates on an order-one task action. Increases in the hierarchical complexity of actions result from a dialectical process of stage transition. (Commons & Richards 2002).

To stimulate human intellectual capacities in computer systems, we design stacked neural networks that recapitulate the developmental process. This approach is necessary because currently we lack the knowledge to build into systems the myriad key behaviors formed during the developmental processes. Moreover, we lack the technology to identify the intricate web of neural connections that are created during the developmental process.

These stacked neural networks go through a series of stages analogous to those that occur during human intellectual development. Stages of development function as both theory and process in these systems. Actions (i.e., operations performed by networks resulting in a changed state of the system) are combined to perform tasks with more complex actions, permitting the performance of more complex tasks and thereby scaling up the power. The number of neural networks in a stack is the highest order of hierarchical complexity of task-required actions identified by the model. An example of a six-stage stacked neural network based on the model of hierarchical complexity (Table 1) follows.

Example. A system answers customer telephone calls, transferring them to the proper area within a large organization. Transfers are based on the customer's oral statements and responses to simple questions asked by the system. The system is capable of a three-year-old's language proficiency. A front-end recognition system translates customers' utterances (system inputs) into words that will serve as simple stimuli. It also measures time intervals between words.

Stacked neural networks based on the MHC meet many of Newell's criteria. They are flexible and can learn anything that a person can learn. They are adaptive because their responses are able to adjust when stimuli enter the stack at any level. They are dynamic in that they learn from their mistakes and successes. In the example, the system adjusts the weights throughout the stack of networks if a customer accepts or rejects the selected neural network location. Knowledge integration occurs throughout the networks in the stack. Moreover, networks based on the MHC learn in the same way as humans learn.

Some criteria are less easily met. Given current technology, neural networks cannot function in real time, are unable to trans-

Table 1 (Commons & White). *Stacked Neural Network (Example of Model of Hierarchical Complexity)*

Order of Hierarchical Complexity	What It Uses	What It Does
0. Calculatory	From Humans	Calculates and executes human written programs
1. Sensory and motor	Caller's utterances	A front-end speech recognition system translates customers' utterances into words. These "words" serve as simple stimuli to be detected.
2. Circular sensory motor	Words from speech recognition system	Forms open-ended classes consisting of groups contiguous individual words
3. Sensory-motor	Grouped contiguous speech segments	Labels and maps words to concepts. Networks are initially taught concepts that are central to the company environment: products and departments such as customer service, billing, and repair.
4. Nominal	Concept domains	Identifies and labels relationships between concept domains. Possible interconnections are trained based on the company's functions, products, and services. Interconnections are adjusted based on system success.
5. Sentential	Joint concept domains	Forms simple sentences and understands relationships between two or more named concepts. Finds possible locations to send customer's calls. Constructs statement on whether they want to be transferred to that department. Customer's acceptances or rejection feeds back to lower levels.

fer learning despite abilities to acquire a vast knowledge base, and cannot exhibit adult language skills. Whether we can build evolutions into systems – or even want to – is open to question. Finally, given our current limited understanding of the brain, we can only partially emulate brain function.

Nonclassical connectionism should enter the decathlon

Francisco Calvo Garzón

Department of Philosophy, Indiana University, Bloomington, IN, and University of Murcia, Facultad de Filosofía, Edif. Luis Vives, Campus de Espinardo Murcia 30100, Spain. **fjcalvo@um.es**

Abstract: In this commentary I explore nonclassical connectionism (NCC) as a coherent framework for evaluation in the spirit of the Newell Test. Focusing on knowledge integration, development, real-time performance, and flexible behavior, I argue that NCC's "within-theory rank ordering" would place subsymbolic modeling in a better position. Failure to adopt a symbolic level of thought cannot be interpreted as a weakness.

Granting Anderson & Lebiere's (A&L's) "cognitive decathlon" overall framework, and their proposed operationalizations and grading scheme for theory-evaluation, the aspects of their article that I address here concern the choice of contestants entering the decathlon, and, based on that choice, the exploration of nonclassical connectionism (NCC) as a coherent framework for evaluation in the spirit of the Newell Test. The range of classical connectionist architectures that A&L assess is confined to models that have a feedforward or a recurrent architecture, a locally supervised learning algorithm (e.g., backpropagation), and a simple nonlinear activation function (e.g., sigmoidal). A nonclassical framework, however, can be coherently developed. By NCC, I shall be referring to the class of models that have different combinations of pattern associator/autoassociative memory/competitive network topologies, with bidirectional connectivity and inhibitory competition, and that employ combined Hebbian and activation-phase learning algorithms (O'Reilly & Munakata 2000; Rolls & Treves 1998). Were NCC allowed to enter the competition, it would (or so I shall argue) obtain a "within-theory rank ordering" that could perhaps place it in a better position than the ACT-R theory. To demonstrate this, I will make three points with regard to 4 of the 12 functional constraints on the architecture of cognition that A&L take into consideration: knowledge integration, development, real-time performance, and flexible behavior.

On knowledge integration, classical connectionism (CC) gets a "worse" grade (see Table 1 of the target article). As an "intellectual combination" example of knowledge integration, A&L consider the literature on transfer of learning in infants. Marcus (2001) assessed the relationship between CC and rule-governed behavior by challenging the connectionist to account for experimental data that had been interpreted as showing that infants exploit (rule-governed) abstract knowledge in order to induce the implicit grammar common to different sequences of syllables (Marcus et al. 1999). Calvo and Colunga (in preparation) show how Marcus's infants-data challenge can be met with NCC (see Calvo & Colunga 2003, for a CC replica of this simulation). Our model (Fig. 1) is based on a simple recurrent network (SRN) architecture that has been supplemented with the following nonclassical features: (1) bidirectional (symmetric) propagation of activation, (2) inhibitory competition, (3) an error-driven form of learning (GenRec in McClelland 1994), and (4) the Hebbian model learning.

The fundamental component of our simulation resides in the fact that the network is pretrained with syllables that can be either duplicated or not. These first-order correlations in the environment amount to subregularities that can be exploited by the network in a semideterministic prediction task. During pretraining, the network learns to represent something general about duplication (i.e., sameness). This abstraction is crucial in encoding the patterns during the habituation phase. Like the infants in Marcus et al.'s study, the networks that were pretrained in a corpus in which some syllables were consistently duplicated learned to distinguish ABB patterns from ABA patterns after a brief period of training akin to infant's habituation.

Error-driven learning makes use of an activation-phase algorithm that, via bidirectional connectivity and symmetric weight matrices, permits the network to alter the knowledge acquired in the weights by computing the difference between an initial phase where the networks activations are interpreted as its "expectation" of what's to happen, and a later phase in which the environment provides the output response to be taken as the teaching signal. Activation-based signals in a prediction task are not to be interpreted in Marcus's terms. The ecologically grounded prediction task of the networks does not incorporate universally open-ended rules. Unsupervised Hebbian learning, on the other hand, makes its contribution by representing in hidden space the first-order correlational structure of the data pool. Our NCC architecture delivers a correct syntactic interpretation of the infants' data. The data are accounted for without the positing of rule-fitting patterns of behavior (allegedly required to constrain novel data).

On development, where CC is graded as "better," the score may be made even more robust. A&L remark upon CC's anti-nativist stance on the nature/nurture debate. Marín et al. (2003) argue, in the context of poverty-of-stimulus arguments in Creole genesis, that CC eschews any form of nativism. Creole genesis, nativists contend, can only be explained by appealing to a Chomskian Universal Grammar (UG). Substratists contend that Creole genesis is influenced, crucially, by substratum languages. We show how the process by which a Pidgin develops into a Creole can be modelled by an SRN exposed to a dynamic (substratum-based) environment. In this way, an empiricist approach is able to account for Creole grammar as a by-product of general-purpose learning mechanisms. Connectionist theory, we argue, furnishes us with a (statistical) alternative to nativism. Taking into account that combined Hebbian and activation-phase learning drives SRN networks to a better performance on generalization than the backpropagation algorithm does (O'Reilly & Munakata 2000), a NCC replica of this simulation would further strengthen connectionist's stronghold on the development criterion.

Biologically plausible NCC would cast light as well upon other Newell Test criteria: Real-time Performance, where classical con-

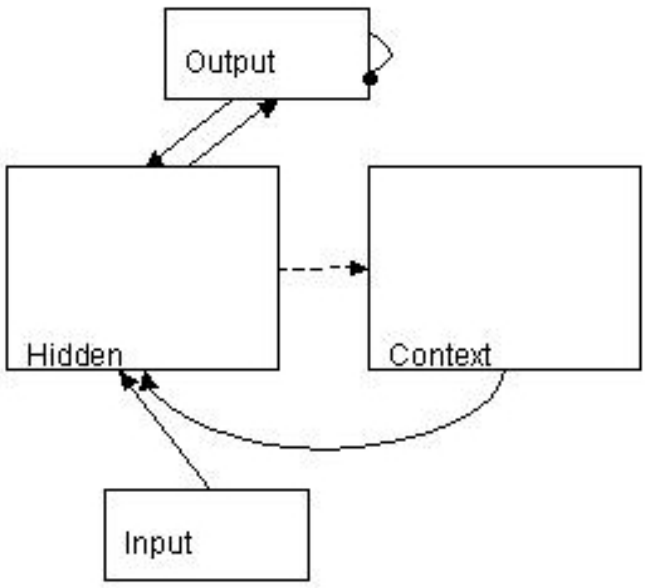

Figure 1 (Garzón). NCC network, with bidirectional connectivity and inhibitory competition, trained on a prediction task.

nectionism gets a "worse" grade, can be improved if we consider online dynamical coupling. In NCC models that do not depend on synchronous updates, it may be assumed, as A&L note, that "the time for a unit to reach full activation is a function of that activation" (sect. 5.2). Moreover, one-shot Hebbian learning (Rolls & Treves 1998), where a few event co-occurrences can contribute to fast recall, can also be seen as a motivation for not having to endorse a hybrid architecture. On the other hand, performance on the flexible behavior criterion would be enhanced as well. Notice that nonclassical, dynamical networks can compute any function to an arbitrary level of accuracy, while allowing for breakdowns in performance.

In general, I can see no good reason not to allow NCC to enter the decathlon. The best connectionist contestant should enter the competition, not a straw man (classical connectionism). It is usually argued that the endorsement of a symbolic-cum-subsymbolic stance would permit connectionism to remain at an appropriate level of realism (Palmer-Brown et al. 2002). However, "failure to acknowledge a symbolic level to thought" (target article, Abstract) cannot be interpreted as a weakness of connectionism when the score is revised as just described.

ACKNOWLEDGMENTS

The author was supported by a Ramón y Cajal research contract (Spanish Ministry of Science and Technology) during the preparation of this commentary. I thank Andy Clark, Eliana Colunga, and Javier Marín for helpful discussions.

Criteria and evaluation of cognitive theories

Petros A. M. Gelepithis

Cognitive Science Laboratory, Kingston University, Kingston upon Thames, KT1 2EE, England. **Petros@kingston.ac.uk**

Abstract: I have three types of interrelated comments. First, on the choice of the proposed criteria, I argue against any *list* and for a *system* of criteria. Second, on grading, I suggest modifications with respect to consciousness and development. Finally, on the choice of "theories" for evaluation, I argue for Edelman's theory of neuronal group selection instead of connectionism (classical or not).

Introduction. Anderson & Lebiere's (A&L's) target article is a useful contribution on the necessity and grading of criteria for a cognitive theory and their application of the Newell Test to classical connectionism and ACT-R a worthwhile exercise. The following comments are partly a criticism on their proposed list of criteria, partly a response to their invitation for modifications of their proposed grading, and partly a critique of their choice of theories for evaluation.

On the choice of criteria for a Theory of Mind (ToM).[1] A&L state that "[t]wice, Newell (1980; 1990) offered slightly different sets of 13 criteria on the human mind" and a bit further down that their table "gives the first 12 criteria from [Newell's] 1980 list, which were basically restated in the 1990 list" (target article, sect. 1: Introduction, 1st para.). Neither of these two statements is correct (as Table 1 confirms).

Furthermore, A&L's list is closer to Newell 1980 than to Newell 1990. No justification for this proximity is provided. Given that Newell's (1990) seminal book is incomparably more comprehensive than his 1980 paper, one wonders about the reasons for A&L's choice. Clearly, their claim of having *distilled* (emphasis added) Newell's two lists (cf. target article, Abstract) cannot be justified either. Although I agree that A&L's list is adequate to avoid "theoretical myopia" (Introduction, 2nd para.), it will create distortions in our quest for a ToM on account of being restricted to a fundamentally impoverished coverage of human phenomena (excluding, e.g., emotion, creativity, social cognition, and culture). It is worth noting that although Newell (1990, sect. 8.4) considered the extension of a unified theory of cognition (UTC) into the social band an important measure of its success, A&L chose to exclude from their list the one constraint with a social element that Newell had included (see item 9 in Table 2).

In contrast, evolution should not be a criterion! Humans are physical objects, but biology is fundamentally different from physics. Similarly, humans are biological systems, but psychology is fundamentally different from biology. The nature of human understanding (Gelepithis 1984; 1991; 1997) transcends the explanatory framework of modern Darwinism and, most importantly, of any future evolutionary theory. (For similar conclusions drawn upon different premises, see Mayr 1988; O'Hear 1997.)

Finally, a fourth list – very different from all previous three – has been offered by Gelepithis (1999). Of the four proposed lists, Table 2 juxtaposes the latest three. The reader can easily spot a number of obvious and significant differences among the three lists. For some of the less obvious, their corresponding serial numbers are in boldface. What all three have in common is that they do not provide necessary and sufficient conditions for a ToM. Still, the mind is a system (Bunge 1980; Hebb 1949; Sherrington 1906). We need, therefore, a *system* (not a list) of criteria characterising mind. A recent promising effort along this route is exemplified by Gelepithis (2002), which presents an *axiomatic system* delineating the class of intelligent systems as a foundation for the development of a ToM[2].

On some "objective measures." *Consciousness.* There are many volumes of readings (e.g., Hameroff et al. 1998; Revonsuo & Kampinnen 1994; Velmans 1996) at least as good as the one cited by A&L. Suggestions of measures on the basis of consciousness-related phenomena in one volume of readings should be avoided. Although universal agreement on what constitutes consciousness is nonexistent, Gelepithis (2001) has provided a list of

Table 1 (Gelepithis). *Extent of the overlap among the proposed sets of criteria by Newell and A&L*

Criteria	Comparisons with Respect to Newell's 1980 List		Comparison with Respect to Newell's 1990 List
	Newell 1990	A&L 2003	A&L 2003
New criteria	2	0	0
Significantly different criteria	3	2	5 or 6
Essentially equivalent criteria	3	3	3 or 2
Identical criteria	5	7	4

Table 2 (Gelepithis). *Three different lists of criteria on human mind.*

	Newell (1990)	Gelepithis (1999)	A&L (2003)
1	Behave flexibly as a function of the environment.		Flexible behaviour (~ Computational Universality).
2	Exhibit adaptive (rational, goal-oriented) behaviour.		Adaptive behaviour.
3	Operate in real time.		Operate in real time.
4	Operate in a rich, complex, detailed environment. Perceive an immense amount of changing detail. Use vast amounts of knowledge. Control a motor system of many degrees of freedom.	Be able to operate in environments of, at least, Earth-level complexity.	Vast knowledge base (sect. 2.4). Dynamic behaviour (sect. 2.5)
5	Use symbols and abstractions.		Knowledge integration.
6	Use language, both natural and artificial.	Acquisition and use of language to, at least, human-level complexity.	Use (natural) language.
7	Learn from the environment and from experience.		Learn from its environment.
8	Acquire capabilities through development.	Explain human neonate's capabilities for development.	Acquire capabilities through development.
9	Operate autonomously, but within a social community.	Operate autonomously, but within a social community.	
10	Be self-aware and have a sense of self.	Be conscious.	Exhibit self-awareness and a sense of self.
11	Be realisable as a neural system.		Be realisable within the brain.
12	Be constructable by an embryological growth process.		
13	Arise through evolution.		Arise through evolution.
14		Use of: (1) domain knowledge and (2) commonsense knowledge for problem solving.	
15		Able to communicate.	
16		Be able to develop skills (e.g., through earning) and judgment (e.g., through maturation).	
17		Develop *own* representational system	
18		Combine perceptual and motor information with *own* belief systems.	
19		Be creative.	
20		Be able to have and exhibit emotions.	

"topics that, *presently,* constitute the major issues in the study of consciousness." I propose that list as a measure.

Development. In view of the suggested grading for consciousness, one might be tempted to propose some or all of the phenomena covered in Johnson et al.'s (2002) reader as a measure for development. Instead, I propose as criterion what is generally agreed to be the fundamental objective in the study of development, namely, "unraveling the *interaction* between genetic specification and environmental influence" (Johnson et al. 2002, p. 3., emphasis added). This fundamental objective in the study of development is shared by most scientists in the field, and it is essentially identical with Piaget's (1967/1971) agenda for developmental psychology. Interestingly, Newell (1990, Ch. 8) has also chosen to talk about development in Piagetian terms.

Choice of "theories" for evaluation. Barring a straightforward case of a Rylean category mistake, A&L seem to believe that there is no difference between theories and a class of models. To put it less strongly, they support the school of thought that argues for theories as families of theoretical models. This is highly debatable in the philosophy of science literature (Giere 1998). Furthermore, taking *theory* in its good old-fashioned meaning, no connectionist (classical or not) model will qualify. In contrast, Edelman's (1989; 1992; Edelman & Tononi 2000) theory of neuronal group selection – based on different foundations[3] – would both have qualified and created a debate on the choice of criteria as well as the types of theories that certain criteria may or may not favour.

To conclude, A&L's concern that connectionists may question the reasonableness of their list is rather well based. Let us not forget that any theory (whether cognitive or otherwise) needs to be founded. Chapter 2 of Newell's (1990) *Unified Theories of Cognition* is an excellent starting point. Comparison between ACT-R's foundations (Anderson 1993; Anderson & Lebiere 1998) and those of SOAR would be revealing; further comparisons of a connectionist (classical or not) theoretical framework and of non-computational ToMs will greatly enhance the foundations of cognitive science and, I would argue, point to the need for a *system* – rather than a *list* – of criteria for Newell's Test.

NOTES

1. I use the terms cognitive theory, unified theories of cognition (UTCs), and ToM interchangeably with respect to their coextensive coverage of human phenomena, and UTC and ToM distinctly with respect to their characteristics.

2. For some interesting earlier results of our approach, the reader is referred to Gelepithis (1991; 1997), Gelepithis and Goodfellow (1992), Gelepithis and Parillon (2002).

3. Evolutionary and neurophysiological findings and principles and the synthetic neural modelling approach to the construction of intelligent entities. For a comparison of four ToMs, see Gelepithis (1999).

Meeting Newell's other challenge: Cognitive architectures as the basis for cognitive engineering

Wayne D. Gray, Michael J. Schoelles, and Christopher W. Myers

Cognitive Science Department, CogWorks Laboratory, Rensselaer Polytechnic Institute, Troy, NY 12180-3590.

{grayw; schoem; myersc}@rpi.edu http://www.rpi.edu/~grayw/ http://www.rpi.edu/~schoem/ http://www.rpi.edu/~myersc/

Abstract: We use the Newell Test as a basis for evaluating ACT-R as an effective architecture for cognitive engineering. Of the 12 functional criteria discussed by Anderson & Lebiere (A&L), we discuss the strengths and weaknesses of ACT-R on the six that we postulate are the most relevant to cognitive engineering.

To mix metaphors, Anderson & Lebiere (A&L) have donned Newell's mantle and picked up his gauntlet. The mantle is Newell's role as cheerleader for the cause of unified architectures of cognition (e.g., Newell 1990). The gauntlet is Newell's challenge to the modeling community to consider the broader issues that face cognitive science. Gauntlets come in pairs, so it is not surprising that Newell threw down another one (Newell & Card 1985), namely, hardening the practice of human factors to make it more like engineering and less based on soft science. (Although Newell and Card framed their arguments in terms of human-computer interaction, their arguments apply to human factors in general and cognitive engineering in particular.)

Cognitive engineering focuses on understanding and predicting how changes in the task environment influence task performance. We postulate that such changes are mediated by adaptations of the mix of cognitive, perceptual, and action operations to the demands of the task environment. These adaptations take place at the embodied cognition level of analysis (Ballard et al. 1997) that emerges at approximately ⅓ second. The evidence we have suggests that this level of analysis yields productive and predictive insights into design issues (e.g., Gray & Boehm-Davis 2000; Gray et al. 1993). However, whatever the eventual evaluation of this approach, our pursuit of it can be framed in terms of six of the Newell Test criteria.

Flexible behavior. We understand A&L to mean that the architecture should be capable of achieving computational universality by working around the limits of its bounded rationality. Hence, not every strategy is equally easy, and not every strategy works well in every task environment. ACT-R fits our cognitive engineering needs on this criterion because it provides a means of investigating, by modeling, how subtle changes in a task environment influence the interaction of perception, action, and cognition to form task strategies.

Real-time performance. When comparing models against human data, a common tack is to simulate the human's software environment to make it easier to run the model. Although such a simulation might represent the essential aspects of the human's task environment, the fidelity of the model's task environment is inevitably decreased. ACT-R enables us to run our models in the same software environment in which we run our subjects by providing time constraints at the time scale that perception, action, and cognition interact.

Adaptive behavior. Section 2.3 of the target article emphasizes Newell's complaint regarding the functionality of then extant theories of short-term memory. In our attempts to build integrated cognitive systems, we too have had similar complaints. For example, the work by Altmann and Gray (Altmann 2002; Altmann & Gray 2002) on task switching was motivated by a failed attempt to use existing theories (e.g., Rogers & Monsell 1995) to understand the role played by task switching in a fast-paced, dynamic environment. Hence, one role of a unified architecture of cognition is that it allows a test of the functionality of its component theories.

Section 5.3 emphasizes the ability to tune models to the "statistical structure of the environment." For cognitive engineering, adaptation includes changes in task performance in response to changes in the task environment, such as when a familiar interface is updated or when additional tasks with new interfaces are introduced. In our experience, ACT-R has some success on the first of these, namely, predicting performance on variations of the same interface (Schoelles 2002; Schoelles & Gray 2003). However, we believe that predicting performance in a multitask environment, perhaps by definition, will require building models of each task. Hence, it is not clear to us whether ACT-R or any other cognitive architecture can meet this critical need of cognitive engineering.

Dynamic behavior. The ability to model performance when the task environment, not the human operator, initiates change is vital for cognitive engineering. We can attest that ACT-R does well in modeling these situations (Ehret et al. 2000; Gray et al. 2000; 2002; Schoelles 2002).

Learning. For many cognitive engineering purposes, learning is less important than the ability to generate a trace of a task analysis of expert or novice performance. With all learning "turned off," ACT-R's emphasis on real-time performance and dynamic behavior makes it well suited for such purposes.

Learning is required to adapt to changes in an existing task environment or to show how a task analysis of novice behavior could, with practice, result in expert behavior. ACT-R's subsymbolic layer has long been capable of tuning a fixed set of production rules to a task environment. However, a viable mechanism for learning new rules had been lacking. With the new production compilation method of Taatgen (see Taatgen & Lee 2003) this situation may have changed.

Consciousness. A&L's discussion of consciousness includes much that cognitive engineering does not need, as well as some that it does. Our focus here is on one aspect: the distinction between implicit and explicit knowledge and the means by which implicit knowledge becomes explicit.

Siegler (Siegler & Lemaire 1997; Siegler & Stern 1998) has demonstrated that the implicit use of a strategy may precede conscious awareness and conscious, goal-directed application of that strategy. ACT-R cannot model such changes because it lacks a mechanism for generating top-down, goal-directed cognition from bottom-up, least-effort-driven adaptations.

Conclusions: Meeting Newell's other challenge. Unified architectures of cognition have an important role to play in meeting Newell's other challenge, namely, creating a rigorous and scientifically based discipline of cognitive engineering. Of the six criteria discussed here, ACT-R scores one best, four better, and one worse, whereas classical connectionism scores two better, two mixed, and two worse. We take this as evidence supporting our choice of ACT-R rather than connectionism as an architecture for cognitive engineering. But, in the same sense that A&L judge that ACT-R has a ways to go to pass the Newell Test, we judge that ACT-R has a ways to go to meet the needs of cognitive engineering. As the Newell Test criteria become better defined, we hope that they encourage ACT-R and other architectures to develop in ways that support cognitive engineering.

ACKNOWLEDGMENTS

Preparation of this commentary was supported by grants from the Office of Naval Research (ONR# N000140310046) as well as by the Air Force Office of Scientific Research (AFOSR# F49620–03–1–0143).

Bring ART into the ACT

Stephen Grossberg

Department of Cognitive and Neural Systems, Boston University, Boston, MA 02215. **steve@bu.edu** **http://www.cns.bu.edu/Profiles/Grossberg**

Abstract: ACT is compared with a particular type of connectionist model that cannot handle symbols and use nonbiological operations which do not learn in real time. This focus continues an unfortunate trend of straw man debates in cognitive science. Adaptive Resonance Theory, or ART, neural models of cognition can handle both symbols and subsymbolic representations, and meets the Newell criteria at least as well as these models.

The authors' use of the nomenclature, "classical connectionist models," falsely suggests that such models satisfy the Newell criteria better than other neural models of cognition. The authors then dichotomize ACT with "classical" connectionism based on its "failure to acknowledge a symbolic level to thought. In contrast, ACT-R includes both symbolic and subsymbolic components" (target article, Abstract). Actually, neural models of cognition such as ART include both types of representation and clarify how they are learned. Moreover, ART was introduced before the "classical" models (Grossberg 1976; 1978a; 1980) and naturally satisfies key Newell criteria. In fact, Figures 2 and 3 of ACT are reminiscent of ART circuits (e.g., Carpenter & Grossberg 1991; Grossberg 1999b). But ART goes further by proposing how laminar neocortical circuits integrate bottom-up, horizontal, and top-down interactions for intelligent computation (Grossberg 1999a; Raizada & Grossberg 2003).

Critiques of classical connectionist models, here called CM (Carnegie Mellon) connectionism, show that many such models cannot exist in the brain (e.g., Grossberg 1988; Grossberg et al. 1997; Grossberg & Merrill 1996). We claim that ART satisfies many Newell criteria better, with the obvious caveat that no model is as yet a complete neural theory of cognition.

Flexible behavior. ART models are self-organizing neural production systems capable of fast, stable, real-time learning about arbitrarily large, unexpectedly changing environments (Carpenter & Grossberg 1991). These properties suit ART for large-scale technological applications, ranging from control of mobile robots, face recognition, remote sensing, medical diagnosis, and electrocardiogram analysis to tool failure monitoring, chemical analysis, circuit design, protein/DNA analysis, musical analysis, and seismic, sonar, and radar recognition, in both software and VLSI microchips (e.g., Carpenter & Milenova 2000; Carpenter et al. 1999; Granger et al. 2001). The criticism of CM connectionism "that complex, sequentially organized, hierarchical behavior" cannot be modeled also does not apply to ART (e.g., Bradski et al. 1994; Cohen & Grossberg 1986; Grossberg 1978a; Grossberg & Kuperstein 1989; Grossberg & Myers 2000; also see the section on dynamic behavior later in this commentary).

Real-time performance. ART models are manifestly real-time in design, unlike CM connectionist models.

Adaptive behavior. ART provides a rigorous solution of the *stability-plasticity dilemma,* which was my term for *catastrophic forgetting* before that phrase was coined. "Limitations like short-term memory" (target article, sect. 5.3) can be derived from the LTM Invariance Principle, which proposes how working memories are designed to enable their stored event sequences to be stably chunked and remembered (Bradski et al. 1994; Grossberg 1978a; 1978b).

Vast knowledge base. ART can directly access the globally best-matching information in its memory, no matter how much it has learned. It includes additional criteria of value and temporal relevance through its embedding in START models that include cognitive-emotional and adaptive timing circuits in addition to cognitive ART circuits (Grossberg & Merrill 1992; 1996).

Dynamic behavior. "Dealing with dynamic behavior requires a theory of perception and action as well as a theory of cognition" (sect. 2.5). LAMINART models propose how ART principles are incorporated into perceptual neocortical circuits and how high-level cognitive constraints can modulate lower perceptual representations through top-down matching and attention (Grossberg 1999a; Raizada & Grossberg 2003). ART deals with novelty through *complementary* interactions between attentional and orienting systems (Grossberg 1999b; 2000b), the former including corticocortical, and the latter, hippocampal, circuits. Action circuits also obey laws that are *complementary* to those used in perception and cognition (Grossberg 2000b), notably VAM (Vector Associative Map) laws. VAM-based models have simulated identified brain cells and circuits and the actions that they control (e.g., Brown et al. 1999; Bullock et al. 1998; Contreras-Vidal et al. 1997; Fiala et al. 1996; Gancarz & Grossberg 1999; Grossberg et al. 1997), including models of motor skill learning and performance (Bullock et al. 1993a; 1993b; Grossberg & Paine 2000).

Knowledge integration. ART reconciles distributed and symbolic representations using its concept of resonance. Individual features are meaningless, just as pixels in a picture are meaningless. A learned category, or symbol, is sensitive to the global patterning of features but cannot represent the *contents* of the experience, including their conscious qualia, because of the very fact that a category is a compressed, or symbolic, representation. Resonance between these two types of information converts the *pattern* of attended features into a coherent context-sensitive state that is linked to its symbol through feedback. This coherent state, which binds distributed features and symbolic categories, can enter consciousness. ART predicts that *all conscious states are resonant states.* In particular, resonance binds spatially distributed features into a synchronous equilibrium or oscillation. Such synchronous states attracted interest after being reported in neurophysiological experiments. They were predicted in the 1970s when ART was introduced (see Grossberg 1999b). Recent neurophysiological experiments have supported other ART predictions (Engel et al. 2001; Pollen 1999; Raizada & Grossberg 2003). Fuzzy ART learns explicitly decodable Fuzzy IF-THEN rules (Carpenter et al. 1992). Thus ART accommodates symbols and rules, as well as subsymbolic distributed computations.

Natural language. ART has not yet modeled language. Rather, it is filling a gap that ACT-R has left open: "ACT-R lacks any theory of the processes of speech perception or speech production" (sect. 4.5, para. 3). ART is clarifying the *perceptual units* of speech perception, word recognition, working memory, and sequential planning chunks on which the brain builds language (e.g., Boardman et al. 1999; Bradski et al. 1994; Grossberg 1978a; 1978b; 1999b; Grossberg et al. 1997a; 1997b; Grossberg & Myers 2000; Grossberg & Stone 1986a; 1986b). Such studies suggest that a radical rethinking of psychological space and time is needed to understand language and to accommodate such radical claims as, "Conscious speech is a resonant wave." ACT-R also does not have "mechanisms . . . [of] perceptual recognition, mental imagery, emotion, and motivation" (sect. 4.5). These are all areas where ART has detailed models (e.g., Grossberg 2000a; 2000c). Speech production uses complementary VAM-like mechanisms (Callan et al. 2000; Guenther 1995). After perceptual units in vision became sufficiently clear, rapid progress ensued at all levels of vision (http://www.cns.bu.edu/Profiles/Grossberg). This should also happen for language.

Development. ART has claimed since 1976 that processes of cortical development in the infant are on a continuum with processes of learning in the adult, a prediction increasing supported recently (e.g., Kandel & O'Dell 1992).

Evolution. "Cognitive plasticity . . . What enables this plasticity in the architecture?" (sect. 5.11). ART clarifies how the ability to

learn quickly and stably throughout life implies cognitive properties like intention, attention, hypothesis testing, and resonance. Although Bayesian properties emerge from ART circuits, ART deals with novel experiences where no priors are defined.

Brain. CM connectionism is said to be "best," although its main algorithms are biologically unrealizable. ART and VAM are realized in verified brain circuits.

It might be prudent to include more ART in ACT. I also recommend eliminating straw man "debates" that do not reflect the true state of knowledge in cognitive science.

ACKNOWLEDGMENTS

Preparation of the this commentary was supported in part by the Air Force Office of Scientific Research (AFOSR F49620–01–1–0397) and the Office of Naval Research (ONR N00014–01–1–0624).

Developing a domain-general framework for cognition: What is the best approach?

James L. McClelland[a], David C. Plaut[a], Stephen J. Gotts[b], and Tiago V. Maia[c]

[a]*Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213;* [b]*Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University, and Laboratory of Neuropsychology, NIMH/NIH, Bethesda, MD 20892;* [c]*Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213.* **jlm@cnbc.cmu.edu http://www.cnbc.cmu.edu/~jlm plaut@cmu.edu http://www.cnbc.cmu.edu/~plaut gotts@nih.gov http://www.cnbc.cmu.edu/~gotts tmaia@cmu.edu http://www.cnbc.cmu.edu/~tmaia**

Abstract: We share with Anderson & Lebiere (A&L) (and with Newell before them) the goal of developing a domain-general framework for modeling cognition, and we take seriously the issue of evaluation criteria. We advocate a more focused approach than the one reflected in Newell's criteria, based on analysis of failures as well as successes of models brought into close contact with experimental data. A&L attribute the shortcomings of our parallel-distributed processing framework to a failure to acknowledge a symbolic level of thought. Our framework does acknowledge a symbolic level, contrary to their claim. What we deny is that the symbolic level is the level at which the principles of cognitive processing should be formulated. Models cast at a symbolic level are sometimes useful as high-level approximations of the underlying mechanisms of thought. The adequacy of this approximation will continue to increase as symbolic modelers continue to incorporate principles of parallel distributed processing.

In their target article, Anderson & Lebiere (A&L) present a set of criteria for evaluating models of cognition, and rate both their own ACT-R framework and what they call "classical connectionism" on the criteria. The Parallel Distributed Processing (PDP) approach, first articulated in the two PDP volumes (Rumelhart et al. 1986) appears to be close to the prototype of what they take to be "classical connectionism." While we cannot claim to speak for others, we hope that our position will be at least largely consistent with that of many others who have adopted connectionist/PDP models in their research.

There are three main points that we would like to make.

1. We share with A&L (and with Newell before them) the effort to develop an overall framework for modeling human cognition, based on a set of domain-general principles of broad applicability across a wide range of specific content areas.

2. We take a slightly different approach from the one that Newell advocated, to pursuing the development of our framework. We think it worthwhile to articulate this approach briefly and to comment on how it contrasts with the approach advocated by Newell and apparently endorsed by A&L.

3. We disagree with A&L's statement that classical connectionism denies a symbolic level of thought. What we deny is only the idea that the symbolic level is the level at which the principles of processing and learning should be formulated. We treat symbolic cognition as an emergent phenomenon that can sometimes be approximated by symbolic models, especially those that incorporate the principles of connectionist models.

In what follows, we elaborate these three points, addressing the first one only briefly since this is a point of agreement between A&L and us.

The search for domain-general principles. There is a long-standing tradition within psychological research to search for general principles that can be used to address all aspects of behavior and cognition. With the emergence of computational approaches in the 1950s and 1960s, and with the triumph of the von Neumann architecture as the basis for artificial computing devices, this search could be formulated as an effort to propose what Newell called "a unified architecture for cognition." An architecture consists of a specification of (1) the nature of the building blocks out of which representations and processes are constructed, (2) the fundamental rules by which the processes operate, and (3) an overall organizational plan that allows the system as a whole to operate. Newell's SOAR architecture and A&L's ACT-R architecture are both good examples of architectures of this type. For our part, we have sought primarily to understand (1) the building blocks and (2) the fundamental rules of processing. Less effort has been devoted to the specifics of the overall organizational plan as such, although we do take a position on some of the principles that the organizational plan instantiates. Because the organization is not fully specified as such, we find it more congenial to describe what we are developing as a framework rather than an architecture. But this is a minor matter; the important point is the shared search for general principles of cognition.

We are of course well aware that this search for general principles runs counter to a strong alternative thread that treats distinct domains of cognition as distinct cognitive modules that operate according to domain-specific principles. Such a view has been articulated for language by Chomsky; for vision, by Marr. Fodor and Keil have argued the more general case, and a great deal of work has been done to try to elucidate the specific principles relevant to a wide range of alternative domains. Although we cannot prove that this approach is misguided, we have the perspective that the underlying machinery and the principles by which it operates are fundamentally the same across all different domains of cognition. While this machinery can be tuned and parameterized for domain-specific uses, understanding the broad principles by which it operates will necessarily be of very broad relevance.

How the search for domain-general principles is carried out. If one's goal is to discover the set of domain-general principles that govern all aspects of human cognition, how best is the search for such principles carried out? Our approach begins with the fundamental assumption that it is not possible to know in advance what the right set of principles are. Instead, something like the following discovery procedure is required.

1. Begin by formulating a putative set of principles.
2. Develop models based on these principles and apply them to particular target domains (i.e., bodies of related empirical phenomena).
3. Assess the adequacy of the models so developed and attempt to understand what really underlies both successes and failures of the models.
4. Use the analysis to refine and elaborate the set of principles, and return to step 2.

In practice this appears to be the approach both of Newell and of A&L. Newell and his associates developed a succession of cognitive architectures, as has Anderson; indeed, Newell suggested that his was only really one attempt, and that others should put forward their own efforts. However, Newell argued for broad application of the framework across all domains of cognition, suggesting that an approximate account within each would be satisfactory. In contrast, we advocate a more focused exploration of a few informative target domains, using failures of proposed models to guide further explorations of how the putative set of principles should be elaborated. To illustrate the power of this approach,

we briefly review two cases. Note that we do not mean to suggest that A&L explicitly advocate the development of approximate accounts. Rather, our point is to bring out the importance of focus in bringing out important principles of cognition.

1. The interactive activation model (McClelland & Rumelhart 1981) explored the idea that context effects in perception of letters – specifically, the advantage for letters in words relative to single letters in isolation – could be attributed to the bidirectional propagation of excitatory and inhibitory signals among simple processing units whose activation corresponds to the combined support for the item the unit represents. When a letter occurs in a word, it and the other letters will jointly activate the unit for the word, and that unit will in turn send additional activation back to each of the letters, thereby increasing the probability of recognition. Similar ideas were later used in the TRACE model of speech perception (McClelland & Elman 1986) to account for lexical influences on phoneme identification. Massaro (1989; Massaro & Cohen 1991) pointed out that the interactive activation model failed to account for the particular quantitative form of the influence of context on the identification of a target item. He argued that the source of the problem lay specifically in the use of bidirectional or interactive activation between phoneme or letter units on the one hand and word units on the other. Since the interactive activation model fit the data pretty well, Newell might have advocated accepting the approximation, and moving on to other issues. However, close investigation of the issue turned out to lead to an important discovery. Subsequent analysis (McClelland 1991; Movellan & McClelland 2001) showed that the failure of the interactive activation model arose from faulty assumptions about the source of variability in performance.

Discovering this was made possible by the failure of the model. It then became possible to consider what changes have to be made in order to fit the data. McClelland (1991) showed that the model had a general deficiency in capturing the joint effects of two different sources of influence even if they were both bottom up and activation was only allowed to propagate in a feedforward direction. The problem was attributed instead to the fact that in the original McClelland and Rumelhart model, the interactive activation process was completely deterministic, and activations were transformed into response probabilities only at the moment of response selection. This led to the discovery of what we take to be an important principle: that the activation process is not only graded and interactive but also intrinsically variable. Reformulated versions of the model incorporating intrinsic variability, in addition to graded representation and interactive processing, were shown through simulations (McClelland 1991) and mathematical analysis (Movellan & McClelland 2001) to produce the right quantitative form of contextual influence on phoneme and letter identification. This principle of intrinsic variability has been incorporated in several subsequent models, including a model that addresses in detail the shapes of reaction time distributions and the effects of a variety of factors on these distributions (Usher & McClelland 2001).

2. Seidenberg and McClelland (1989) introduced a model that accounted for frequency, regularity, and consistency effects in single word reading. The model relied on a single network that mapped distributed input representations of the spellings of words, via one layer of hidden units, onto a set of output units representing the phonemes in the word's pronunciation. However, as two independent critiques pointed out (Besner et al. 1990; Coltheart et al. 1993), the model performed far worse than normal human subjects at reading pronounceable nonwords. Both critiques attributed this shortcoming of the model to the fact that it did not rely on separate lexical and rule-based mechanisms. However, subsequent connectionist research (Plaut et al. 1995; 1996) demonstrated that the particular choice of input and output representations used by Seidenberg and McClelland (1989) was instead the source of the difficulty. These representations tended to disperse the regularity in the mapping from spelling to sound over a number of different processing units. This was because the input units activated by a given letter depended on the surrounding context, and the output units representing a given phoneme were likewise context dependent. Because the learning in the model is in the connections among the units, this led to a dispersion of the information about the regularities across many different connections and created a situation in which letters in nonwords might occur in contexts that had not previously been encountered by the network. This led to the discovery of the principle that to succeed in capturing human levels of generalization performance, the representations used in connectionist networks must condense the regularities. Subsequent models of word reading, inflectional morphology, and other cognitive tasks have used representations that condense the regularities, leading them to achieve human levels of performance with novel items while yet being able to learn to process both regular and exception words.[1]

These two case studies bring out the importance of taking seriously mismatches between a model's behavior and human performance data, even when the model provides an approximate account of most of the relevant phenomena. We believe that such mismatches are important forces in driving the further development of a framework. Of course, such mismatches might also reflect a fundamental inadequacy of the framework as a whole or of its most fundamental grounding assumptions. Analysis is required to determine which; but whatever the outcome, the examination of failures of fit is an important source of constraint on the further development of the framework.

With these comments in mind, we can now turn to the framing of the goals of cognitive modeling as articulated in the sorts of criteria that Newell proposed and A&L have adopted with their own modifications. We agree that it is useful to focus attention on some of these general issues, and that there is more to a good cognitive model than simply a close fit to experimental data. We would note, however, that making the effort at this stage to achieve the sort of breadth that Newell's criteria imply may distract attention from addressing critical discrepancies that can only be revealed through close comparison of models and data. We have chosen to adopt a more focused approach, but we do not deny that a broader approach may reveal other limitations, and that it may be worthwhile for some researchers to follow Newell's strategy.

The importance and nature of the symbolic level. A&L suggest that the shortcomings of the connectionist approach are fundamental, deriving from its failure to acknowledge a symbolic level of thought, whereas the shortcomings of the ACT-R theory are temporary, and derive from its failure as yet to address certain of Newell's criteria. We have a very different reading of the situation.

First of all, our PDP approach does not deny a symbolic level of thought. What we deny is only that the symbolic level is the appropriate level at which the principles of processing and learning should be formulated. We treat symbolic thought as an emergent phenomenon which can sometimes be approximated to a degree by a model formulated at the symbolic level, but which, on close scrutiny, does not conform exactly to the properties that it should have according to symbolic models.

As is well known, the issue here is one that has been extensively explored in the context of research on the formation of past tenses and other inflections of nouns and verbs. A recent exchange of articles contrasts the PDP perspective (McClelland & Patterson 2002a; 2002b) and Pinker's symbolic, dual-mechanism account (Pinker & Ullman, 2002a; 2002b). Here we will present the PDP perspective.

In several places, Pinker and his colleagues have argued that the past tense of English is characterized by two mechanisms, one involving symbolic rules, and the other involving a lexical mechanism that operates according to connectionist principles. A symbolic rule, according to Pinker's approach, is one that applies uniformly to all items that satisfy its conditions. Furthermore, such conditions are abstract and very general. For example, the past-tense rule applies uniformly to any string of phonemes, provided only that it is the stem of a verb. In many places Pinker also states that symbolic rules are acquired suddenly; this conforms to

the idea that a rule is something that one either has or does not have. Finally, the symbolic rule is thought to require a completely different kind of mechanism than the one underlying the inflection of exceptions, leading to the prediction that brain lesions could selectively impair the ability to use the rule while leaving the inflection of irregular forms intact.

Although Pinker and his colleagues have pointed to evidence they believe supports their characterization of the mechanism that produces regular past-tense inflections, in their review of that evidence McClelland and Patterson (2002a) found instead that in every case the evidence supports an alternative characterization, first proposed by Rumelhart and McClelland (1986a), in which the formation of an inflected form arises from the interactions of simple processing units via weighted connections learned gradually from exposure to example forms in the language.[2] First, the evidence indicates that the onset of use of regular forms is gradual (extending over a full year; see Brown 1973; Hoeffner 1996). It is initially restricted to verbs characterized by a set of shared semantic properties, and then gradually spreads to other verbs starting with those sharing some of the semantic properties of the members of the initial set (Shirai & Anderson 1995). Second, usage of the regular past tense by adults is not insensitive to phonology but instead reflects phonological and semantic similarity to known regular verbs (Albright & Hayes 2001; Ramscar 2002). Third, purported dissociations arising from genetic defects (Gopnik & Crago 1991) or strokes (Ullman et al. 1997) disappear when materials are used that control for frequency and phonological complexity (Bird et al. 2003; Vargha-Khadem et al. 1995); individuals with deficits in inflection of regular forms show corresponding deficits with appropriately matched exceptions. In short, the acquisition and adult use of the regular past tense exhibits exactly those characteristics expected from the connectionist formulation. Ultimate adult performance on regular items conforms approximately to the predictions of the rule; for example, reaction time and accuracy inflecting regular forms is relatively insensitive to the word's own frequency. But exactly the same effect also arises in the connectionist models; as they learn from many examples that embody the regular pattern, the connection weights come to reflect it in a way that supports generalization to novel items and makes the number of exposures to the item itself relatively unimportant.

In summary, the characteristics expected on a connectionist approach, but not the symbolic rule approach of Pinker, are exhibited by human performance in forming inflections. Such characteristics include fairly close approximation to what would be expected from use of a symbolic rule under specifiable conditions, but allow for larger discrepancies from what would be predicted from the rule under other conditions (i.e., early in development, after brain damage of particular kinds, and when the language environment is less systematic).[3]

What implications do the characteristics of human performance in forming inflections have for the ACT-R approach of A&L? They have already described an ACT-R model (Taatgen & Anderson 2002) of past-tense formation in which the acquisition of the regular past tense occurs fairly gradually, and we have no doubt that with adjustment of parameters even more gradual acquisition would occur. Furthermore, we see relatively little in A&L's formulation that ties them to the assumption that the conditions for application of symbolic rules must be abstract as Pinker (1991; Pinker & Ullman 2002a) and Marcus (2001) have claimed. Nor is there anything that requires them to posit dissociations, since production rules are used in their model for both regular and exceptional forms. Thus, although the past tense rule actually acquired in the Taatgen and Anderson model is as abstract and general as the one proposed by Pinker, a modified version of their model could surely be constructed, bringing it closer to the connectionist account. To capture the graded and stochastic aspects of human performance, they have introduced graded strengths that are tacked onto symbolic constructs (propositions and productions), thereby allowing them to capture graded familiarity and regularity effects. To capture similarity effects, there is no reason why the condition-matching operation performed by rule-like productions could not be formulated as graded constraints, so that the degree of activation of a production would depend on the degree to which its conditions match current inputs. Indeed, A&L note that by allowing graded condition matching in ACT-R, they can capture the graded, similarity-based aspects of human performance that are naturally captured within the connectionist framework.

Even these adjustments, however, would leave one aspect of connectionist models unimplemented in the Taatgen and Anderson model. This is the ability of connectionist models to exploit multiple influences simultaneously, rather than to depend on the output generated by just one production at a time. Specifically, in the Taatgen and Anderson account of past-tense formation, a past-tense form is generated either by the application of the general *-ed* rule or by the application of an item-specific production; the form that is generated depends on only one of these productions, not on their simultaneous activation. We argue that this is a serious weakness, in that it prevents the Taatgen and Anderson model from exploiting the high degree of conformity with the regular pattern that exists among the exceptions. In our view this is an important and general limitation of many symbolic models, even ones like ACT-R that have moved a long way toward incorporating many of the principles of processing espoused by connectionists.

As McClelland and Patterson (2002b) have noted, fully 59% of the exceptional past-tense verbs in English end in /d/ or /t/. In the connectionist models, the same connection-based knowledge that imposes the regular inflection on fully regular verbs also operates in the inflection of these exceptional cases. That is, the same connections that add /t/ to the regular verb *like* to make *liked* also add /t/ to the irregular verb *keep* to make *kept.* In the case of *kept,* additional influences (from experience with *kept* itself and other similar cases) also operate to allow the model to capture the alteration of the vowel that makes this item an exception. In contrast, in the Taatgen and Anderson model and many other dual-mechanism models, only one production at a time can fire, so that a past-tense form is either generated by the rule (in which case it will be treated as regular) or by a production specific to it as an exception. Given this, no benefit accrues to an exception for sharing properties of the regular past tense, and all exceptions might as well be completely arbitrary. This is problematic because it leaves unexplained important aspects of the distributions of word forms. Across languages, there are many forms that are partially regular and very few that are completely arbitrary, and those that are completely arbitrary are of very high frequency (Plunkett & Marchman 1991); the same is true for irregular spelling-to-sound correspondences. This suggests that human language users are highly sensitive to the degree to which exceptions share properties with regular items, contrary to the properties of the Taatgen and Anderson model.

In response to this, we anticipate that A&L might be tempted to modify the ACT-R framework even further in the direction of connectionist models by allowing application of multiple productions to work together to produce an individual inflected word form. We certainly think this would lead to models that would be more likely than current ACT-R–based accounts to address the influence of regularities in exceptions, and would bring ACT-R more fully into line with the fundamental idea of parallel distributed processing. After all, the essence of PDP is the idea that every act of cognition depends on and is distributed over a large number of contributing units, quite different from what happens presently in ACT-R, where any given output is the product of the application of a single production.

While such a change to ACT-R would, we believe, improve it considerably, we want to simply note two points in this context. First, this would continue the evolution of symbolic models of human cognition even further in a connectionist-like direction. This evolution, which has been in process for some time, is not, in our view, accidental, because with each step in this direction, symbolic

models have achieved a higher degree of fidelity to the actual properties of human cognition. What this indicates to us is that, although the shortcomings of symbolic models may be temporary (as A&L suppose), they are most likely to be overcome by incorporation of the very principles that govern processing as defined at the connectionist level.

Second, as symbolic modelers take each new step in the direction of connectionist models, they do so accepting the fact that the phenomena to be explained have the characteristics that served to motivate the exploration of connectionist models in the first place. This, in turn, undermines the stance that the fundamental principles of human cognition should be formulated at the symbolic level, and instead further motivates the exploration of principles at the connectionist level. While we acknowledge that connectionist models still have many limitations, we nevertheless feel that this does not arise from any failure to acknowledge a symbolic level of thought. Instead we suggest that it arises from the fact the connectionists (like symbolic modelers) have not yet had the chance to address all aspects of cognition or all factors that may affect it.

In spite of our feeling that the facts of human cognition are completely consistent with the principles of parallel distributed processing, we do not wish to give the impression that we see no merit in modeling that is directed at the symbolic level. Given that symbolic formulations often do provide fairly good approximations, it may be useful to employ them in cases where it would be helpful to exploit their greater degree of abstraction and succinctness. We believe that work at a symbolic level will proceed most effectively if it is understood to be approximating an underlying system that is much more parallel and distributed, because at that point insights from work at the connectionist level will flow even more freely into efforts to capture aspects of cognition at the symbolic level.

ACKNOWLEDGMENTS

Preparation of this commentary was supported by Interdisciplinary Behavioral Science Center Grant MH 64445 from the National Institute of Mental Health (USA). Tiago V. Maia was supported by the Foundation for Science and Technology (Portugal). We thank the other members of the PDP Research Group at Carnegie Mellon for useful discussions.

NOTES

1. It is necessary to note that none of the models we have discussed fully embody all the principles of the PDP framework. For example, the interactive activation and TRACE models use localist, not distributed, representations, while the models of spelling-to-sound mapping (Seidenberg & McClelland 1989; Plaut et al. 1996) do not incorporate intrinsic variability. This fact can lead to confusion about whether there is indeed a theoretical commitment to a common set of principles.

In fact, we do have such a commitment. The fact that individual models do not conform to all of the principles is a matter of simplification. This leads to computational tractability and can foster understanding, and we adopt the practices only for these reasons. Everyone should be aware that models that are simplified embodiments of the theory do not demonstrate that models incorporating all of its complexity will be successful. In such cases further research is necessary, especially when the possibility of success is controversial. For example, Joanisse and Seidenberg (1999) used localist word units in their model of past-tense inflection, and Pinker and Ullman (2002a; 2002b) have argued that this is essential. In this context, we fully accept that further work is necessary to demonstrate that a model using distributed semantic representations can actually account for the data.

2. It should be noted here that none of these models assume that learning occurs through correction of overtly generated errors. Instead, it is assumed that exposure provides examples of appropriate usage in context. The learner uses the context as input to generate an internal representation corresponding to the expected phonological form. Learning is driven by the discrepancy between this internal representation and the actual perceived form provided by the example.

3. Marcus et al. (1995) claimed that German has a regular plural (the so-called +s plural) that conforms to the expectation of the symbolic approach, in spite of the fact that it is relatively infrequent. However, subsequent investigations indicate that the +s plural does not exhibit the properties one would expect if it were based on a symbolic rule (Bybee 1995; Hahn & Nakisa 2000).

Evaluating connectionism: A developmental perspective

Claire F. O'Loughlin[a] and Annette Karmiloff-Smith[b]

[a]*Department of Psychology, University of Canterbury, Christchurch, New Zealand;* [b]*Neurocognitive Development Unit, Institute of Child Health, London WC1N 1EH, UK.* **aallardy@xtra.co.nz**
a.karmiloff-smith@ich.ucl.ac.uk

Abstract: This commentary questions the applicability of the Newell Test for evaluating the utility of connectionism. Rather than being a specific theory of cognition (because connectionism can be used to model nativist, behaviorist, or constructivist theories), connectionism, we argue, offers researchers a collection of computational and conceptual tools that are particularly useful for investigating and rendering specific fundamental issues of human development. These benefits of connectionism are not well captured by evaluating it against Newell's criteria for a unified theory of cognition.

In this commentary, we question Anderson & Lebiere's (A&L's) project of grading connectionism according to the Newell Test as an appropriate means of assessing its utility for cognitive science. In our view, connectionism, unlike ACT-R, is not a specific theory of cognition. It can be used to model nativist, behaviourist, or constructivist theories by modifying parameters with respect to built-in representational and architectural or computational structures. Rather, connectionism is a set of computational and conceptual tools that offer researchers new and precise ways of thinking about and investigating complex emergent behaviour. From this standpoint, if we take the view that theory evaluation in science is best conceived as a comparative affair in which mature theories are evaluated along a number of dimensions to determine which provides the best explanation of the phenomena in question (e.g., Lakatos 1970; Thagard 1992), then connectionism does not offer an appropriate theoretical alternative against which to evaluate ACT-R. Moreover, the current appraisal of connectionism against Newell's criteria actually misses many of the positive applications of connectionist tools in cognitive science research. In developmental psychology, for example, this methodological and conceptual toolbox has been put to use in the service of tackling long-standing issues about the mechanisms responsible for developmental change and, more generally, has supported renewed efforts to construct a genuinely interactional account as a theoretical framework for cognitive development (Elman et al. 1996; Karmiloff-Smith 1992; Newcombe 1998). It has also been successfully used to clarify the fundamental differences between adult neuropsychological patients and children with developmental disorders (Karmiloff-Smith 1997; 1998; Karmiloff-Smith et al. 2002; 2003; Thomas & Karmiloff-Smith 2002) and to model how language acquisition can follow atypical developmental trajectories (Thomas & Karmiloff-Smith 2003).

Connectionist models have been shown to be highly relevant to the concerns of developmental researchers, first, because they offer a valuable means of investigating the necessary *conditions* for development. That is, connectionist models provide concrete demonstrations of how the application of simple, low-level learning algorithms operating on local information can, over developmental time, give rise to high-level emergent cognitive outcomes (Elman et al. 1996; Karmiloff-Smith 1992; Karmiloff-Smith et al. 1998; Plunkett et al. 1997). These demonstrations in turn have forced researchers to revisit assumptions about what can actually be learned as opposed to what has to be prespecified, and to recognize that far more structure is latent in the environmental input and capable of being abstracted by basic learning algorithms than previously imagined.

Concerning assumptions about the nature of the starting state in the developing individual, explorations with connectionist models have been pivotal in clarifying the issue of innateness and identifying a range of potential ways in which innate constraints can be realised (Karmiloff-Smith et al. 1998). As Elman et al. (1996) make clear, despite the current dominance of nativist approaches

to the development of language and cognition, scant attention has been given to the issue of biological plausibility in discussions of innate properties, and there has been little investigation of the potential variety of ways in which something could be innate. In contrast, and as a direct result of their experience with connectionist models, Elman et al. (1996) not only present a case against the plausibility of "representational nativism," but also offer a framework for developing alternative conceptions of innate constraints on development that draws on architectural and timing constraints in connectionist models as a guide.

In addition to clarifying the necessary conditions for development, connectionist models also provide a vehicle for exploring the *dynamics* of development. One of the key insights provided by connectionist models is that the mapping between overt behaviour and underlying mechanism is often nonlinear. As Elman et al. (1996) emphasize, contrary to assumptions underpinning much developmental research, qualitative changes in behaviour do not necessarily signal qualitative changes in the mechanisms responsible for that behaviour. Instead, these models demonstrate that sudden dramatic effects in terms of the output of a system can be produced by tiny, incremental changes in internal processing over time. In the case of ontogenetic development, this suggests that apparent discontinuities in conceptual or linguistic understanding or output may not be the result of new mechanisms coming online at certain points in development as has often been assumed, but instead reflect the continuous operation of the same mechanism over time.

Added to demonstrations of how the same mechanism can be responsible for multiple behaviours, connectionist models can also illuminate the reverse case in which a single outcome or behaviour arises through the action of multiple interacting mechanisms. Further, Elman et al. (1996) point to instances where the same behavioural outcome can be produced in a number of different ways, as in the case of degraded performance in artificial neural networks. (See Karmiloff-Smith 1998 for how crucial this is in understanding so-called behaviour in the normal range in some developmental disorders). Precisely because connectionist models allow researchers to probe the potential range of relations that can exist between behavioural outcomes and their underlying causes, they overturn assumptions of straightforward one-to-one mapping between mechanisms and behaviour and are therefore useful in revealing the "multiplicity underlying unity" in development (Elman et al. 1996, p. 363).

The preceding are but a few examples that identify specific issues in developmental psychology where connectionist tools have demonstrated natural applications. More generally, the resources of connectionism have also been a critical factor in recent attempts to develop a viable interactionist framework for cognitive developmental research. Commenting on the connectionist inspired framework advocated by Elman et al. (1996), Newcombe (1998) points to a recent trend in cognitive developmental theorising that eschews the extremes of nativist and empiricist approaches to learning and cognition, in favour of an account that offers some substantive ideas about the reciprocal actions of organism and environment in producing developmental change. From this standpoint, the resources of connectionism can be seen to contribute to this project by offering researchers a specified, formal account of the developmental process that goes well beyond the verbal accounts typical of developmental theory. Moreover, as Elman et al. (1996) point out, the striking resemblance between the process of error reduction in artificial neural networks and earlier attempts to depict epigenesis in natural systems (e.g., Waddington 1975) offers further evidence of the utility of connectionism for attempts to formalize the interactional nature of development.

The preceding sketch serves to highlight some of the variety of ways in which the computational and conceptual resources of connectionism have been usefully applied in developmental psychology. Yet these pragmatic benefits of connectionist models are not readily apparent in A&L's present evaluation of connectionism against the Newell Test designed to reveal an adequate theory of cognition. As it stands, their evaluation falls short of a comprehensive comparative appraisal of ACT-R as a candidate theory of cognition, and it fails to bring forth the utility of the connectionist toolbox for cognitive science research.

On the encompassing of the behaviour of man

Morten Overgaard[a] and Soeren Willert[b]

[a]*Department of Psychology, University of Aarhus, Asylvej 4, Risskov 8240, Denmark;* [b]*Center for System Development, University of Aarhus, Katrinebjergvej 89GAarhus N 8200, Denmark.* **Overgard@pet.au.dk swi@psy.au.dk www.psy.au.dk/phd/morten www.psy.au.dk/ompi/soeren**

Abstract: One supposition underlying the Anderson & Lebiere (A&L) target article is that the maximally broad "encompassing of its subject matter – the behavior of man" (cf. sect. 6, last para.) is regarded as an unquestioned quality criterion for guiding cognitive research. One might argue for an explicit specification of the limitations of a given paradigm, rather than extending it to apply to as many domains as possible.

Anderson & Lebiere (A&L) set out on an important and admirable mission: to evaluate theories within the more or less well-defined area of cognitive science from one set of criteria in order to avoid a dissolving of theories into disconnected paradigms. We shall not criticise their general idea of measuring comparable theories with a common yardstick, nor the actual grading of ACT-R and connectionism presented by A&L. However, the very approach implies that there is a set of theories that can legitimately be labelled "cognitive theories." To decide whether a given theory falls under the category "cognitive science" and thus decide which theories it would be meaningful to grade with the Newell Test, certain basic requirements must be fulfilled. One could ask whether such basic requirements would be identical to the criteria in the A&L version of the Newell Test. If that were indeed the case, we could have no theory that could truly be called *cognitive* to this day. For instance, we have no theory to explain why consciousness is "a functional aspect of cognition" (let alone one that also explains dynamic behaviour, knowledge integration, etc.) (Chalmers 1996; Velmans 1991). Furthermore, it would be a circular enterprise indeed to measure a theory according to criteria identical to the ones it must already fulfil.

Most likely, however, one would not equate the basic requirements for cognitive science with the criteria of the Newell Test. For such a purpose, the criteria seem to be set much too high. Rather, one would look at the many *different* usages of the term *cognitive* within the research field in general and establish relevant criteria on this basis. This, however, leads us into the situation where we presently stand, that is, a situation where "cognitive science" is loosely defined. We have a number of core theories that definitely are cognitive – such as Treisman's attenuation model (Treisman & Gelade 1980) or the SAS model of visual attention (Norman & Shallice 1986) – and several borderline cases – such as Gibson's ecological perception theory (Gibson 1979) – where it is unclear whether the theory is truly a cognitive psychological theory.

Although our conceptualisation of cognitive science does not seem very exact, it seems safe to say that it has developed historically as an attempt to explain the transition from stimulus to response by "internal variables" (see Tolman 1948). Thus, all cognitive theories – the core cases as well as the less clear-cut ones – intend to give explanations in terms of functions. No matter how the specific theories are construed, all cognitive theories explain the function of some mental phenomenon, whether they collect empirical data from behavioural measures, computer simulations, mathematical models, or brain scannings. This common point of departure has certain consequences for the kind of theory that can be developed. First and foremost, any cognitive theory must be

able to model or causally explain observable behaviour. Response times, button presses, verbal reports, and the like, must be the basis of any such theory; without such third-person information, a cognitive science theory would have nothing to explain.

Returning to the problem of consciousness (or the mind-body problem): Why do certain cognitive and emotional processes have specific experiential or so-called qualitative features? Block (1995) has argued for a difference between so-called access-consciousness (A) and phenomenal consciousness (P). A mental state is A-conscious if it can be poised as premise in reasoning, rational control of action and speech. A mental state is P-conscious if there is something it is *like* to be in that state (Nagel 1974). The mind-body problem is, then, normally interpreted as a problem of explaining how P is related to (other) physical matter.

Any cognitive theory should be able to explain or model what happens when subjects report about consciousness, or about anything else, for that matter. In themselves, however, such explanations or modelling exercises do not necessarily point at anything more than correlations between two sets of psychological third-person data, for example, verbal reports and brain activity. At best, this will give us an understanding of A-consciousness, but not necessarily of P. When describing a cognitive process in terms of its functions or causal processes, P does not fit in unproblematically. Even when turning to some of the more optimistic accounts, one finds arguments that cognitive science can *inform* a solving of the mind-body problem but not actually solve it (Overgaard, in press). Epistemologically speaking, one can easily describe one's experiences exactly without ever referring to the kinds of descriptions and models used by cognitive scientists. Vice versa, one can make a full description of a cognitive process in terms of mathematical models or the often-seen "boxes with arrows between them" without ever referring to experiential qualities. On this basis, one might reasonably question whether an explanation of consciousness is a realistic goal for cognitive science.

For this reason, we are sceptical of one basic supposition underlying the A&L target article: that the maximally broad "encompassing of its subject matter – the behavior of man" (Newell 1973, p. 288, cited in sect. 6, Conclusion, last para.) shall be regarded as an unquestioned quality criterion for theoretical models guiding cognitive research. On the contrary, one might argue that it would be a more theoretically sound approach to explicitly specify the limitations of a given paradigm and its possible openness and connectedness with other paradigms, rather than trying to extend it to apply to as many domains as possible.

The one existing type of language in which *everything* can be spoken about is natural, everyday language. The all-encompassing semantic capacity of natural, everyday language is bought at the price of a low degree of specificity as far as the identification of statements' truth conditions is concerned. The potential utility value of theoretical languages lies in their capacity to isolate and specify knowledge domains characterised by high degrees of epistemic consistency (for scientific purposes) and action predictability (for technological purposes). Definitely, at this stage of cognitive science, we fear this utility value may become jeopardised if success in theory building gets simplistically equated with breadth of coverage.

Connectionism, ACT-R, and the principle of self-organization

Pavel N. Prudkov

Ecomon Ltd., Selskohosyastvennaya str 12-A, Moscow, Russia.
Pnprudkov@mtu-net.ru

Abstract: The target article is based upon the principle that complex mental phenomena result from the interactions among some elementary entities. Connectionist nodes and ACT-R's production rules can be considered as such entities. However, before testing against Newell's macro-criteria, self-organizing models must be tested against criteria relating to the properties of their elementary entities. When such micro-criteria are considered, they separate connectionism from ACT-R and the comparison of these theories against Newell's Tests is hardly correct.

The target article by Anderson & Lebiere (A&L) is devoted to the demonstration of the possibilities of the ACT-R theory. To this end, the authors compare their theory against connectionism on the basis of Newell's criteria for a theory of cognition. However, it is difficult to understand from the article why A&L have decided to select connectionism as a competitor of ACT-R. Indeed, if ACT-R is an unified framework, but the term "connectionism" is "used in the field to refer to a wide variety of often incompatible theoretical perspectives" (target article, sect. 3, para. 7), then A&L could test ACT-R against, for example, a bunch of symbolic models sharing certain common characteristics.

It seems that the main reason for A&L's choice (acknowledged by A&L only partially) is the principle of self-organization, that is, the assumption that complex mental phenomena can be described as a result of the interactions among some elementary entities. This principle has been suggested by me elsewhere (cf. Prudkov 1994), and it was based on the following two facts. First, we know that mental processes are heavily connected to various aspects of brain functioning, though the mechanism of this connection is still unclear. Second, neuroscience data demonstrate that the complex forms of brain activity result from the interactions among some elementary brain entities. Brain areas, single neurons, parts of a neuron, distributions of electrical fields, and the like, can be treated as such entities in accordance with the level of brain functioning considered. It seems impossible to reduce all neural levels to a basic one.

The principle of self-organization requires no correspondence between cognitive elementary entities and any of their neural counterparts, though such correspondence is possible. But all characteristics of a cognitive self-organizing process must result from the properties of its elementary entities and interactions among them, without involving any factors external to the system. The architecture of a self-organizing system is defined by three sorts of characteristics (Prudkov 1994). First, it is necessary to define the elementary entities of the system. Second, the results of the interactions between the entities must be determined. Because the idea of interaction supposes changes in components of the entities, one can say self-organizing models by definition are hybrid. And, third, all conditions or probabilities of the interactions to occur must be described. Learning, then, corresponds to long-term changes in a self-organizing system.

With connectionist nodes as elementary entities, it is intuitively clear that connectionism complies with the principle (a more detailed representation is in Prudkov 1994). With the biological implausibility of many connectionist methods, the principle is likely to be the main reason to use connectionism for understanding cognition (Green 1998). To convert the ACT-R theory into self-organization terms, suppose that production rules are elementary entities, matching the conditions of production rules, and the state of declarative memory determines which entities can interact at the given time. Finally, the rule selected for firing, the result of the firing along with the corresponding changes in declarative memory, is the consequence of an interaction.

Of course, this principle must be considered as a heuristic

rather than an established theory. It allows one to construct a wide variety of models and theories, but their efficiency should be tested against various criteria in order to construct adequate models. To some extent, this principle corresponds to the idea that various physical phenomena stem from the interactions among atoms or molecules. Before 1905, when Einstein proved the existence of these particles, this idea was also a heuristic, but its usefulness for physics is obvious.

However, the idea itself is not sufficient to construct physical models, so these interactions must correspond to various physical laws, such as the laws of thermodynamics. In a similar vein, the self-organizing models of cognition initially must be tested against some criteria relating to the properties of its architecture. Such micro-criteria seem absent (or not stated explicitly) in the target article; however, without using them, the comparison against macro-criteria such as Newell's is hardly correct because of the considerable arbitrariness in the models constructed. For instance, different models can merely describe various levels of the phenomenon under study.

Of course, the theory of cognition still does not have such strict laws as in physics, but several micro-criteria appear useful to judge self-organizing models. The first micro-criterion is the similarity in relevant brain functioning. Since self-organizing models of cognition implicitly refer to self-organizing brain activity which can involve various levels of brain functioning, various models can be compared if their architecture meets the same levels of brain functioning. The architecture of connectionism meets the level of single neurons, but the ACT-R architecture corresponds to cortical regions.

The second micro-criterion is the similarity in the determination of initial settings. Various models can be compared when similar efforts are necessary to establish their initial settings and these settings are equally robust to their changes. The robustness of connectionist settings is well known; ACT-R seems to require more precise but vulnerable settings. For example, the ACT-R model of learning the past tense in English (Taatgen & Anderson 2002) performs well, but the model seems to be vulnerable to the choice of the production rules and learning mechanisms used. It is not obvious that the model with slightly different characteristics could show similar results.

The last micro-criterion assumes that the complexity of entities, interactions, and conditions must be approximately the same in the models judged, or the architecture of one model must naturally result from emergent processes in the other. The architecture of connectionist models is simpler than ACT-R's and, realizing this, A&L describe another model, ACT-RN, which implements ACT-R by standard connectionist methods. However, this implementation seems artificial, for A&L simply predetermine the existence of ACT-R's slots and production rules instead of deriving them from more primitive features of a connectionist model. In principle, A&L simply demonstrate that ACT-RN (and, accordingly, ACT-R) meets the principle of self-organization.

One can conclude that three micro-criteria separate connectionism from ACT-R; these theories describe different levels of cognition; therefore, their direct comparison is hardly correct.

Dual-process theories and hybrid systems

Ilkka Pyysiäinen

Helsinki Collegium for Advanced Studies, University of Helsinki, FIN-00014, Finland. **ilkka.pyysiainen@helsinki.fi**
http://www.helsinki.fi/collegium/eng/staff.htm

Abstract: The distinction between such differing approaches to cognition as connectionism and rule-based models is paralleled by a distinction between two basic modes of cognition postulated in the so-called dual-process theories. Integrating these theories with insights from hybrid systems might help solve the dilemma of combining the demands of evolutionary plausibility and computational universality. No single approach alone can achieve this.

Not only are cognitive scientific "paradigms" disconnected; it also seems to be difficult for a theory of cognition to meet both Newell's criteria 1 and 11. An evolved cognitive architecture apparently cannot be computationally universal (e.g., Bringsjord 2001). Anderson & Lebiere (A&L) thus emphasize that humans can learn to perform almost arbitrary cognitive tasks, but they do not explain why some tasks are easier to learn than others. They suggest that applying a broad enough range of criteria might help us construct an exhaustive theory of cognition, referring to Sun's (1994; 2002) hybrid systems integrating connectionism and a rule-based approach as an example (see also Sun & Bookman 1995). I argue that the distinction between connectionist and functionalist models is paralleled by a distinction between two types of actual cognitive processing, as postulated within the so-called dual-process theories. These theories, developed in social psychology, personality psychology, and neuropsychology, for example, strongly suggest that there are two different ways of processing information, variously labeled

Intuition and implicit learning versus deliberative, analytic strategy (Lieberman 2000)

A reflexive and a reflective system (Lieberman et al. 2002)

Associative versus rule-based systems (Sloman 1996; 1999)

An experiential or intuitive versus a rational mode of thinking (Denes-Raj & Epstein 1994; Epstein & Pacini 1999; Epstein et al. 1992; Simon et al. 1997)

An effortless processing mode that works through associative retrieval or pattern completion in the slow-learning system elicited by a salient cue versus a more laborious processing mode that involves the intentional retrieval of explicit, symbolically represented rules from either of the two memory systems to guide processing (Smith & DeCoster 2000)

Implicit versus explicit cognition (Holyoak & Spellman 1993)

Intuitive versus reflective beliefs (Cosmides & Tooby 2000a; Sperber 1997)

Although the terminologies vary, there is considerable overlap in the substance of these distinctions. The two systems serve different functions and are applied to differing problem domains. They also have different rules of operation, correlate with different kinds of experiences, and are carried out by different brain systems. Some consider these two mechanisms as endpoints on a continuum, whereas Lieberman et al. (2002) argue that they are autonomous systems (see, e.g., Chaiken & Trope 1999; Holyoak & Spellman 1993).

By synthesizing the extant theories, with a special focus on Sloman (1996) and Lieberman et al. (2002), we may characterize the spontaneous system as follows. It operates reflexively, draws inferences, and makes predictions on the basis of temporal relations and similarity; and employs knowledge derived from personal experience, concrete and generic concepts, images, stereotypes, feature sets, associative relations, similarity-based generalization, and automatic processing. It serves such cognitive functions as intuition, fantasy, creativity, imagination, visual recognition, and associative memory (see especially, Sloman 1996). It involves such brain areas as the lateral temporal cortex, amygdala, and basal ganglia. The lateral temporal cortex is, for example, most directly in-

volved in the construction of attributions, whereas the amygdala and basal ganglia are responsible for trying to predict possible punishments and rewards related to one's actions (Lieberman et al. 2002; cf. Rolls 2000).

This system consists of a set of neural mechanisms tuned by a person's past experience and current goals; it is a subsymbolic, pattern-matching system that employs parallel distributed processing. It produces that continuous stream of consciousness we experience as "the world out there," whereas the rational system reacts to the spontaneous system, producing conscious thoughts experienced as reflections on the stream of consciousness (Lieberman et al. 2002). As a pattern-recognition system, the spontaneous system tries to combine all perceived features into a coherent representation; this is because the relevant neurons have been so paired by past experience that the activation of some will also activate others. The spontaneous system cannot consider the causal or conditional relationships between percepts because it does not operate by symbolic logic and because its links are bidirectional. Thus, simply asking a dispositional question (e.g., "Is this man prone to violent behavior?") may easily lead to an affirmative answer (Lieberman et al. 2002).

The rational system involves such brain areas as the anterior cingulate, prefrontal cortex, and hippocampus (Lieberman et al. 2002). It is a rule-based system able to encode any information that has a well-specified formal structure. Such a structure also allows the generation of new propositions on the basis of systematic inferences carried out in a language of thought which has a combinatorial syntax and semantics. It explicitly follows rules. This system thus seeks for logical, hierarchical, and causal-mechanical structure in its environment; operates on symbol manipulation; and derives knowledge from language, culture, and formal systems. It employs concrete, generic, and abstract concepts; abstracted features; compositional symbols; as well as causal, logical, and hierarchical relations. It is productive and systematic; abstracts relevant features; is strategic, not automatic; and serves such cognitive functions as deliberation, explanation, formal analysis, verification, ascription of purpose, and strategic memory (Sloman 1996).

The rational system either generates solutions to problems encountered by the spontaneous system, or it biases its processing in a variety of ways. A pre-existing doubt concerning the veracity of one's own inferences seems to be necessary for the activation of the rational system. The rational system thus identifies problems arising in the spontaneous system, takes control away from it, and remembers situations in which such control was previously required. These operations consist of generating and maintaining symbols in working memory, combining these symbols with rule-based logical schemes, and biasing the spontaneous system and motor systems to behave accordingly (Lieberman et al. 2002).

It could thus be argued that the spontaneous system is a collection of evolved mechanisms with an adaptive background, whereas computational universality is based on the ability of the rational system to exploit the evolved mechanisms to create algorithms for the performance of any cognitive task (see Pinker 1997, pp. 358–359; Atran 2002). This explains the fact that in many areas of everyday life people rely both on evolutionary intuitions and explicit theories. This distinction has recently been studied with regard to peoples' religious intuitions and their theological theories (e.g., Barrett 1998; 1999; Barrett & Keil 1996; Boyer 2001; Pyysiäinen 2003; Whitehouse 2002). Interaction between work on these types of real-life problem fields and on construction of hybrid systems might help us develop more integrated theories of human cognition.

ACKNOWLEDGMENTS

I thank Matthew Lieberman, Marjaana Lindeman, and Markku Niemivirta for help in writing this commentary.

The hardest test for a theory of cognition: The input test

Asim Roy

School of Information Systems, Arizona State University, Tempe, AZ 85287-3606. **asim.roy@asu.edu**

Abstract: This commentary defines an additional characteristic of human learning. The nature of this test is different from the ones by Newell: This is a hard, pass/fail type of test. Thus a theory of cognition cannot partially satisfy this test; it either conforms to the requirement fully, or it doesn't. If a theory of cognition cannot satisfy this property of human learning, then the theory is not valid at all.

The target article by Anderson & Lebiere (A&L) is very refreshing in the sense that it turns the focus back on accountability and tests for any theory of cognition. In examining theories of cognition, a look at system identification in science and engineering may be in order. In system identification, the basic idea is to construct an equivalent system (model) that can produce "behavior" that is similar to the actual system. So the key idea is to produce "matching external behavior." The equivalent system may not necessarily match the internal details of the system to be identified, but that is fine as long as it matches the system's external properties. And the external properties to match may be many. This is not to say that one should not take advantage of any information about the internals of the system.

Therefore, the crucial task for this science is to define the external behavioral characteristics that any system of cognition is supposed to exhibit. Understanding and characterizing the phenomenon to be modeled and explained is clearly the first and main step towards developing a theory for it. If that is not done, it is very likely that wrong theories will be proposed, because it is not known exactly what the theory should account for. This commentary defines an additional characteristic of human learning other than the ones in the Newell Test (Roy et al. 1997). In the spirit of the Newell Test, this is a characteristic of the brain that is "independent of" (1) any conjectures about the "internal" mechanisms of the brain (theories of cognition) and (2) the specific learning task. That is, this property of human learning is independent of a specific learning task like learning a language, mathematics, or a motor skill. The nature of this test is quite different from the ones provided by Newell: This is a hard, pass/fail type of test. In that sense, a theory of cognition cannot partially satisfy this test; it either conforms to its requirement fully, or it doesn't. This pass/fail test would allow one to quickly check the validity of alternative theories of cognition. If a theory of cognition cannot satisfy this property of human learning, then the theory is not valid at all. So this particular test is good enough for initial screening of theories. As explained in the following paragraphs, classical connectionism fails this test. One has to take a closer look at ACT-R and ACT-RN to pass judgment on them.

So what is this real hard test for theories of cognition? It can be summarized as follows: A brain-like system, constructed on the basis of some theory of cognition, is not permitted to use any inputs in its construction phase that are not normally supplied to a human brain. So the real hard test for any theory is in the inputs required to construct the relevant system of cognition. Let this test be called the "Input Test." The human brain has two sources of inputs during its development, both inside the womb and outside. Biological parents are the first source, and certain structures and systems can be inherited through that source. The other source of inputs for its development is the environment after birth. So any theory of cognition has to clearly delineate what pieces of its functioning system are inherited from biological parents and what pieces are developed subsequently through interactions with the environment. For humans, it is known for a fact that certain functionality of the brain is definitely not inherited, like the ability to speak a certain language, do mathematics, and so on. The modules for these functionalities/tasks do not come pre-built in the hu-

man brain; rather, they are developed and constructed gradually over time. So, to reiterate this point, the first task of a theory of cognition is to clearly delineate what pieces of its functioning system are inherited and what pieces are developed subsequently through interactions with the environment. And with regard to what can come pre-built (inherited), it has to provide sensible arguments.

Once a proposed theory of cognition maps out what is pre-built in the system in the sense of being inherited from biological parents, then the problem for the theory is to show how it develops and constructs the modules that are not pre-built. And whatever the means are for developing and constructing these modules, the hardest test for the theory is this: It has to demonstrate that it is not using any inputs for developing and constructing these modules that are not provided to humans from the environment. This input test can be explained nicely by examining classical connectionism. In classical connectionism, for example, network designs and other algorithmic information have to be externally supplied to the learning system, whereas no such information is ever an external input to the human brain. The well-known back-propagation algorithm of Rumelhart et al. (1986) is a case in point. In fact, many different network designs and other parameter values often have to be supplied to these learning systems on a trial-and-error basis in order for them to learn. However, as far as is known, no one has ever been able to externally supply any network designs or learning parameters to a human brain. So classical connectionism clearly violates this input test and is not a valid theory of cognition.

In general, for previously unknown tasks, the networks could not feasibly come predesigned in human brains; thus network designs cannot be inherited for every possible unknown learning problem faced by the brain on a regular basis. And the networks required for different tasks are different; it is not a one-size-fits-all situation. Since no information about the design of a network is ever supplied to the brain externally, it therefore implies that the brain performs network designs internally. Thus, it is expected that any theory of cognition must also demonstrate the same ability to design networks and adjust its own learning parameters without any outside intervention. But the connectionist learning systems can't demonstrate this capability, and that again implies that classical connectionism is not a valid theory of cognition.

In summary, in this input test, a theory of cognition should be restricted to accepting information that is normally supplied to a human from the environment, nothing more.

Rethinking learning and development in the Newell Test

Sylvain Sirois

Department of Psychology, The University of Manchester, Manchester M13 9PL, United Kingdom. **sylvain.sirois@man.ac.uk**
http://www.psy.man.ac.uk/staff/sirois.htm

Abstract: The Newell Test is an ambitious and promising project, but not without pitfalls. Some of the current criteria are not theoretically neutral, whereas others are unhelpful. To improve the test, the learning and development criteria are reviewed and revised, which suggests adding a maturation criterion as well. Such changes should make the Newell Test more general and useful.

Anderson & Lebiere (A&L) have certainly embarked on an ambitious project: to transform Newell's (1980; 1990) functional criteria for human cognitive architectures into the ultimate test of cognitive theories. I certainly sympathise with such ambitions, especially given their emphasis on the functional aspects of the criteria that should be used. For example, we recently conducted a similar (albeit substantially more humble) exercise for models of infant habituation (Sirois & Mareschal 2002). We identified a set of seven behavioural and neural criteria that functional models of the phenomena need to satisfy. This proved extremely useful to highlight the limitations of current models, but also (and perhaps more importantly) to suggest what the next generation of models needed to address. Given the relatively small scale of the problem addressed in our work, one could conceivably expect huge and varied rewards from A&L's far more valiant endeavour.

Whereas the rewards may prove an exponential function of those we observe in analogous but restricted projects, so may the problems. The authors quite rightly acknowledge that their criteria (which are a slightly modified version of Newell's) are not the only criteria by which a theory can be assessed. But far more crucial than how many criteria (which makes the test more or less liberal) is the question of *which* criteria (which makes the test more or less useful). If the stated goal of such a test is to avoid theoretical myopia, then a few of the criteria are certainly problematic because they either imply that a model adheres to a specific school of thought or to tests of models against highly disputable standards. For example, *knowledge integration* may have been retitled from Newell (1990) but owes no less to symbolic tradition than when it was proposed by Newell. As such, the grading of this criterion is unduly biased towards models and theories originating from this tradition. The *consciousness* criterion is even more problematic: Whether the criterion has any functional value depends on an eventual theory that would make such a suggestion!

Other commentators will likely address the relevance or appropriateness of the various criteria, if not of the test itself. Despite inherent difficulties in such projects, I believe that a revised formulation of the Newell Test could be quite useful. I would thus like to focus on two criteria that, in my view, should be kept in the Newell Test: *learning* and *development.* Surprisingly, the authors evacuated the functional role of learning in their discussion. Moreover, they discuss development as a (perhaps functional) constraint rather than as a functional mechanism. In fact, what they present as development sounds remarkably like maturation.

The authors should not be blamed too harshly for reproducing a common problem in developmental psychology: confounding learning and development by discussing them in terms of *outcomes* rather than *mechanisms* (Liben 1987). This is most explicit when they present the slow learning of *classical connectionism* as satisfying the development criterion. If anything, and contrary to what the authors suggested, the sort of learning in classical connectionism can be characterised as a nativist learning theory (Quartz 1993; Sirois & Shultz 1999).

Fortunately, the notions of learning and development can be expressed formally as non-overlapping functions (Sirois & Shultz 1999). *Learning* can be defined as parametric changes that enable a given processing structure to adapt to its environment. *Development,* however, can be defined as structural changes that foster more complex adaptations, given learning failure. These definitions not only constrain the contribution of each mechanism to cognitive change, but also specify the relationship between learning and development. Learning causes the current structure to adapt, but when that fails, development alters the structure to promote further learning. It must be noted that either form of change is a function of experience. Within this framework, then, *maturation* becomes an experience-independent structural change that delays learning, in line with what A&L discussed as development.

Like others (including A&L), I believe that an adequate theoretical formulation of cognition must be consistent with learning and developmental issues. Moreover, given the significant changes that can be introduced by maturation (i.e., the cognitive structure increases in complexity), I would suggest that the Newell Test also incorporates maturation as one of its criteria. The grading is relatively straightforward for the learning, development, and maturation criteria. If a theory allows for parametric changes as a function of experience, it can learn. If it allows for experience-dependent structural changes that support further learning, it satisfies development. Finally, if it allows for experience-independent, programmed structural changes that modify the learning space, it satisfies maturation.

These learning, development, and maturation criteria are general by design, and so are the grading proposals, in line with Newell's wish to avoid theoretical myopia. A cognitive theory should be granted with the ability to satisfy any of these criteria if it satisfies the relevant functional properties, irrespective of how the mechanisms are actually realised. This general nature does not imply that the criteria are vague, however. We initially proposed these definitions to discuss various classes of neural networks as they are applied to developmental problems. We found that the classical connectionist framework only satisfied the learning criteria (Sirois & Shultz 1999). But we applied the same framework to discuss the various mechanisms of Piagetian theory, clarifying them in the process, and allowing for a formal distinction between learning and developmental notions in Piaget's work (Sirois & Shultz 2003). If we apply these definitions to ACT-R as discussed by A&L, we could grant ACT-R with the ability to satisfy learning and developmental criteria (the latter through the construction of new rules).

To summarise, the idea of a Newell Test is quite attractive but not without design pitfalls. Whereas there may be some inadvertent myopia in the choice of criteria, most of these may well be retained (but perhaps reformulated). The peer commentaries in this journal will hopefully provide the next few steps towards the design of a generally satisfying test of cognitive theories.

ACKNOWLEDGMENT

I thank Isabelle Blanchette for useful comments on an earlier draft.

What about embodiment?

David Spurrett

Philosophy Department, University of Natal, Durban, 4041, South Africa.
spurrett@nu.ac.za **http://www.nu.ac.za/undphil/spurrett/**

Abstract: I present reasons for adding an *embodiment* criterion to the list defended by Anderson & Lebiere (A&L). I also entertain a likely objection contending that embodiment is merely a type of *dynamic behavior* and is therefore covered by the target article. In either case, it turns out that neither connectionism nor ACT-R do particularly well when it comes to embodiment.

The principle that cognitive theories should be evaluated according to multiple criteria is worth adopting, and Anderson & Lebiere's (A&L's) development of Newell's proposals in this regard is useful. One important criterion seems to be missing, though, and that is *embodiment*.

By embodiment, I understand, loosely, physical implementation in an environment. Humans, clearly a key consideration of the target article, are, of course, embodied. They exhibit striking virtuosity at moving around the world and exploiting the resources available in it. Perhaps more important for present purposes, we are talented at exploiting the structure of environments (and of our bodies in them) for cognitive ends, or as some would have it, engaging in "distributed cognition" (e.g. Hutchins 1995). One example is locomotion, where recent research (Thelen & Smith 1994) indicates that the architecture of the body, and the properties of the body in interaction with the environment, play significant roles in control of behavior. Another example, rather closer to the concerns of traditional cognitive science, is the game of Tetris, where it has been shown (Kirsh & Maglio 1994) that human players use external actions to improve the efficiency (speed, accuracy, error rate) of the spatial manipulations and judgements demanded by the game. External rotation of a Tetris piece, along with inspection to establish whether the rotated piece is in a preferable orientation (compared to before), is often faster and less error-prone than mental rotation for the same purpose. This suggests that at least some cognitive problems are tackled using a coalition of internal and external resources, and that an important feature of our cognitive makeup is that we can detect opportunities for this. (Further examples in humans, other animals, and (some) robots abound. Clark [1997] is a useful survey.) This in turn indicates that a theory of cognition that fails to take embodiment seriously is unlikely to capture such features of our own cognitive performance.

A likely objection here notes that A&L's criterion 5 is "dynamic behavior." Since this criterion concerns the relationship between a cognitive system and an environment, perhaps, properly understood, it includes embodiment and distributed cognition. Distributed cognition just *is,* the objection goes – a kind of dynamical coupling between an information-processing system and a structured body and environment. This objection may be taking charitable interpretation too far. A&L's discussion of their "dynamic behavior" criterion (sect. 2.5 of the target article) places considerable emphasis on dealing with the unexpected, and relatively less on exploiting external structure. When evaluating the relative performance of classical connectionism and ACT-R with respect to the dynamic behavior criterion (sect. 5.5 of the target article), their emphasis is on real-time control, not embodiment. Rather than try to settle the question whether embodiment is or is not a version of dynamic behavior, I propose to consider how connectionism and ACT-R fare in the case where embodiment is added as a separate criterion, and where dynamic behavior is interpreted to include it.

Were embodiment added as a criterion, I suggest that connectionism would achieve mixed results. In some cases it does extraordinarily well. Consider Quinn and Espenschied's (1993) neural network for controlling a hexapod robot. The success of this system depends to a significant extent on allowing features of the physical construction of the robot, in interaction with the environment, to play a role in control – so that the motion of individual feet will be inhibited if other specific feet do not yet have secure positions. One way of understanding this is to regard the changing physical links between some neurons, parts of the robot anatomy, the physical environment, other parts of the anatomy and (eventually, and sometimes) other neurons, as functioning like additional neurons, or interneuron connections, transforming or transmitting information about footing, load on joints, and so on. In other cases, though, it is not (yet) clear how to go about building a network, embodied or otherwise, to handle tasks (such as air traffic control) involving fairly specific and detailed functional decomposition, tasks for which systems such as ACT-R seem well suited.

ACT-R, I argue, scores worse for embodiment. Its successes at, for example, modelling driving are in constrained simulation environments, where embodied interaction with the "feel" of the vehicle and its relation to the road surface, are absent, and where attendant opportunities for exploiting environmental structure (engine tone, vibration) to help cue such actions as gear changes are absent for both the human subjects who provide the target data, and the ACT-R models of driving behavior which do well at approximating the behavior of such humans.

However, we might reinterpret A&L's "dynamical behavior" criterion in a way that includes embodiment as a subtype of dynamic behavior. In this case, and in the light of what is said in the target article and so far in this commentary, connectionism should retain its mixed score. In this case ACT-R should also, I argue, receive a mixed score: It doesn't do well at plain embodiment, but does better at non-embodied forms of dynamic behavior. In either case, the moral to draw is that if embodiment is a genuinely important criterion, then *neither* connectionism nor ACT-R seem, as they stand, in a good position to perform consistently well on it.

Conceptions and misconceptions of connectionism

Ron Sun

CECS Department, University of Missouri-Columbia, Columbia, MO 65211.
rsun@cecs.missouri.edu **http://www.cecs.missouri.edu/~rsun**

Abstract: This commentary examines one aspect of the target article – the comparison of ACT-R with connectionist models. It argues that conceptions of connectionist models should be broadened to cover the whole spectrum of work in this area, especially the so-called hybrid models. Doing so may change drastically ratings of connectionist models, and consequently shed better light on the developing field of cognitive architectures.

John Anderson has been one of the pioneers of cognitive architectures. His and Christian Lebiere's work on ACT-R has been highly influential. In many ways, their work defines this field today.

However, instead of going on praising ACT-R, I shall here focus on shortcomings of the target article. One shortcoming, as I see it, is in Anderson & Lebiere's (A&L's) treatment of connectionist models or, more precisely, in their very conception of connectionist models. In the target article, as a comparison to ACT-R, A&L focus exclusively on what they term "classical connectionism" (which I would call "strong connectionism") – the most narrowly conceived view of connectionist models, from the mid-1980s, as articulated by the classic PDP book (Rumelhart & McClelland 1986). In this view, connectionist models are the ones with regular network topology, simple activation functions, and local weight-tuning rules. A&L claim that this view "reflects both the core and the bulk of existing neural network models while presenting a coherent computational specification" (target article, sect. 3, last para.).

However, it appears that connectionist models conforming to this view have some fundamental shortcomings. For example, the limitations due to the regularity of network topology led to difficulty in representing and interpreting symbolic structures (despite some limited successes so far). Other limitations are due to learning algorithms used by such models, which led to lengthy training (with many repeated trials), requiring a priori input/output mappings, and so on. They are also limited in terms of biological relevance. These models may bear only remote resemblance to biological processes.

In coping with these difficulties, two forms of connectionism became rather separate: Strong connectionism adheres closely to the above strict precepts of connectionism (even though they may be unnecessarily restrictive), whereas weak connectionism (or hybrid connectionism) seeks to incorporate both symbolic and subsymbolic processes – reaping the benefit of connectionism while avoiding its shortcomings. There have been many theoretical and practical arguments for hybrid connectionism (see, e.g., Sun 1994). Considering our lack of sufficient neurobiological understanding at present, a dogmatic view on the "neural plausibility" of hybrid connectionist models is not warranted. It appears to me (and to many other people) that the death knell of strong connectionism has already been sounded, and it is time now for a more open-minded framework without the strait-jacket of strong connectionism.

Hybrid connectionist models have, in fact, been under development since the late 1980s. Initially, they were not tied into work on cognitive architectures. The interaction came about through some focused research funding programs by funding agencies. Several significant hybrid cognitive architectures have been developed (see, e.g., Shastri et al. 2002; Sun 2002; Sun et al. 2001).

What does this argument about the conception (definition) of connectionism have to do with ratings on the Newell Test? In my own estimate, it should affect ratings on the following items: "a vast amount of knowledge," "operating in real time," "computational universality," "integrating diverse knowledge," and possibly other items as well. Let's look into "a vast amount of knowledge," as an example. What may prevent neural networks from scaling up and using a vast amount of knowledge is mainly the well-known problem of catastrophic interference in these networks. However, the problem of scaling and "catastrophic interference" in neural networks may in fact be resolved by modular neural networks, especially when symbolic methods are introduced to help partition tasks (Sun 2002). With different subtasks assigned to different networks that are organized in a modular fashion, catastrophic interference can be avoidable. Thus, if we extend the definition of connectionist models, we can find some (partial) solutions to this problem, which are (at least) as good as what is being offered by ACT-R to the same problem. Similar things may be said about "integrating diverse knowledge" or "operating in real time," and so on. Overall, when our conceptions of connectionist models are properly expanded, our ratings of connectionist models will have to be changed accordingly too; hence the significance of this issue to the target article.

A related shortcoming of the target article is the lack of adequate discussion and rating of hybrid connectionist models besides ACT-R. Ratings of these models and comparisons with ACT-R can shed further light on the strengths and weaknesses of different approaches. There have been some detailed analyses and categorizations of hybrid connectionist models, which include "classical" connectionist models as a subset, that one might want to look into if one is interested in this area (see, e.g., Sun & Bookman 1994; Wermter & Sun 2000).

Finally, I would like to echo the authors' closing remarks in the conclusion (sect. 6) of the article: If researchers of all theoretical persuasions try to pursue a broad range of criteria, the disputes among theoretical positions might simply dissolve. I am confident that the target article (and more importantly, this entire treatment) may in fact contribute toward this end.

ACKNOWLEDGMENT

This work was supported in part by ARI contract DASW01–00–K–0012.

Poppering the Newell Test

Niels A. Taatgen

Department of Artificial Intelligence, University of Groningen, 9712 TS Groningen, The Netherlands. **niels@ai.rug.nl**
http://www.ai.rug.nl/~niels

Abstract: The Newell Test as it is proposed by Anderson & Lebiere (A&L) has the disadvantage of being too positivistic, stressing areas a theory should cover, instead of attempting to exclude false predictions. Nevertheless, Newell's list can be used as the basis for a more stringent test with a stress on the falsifiability of the theory.

The idea of the Newell Test is obviously inspired by its illustrious predecessor, the Turing Test (Turing 1950) and can be considered as an elaboration of the topics that have to be addressed by a theory to make it a plausible basis for an intelligent machine. There is a subtle difference between the two tests: Although the Turing Test stresses the fact that the computer should be able to make meaningful conversation, the main point is that the judge in the Turing Test is supposed to do everything possible to expose the computer as a fraud. This aspect of the test is very important, because noncritical discussion partners of the computer can easily be fooled by programs like ELIZA (Weizenbaum 1966; also see Lodge 1984) and its successors. Analogous to the Turing Test, the Newell Test has two aspects: a positivistic aspect (i.e., the theory should allow models of all areas of cognition) and a falsifiability aspect (i.e., the theory should restrict and eventually disallow all "false" models) (Popper 1963). The latter aspect, however, has much less prominence in the Newell Test than the former. I would like to criticize this and argue that the aspect of excluding false models is at least as important, and maybe much more important, than permitting true models.

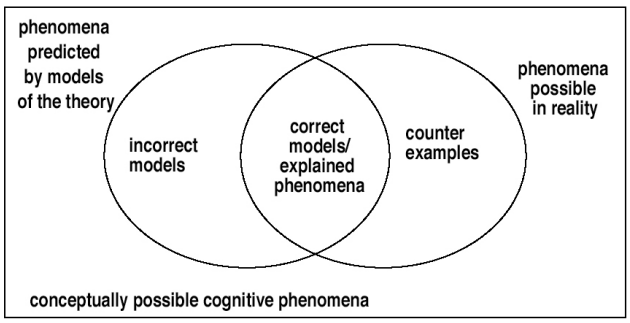

Figure 1 (Taatgen). Diagram to illustrate successes and problems of a theory of cognition.

Figure 1 illustrates the issue. Consider the set of all possibly conceivable cognitive phenomena, of which only a subset contains phenomena that can actually occur in reality. Then the goal of a theory is to predict which of the conceivable phenomena are actually possible, and the success of a theory depends on the overlap between prediction and reality. The problems of a theory can be found in two categories: counterexamples, phenomena that are possible in reality but are not predicted by the theory, and incorrect models, predictions of the theory that are not possible in reality. The issue of incorrect models is especially important, because an unrestricted Turing Machine is potentially capable of predicting any conceivable cognitive phenomenon. One way to make the Newell Test more precise would be to stress the falsifiability aspects for each of the items on the test. For some items this is already more or less true in the way they are formulated by Anderson & Lebiere (A&L), but others can be strengthened, for example:

Flexible behavior. Humans are capable of performing some complex tasks after limited instructions, but other tasks first require a period of training. The theory should be able to make this distinction as well and predict whether humans can perform the task right away or not.

Real-time performance. The theory should be able to predict human real-time performance, but should not be able to predict anything else. Many theories have parameters that allow scaling the time predictions. The more these parameters are present, the weaker is the theory. Also the knowledge (or network layout) that produces the behavior can be manipulated to adjust time predictions. Restricting the options for manipulation strengthens the theory.

Knowledge integration. One property of what A&L call "intellectual combination" is that there are huge individual differences. This gives rise to the question how the theory should cope with individual differences: Are there certain parameters that can be set that correspond to certain individual differences (e.g., Lovett et al. 1997; Taatgen 2002), or is it mainly a difference in the knowledge people have? Probably both aspects play a role, but it is of chief importance that the theory should both predict the breadth and depth of human behavior (and not more).

Use natural language. The theory should be able to use natural language but should also be able to assert what things cannot be found in a natural language. For example, the ACT-R model of learning the past tense shows that ACT-R would not allow an inflectional system in which high-frequency words are regular and low-frequency words are irregular.

Learning. For any item of knowledge needed to perform some behavior, the theory should be able to specify how that item has been learned, either as part of learning within the task, or by showing why it can be considered as knowledge that everyone has. By demanding this constraint on models within a theory, models that have unlearnable knowledge can be rejected. Also, the learning system should not be able to learn knowledge that people cannot learn.

Development. For any item of knowledge that is not specific to a certain task, the theory should be able to specify how that item of knowledge has been learned, or to supply evidence that that item of knowledge is innate. This constraint is a more general version of the learning constraint. It applies to general strategies like problem solving by analogy, perceptual strategies, memorization strategies, and the like.

Another aspect that is of importance for a good theory of cognition is parsimony. This is not an item on Newell's list, because it is not directly tied to the issue of cognition, but it was an important aspect of Newell's research agenda. This criterion means that we need the right number of memory systems, representations, processing, and learning mechanisms in the theory, but not more. An advantage of parsimony is that is makes a stronger theory. For example, SOAR has only one learning mechanism, chunking. This means that all human learning that you want to explain with SOAR has to be achieved through chunking, as opposed to ACT-R, which has several learning mechanisms. Of course, SOAR's single mechanism may eventually be found lacking if it cannot account for all human learning.

To conclude, research in cognitive modeling has always had a positivistic flavor, mainly because it is already very hard to come up with working models of human intelligence in the first place. But as cognitive theories gain in power, we also have to face the other side of the coin: to make sure that our theories rule out wrong models. This is not only an issue for philosophers of science, but a major issue if we want to apply our theories in human-computer interaction and education. There, it is of vital importance that we should be able to construct models that can provide reliable predictions of behavior without having to test them first.

Cognitive architectures have limited explanatory power

Prasad Tadepalli

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331-3202. **tadepall@cs.orst.edu**
http://www.eecs.orst.edu/~tadepall

Abstract: Cognitive architectures, like programming languages, make commitments only at the implementation level and have limited explanatory power. Their universality implies that it is hard, if not impossible, to justify them in detail from finite quantities of data. It is more fruitful to focus on particular tasks such as language understanding and propose testable theories at the computational and algorithmic levels.

Anderson & Lebiere (A&L) undertake the daunting task of evaluating cognitive architectures with the goal of identifying their strengths and weaknesses. The authors are right about the risks of proposing a psychological theory based on a single evaluation criterion. What if the several micro-theories proposed to meet different criteria do not fit together in a coherent fashion? What if a theory proposed for language understanding and inference is not consistent with the theory for language learning or development? What if a theory for playing chess does not respect the known computational limits of the brain? The answer, according to Newell, andA&L, is to evaluate a cognitive theory along multiple criteria such as flexibility of behavior, learning, evolution, knowledge integration, brain realization, and so forth. By bringing in multiple sources of evidence in evaluating a single theory, one is protected from *overfitting,* a problem that occurs when the theory has too many degrees of freedom relative to the available data. Although it is noncontroversial when applied to testable hypotheses, I believe that this research strategy does not work quite as well in evaluating cognitive architectures.

Science progresses by proposing testable theories and testing them. The problem with cognitive architectures is that they are not theories themselves but high-level languages used to imple-

ment theories, with only some weak architectural constraints. Moreover, these languages are computationally universal and thus are equivalent to one another in the sense that one language can simulate the other. How does one evaluate or falsify such universal languages? Are the multiple criteria listed by the authors sufficient to rule out anything at all, or do they simply suggest areas to improve on? The authors' grading scheme is telling in this respect. It only evaluates how an architecture satisfies one criterion better than another criterion, and does not say how to choose between two architectures. One cannot, of course, duck the question merely by choosing an architecture based on the criterion one is interested in explaining. This is precisely the original problem that Newell was trying to address through his multiple criteria.

The authors suggest that timing constraints and memory limitations imply that one cannot only program arbitrary models in ACT-R. But that still leaves room for an infinite variety of models, and ACT-R cannot tell us how to choose between them. To take an analogy to programming languages: it is possible to design an infinite variety of cognitive architectures and implement an infinite variety of models in each one. Can we ever collect enough evidence to be able to choose one over another?

This suggests to me that a cognitive theory must be carefully distinguished from the concrete implementation and the underlying architecture. Just as a programming language can implement any given algorithm, a cognitive architecture can instantiate any cognitive theory (albeit with some variations in time efficiencies). This should not count as evidence for the validity of the architecture itself, any more than good performance of an algorithm should count as evidence for the validity of the programming language. Cognitive science can make better progress by carefully distinguishing the algorithm from the architecture and confining the claims to those parts of the algorithm that are in fact responsible for the results. Consider, for example, ACT-R's theory of past-tense learning by children. More specifically, consider the empirical observation that the exceptions tend to be high-frequency words. A&L attribute this to the fact that only high-frequency words develop enough base-level activation to be retrieved in ACT-R. In more general terms, only high-frequency words provide sufficient training data for the system to be able to learn an exception. How much of this explanation is a result of the particulars of ACT-R theory as opposed to being a necessary consequence of learning in a noisy domain? If any learning system that operates in a noisy environment needs more training data to learn an exception, why should this be counted as evidence for the ACT-R theory? Similar criticisms can be leveled against other cognitive architectures and mechanisms such as SOAR and chunking, connectionism and backprop.

In other words, even when multiple criteria are used to evaluate a cognitive architecture, there still remains an explanatory gap (or a leap of faith) between the evidence presented and the paradigm used to explain it. To guard against such over-interpretation of the evidence, Ohlsson and Jewett propose "abstract computational models," which are computational models that are designed to test a particular hypothesis without taking a stand on all the details of a cognitive architecture (Ohlsson & Jewett 1997). Similar concerns are expressed by Pat Langley, who argues that the source of explanatory power often lies not in the particular cognitive architecture being advanced but in some other fact such as the choice of features or the problem formulation (Langley 1999). Putting it another way, there are multiple levels of explanations for a phenomenon such as past-tense learning or categorization, including computational theory level, algorithmic level, and implementation level. Computational theory level is concerned with *what* is to be computed, whereas algorithmic level is concerned with *how* (Marr 1982). Cognitive architecture belongs to the implementation level, which is below the algorithmic level. Where the explanatory power of an implementation mostly lies is an open question.

Only by paying careful attention to the different levels of explanations and evaluating them appropriately can we discern the truth. One place to begin is to propose specific hypotheses about the algorithmic structure of the task at hand and evaluate them using a variety of sources of evidence. This may, however, mean that we have to put aside the problem of evaluating cognitive architectures, for now or forever.

ACKNOWLEDGMENTS

I thank Sridhar Mahadevan and Pat Langley for influencing my thinking on this matter and for their comments on an earlier draft.

Cognitive modelling of human temporal reasoning

Alice G. B. ter Meulen

Center for Language and Cognition, University of Groningen, 9700 AS Groningen, The Netherlands. **atm@let.rug.nl** **http://atm.nemit.net**

Abstract: Modelling human reasoning characterizes the fundamental human cognitive capacity to describe our past experience and use it to form expectations as well as plan and direct our future actions. Natural language semantics analyzes dynamic forms of reasoning in which the real-time order determines the temporal relations between the described events, when reported with telic simple past-tense clauses. It provides models of human reasoning that could supplement ACT-R models.

Real-time performance, the second criterion for a human cognitive architecture in Newell (1990), requires the system to operate as fast (or as slow) as humans (target article, sect. 2, Table 1) on any cognitive task. Real time is hence considered a constraint on learning as well as on performance (sect. 5). Although I certainly consider it an advantage of the ACT-R system that it does not rely on artificial assumptions about presentation frequency in the way classical connectionist systems do (Taatgen & Anderson 2002), the limited focus the two systems share on the acquisition of the morphological variability in the simple past-tense inflection in English ignores its obvious common semantic properties, which also must be learned. In this commentary, I propose to include in real-time performance the characteristic human ability to use time effectively when using language to encode information that systematically depends on contextual parameters, such as order of presentation or time of utterance.

Human linguistic competence includes automated processes of temporal reasoning and understanding, evidence of which is presented in our linguistic intuitions regarding the temporal relations that obtain between events described in coherent discourse. The presentation order in which simple past-tense clauses are produced in real time often contains important clues for the correct interpretation. As opposed to the past progressive (*John was leaving*) and the past perfect (*John had left*), the English simple past tense (*John left*) refers to an event that not only precedes the time of utterance but also is temporally located with respect to other events described by prior discourse. The following examples, (1) and (2), show that the order of presentation affects our understanding of what happened.

(1) *John lit a cigarette. He left.*
(2) *John left. He lit a cigarette.*

From (1) we understand that John left after he had lit a cigarette. But (2) makes us understand that the described events occurred in the opposite order. Obviously, the real-time order of presentation in this case determines the temporal relations between the events described. But this is not always so, as we see from examples (3) and (4), where reversing the order of the simple past-tense clauses does not affect the temporal relations between the events.

(3) *John slept for hours. He dreamt of Mary.*
(4) *John dreamt of Mary. He slept for hours.*

Either (3) or (4) makes us understand that John dreamt of Mary while he slept, which is reinforced by the lexical presupposition of dreaming requiring that the dreamer be asleep.

The differences observed between the interpretations of (1)–(4), coincidentally all morphologically strong past-tense inflections, are attributed to the aspectual class of the clauses, which may be telic or atelic (Partee 1984; Hinrichs 1986). Although the compositional characterization of telicity has been a core item on the linguistic research agenda for quite some time, it is generally agreed that in English, clauses that may be modified by durative adverbials, such as *for hours,* are atelic, and clauses that are unacceptable with durative modifiers are telic (ter Meulen 1995; Verkuyl 1996). Temporal precedence effects, which conceptually shift the reference time, are determined by order of presentation of telic clauses in simple past-tense clauses.

Children gradually learn to produce cohesive discourse with simple past-tense clauses, effectively using order of presentation, instead of connecting clauses in their stories with *and the . . . and then . . . É.* It depends on their understanding of logical or causal relations between lexical items; for example, dreaming entails sleeping, leaving entails moving elsewhere. It also requires mastering deductive or abductive forms of reasoning, into which neither classical connectionism nor ACT-R have many modelling insights to offer, as Anderson & Lebiere (A&L) readily admit. Reasoning in context and exploiting the dependencies between tense and other indexical features of linguistic expressions cannot be reduced to conditioned correlations between lexical items and concepts, as classical connectionists may want to argue, because it needs a representation of the agent's own information structured information state, as well as a representation of the external domain described by linguistic input and other agents it communicates with. Human understanding of information communicated in ordinary language discourse should, therefore, constitute a core task on the common agenda of cognitive science, testing not only Newell's criteria of real-time performance and natural language, but also adaptive, dynamic, and flexible behavior, as well as knowledge integration and development. Natural language semantics is studying the structured dependencies between context, information, and described domain (Asher et al. 1994; van Eijck & Kamp 1997; ter Meulen 2000). The "Dynamic Turn" in the semantics of both formal-logical, and natural languages has profoundly changed the agenda of the traditional logical systems to require that a dynamic semantics of natural language ideally provides abstract models of our human cognitive capacities of information processing, envisaged in Partee (1980; 1997;) as the program to "naturalize formal semantics." ACT-R accounts of human cognition may well find it a congenial companion, supplementing its self-proclaimed need for an account of human reasoning.

Real-world behavior as a constraint on the cognitive architecture: Comparing ACT-R and DAC in the Newell Test

Paul F. M. J. Verschure

Institute of Neuroinformatics, University Zürich–Swiss Federal Institute of Technology (ETH), Zürich, 8057, Switzerland. **pfmjv@ini.phys.ethz.ch**
http://www.ini. ethz.ch/~pfmjv

Abstract: The Newell Test is an important step in advancing our understanding of cognition. One critical constraint is missing from this test: A cognitive architecture must be self-contained. ACT-R and connectionism fail on this account. I present an alternative proposal, called Distributed Adaptive Control (DAC), and expose it to the Newell Test with the goal of achieving a clearer specification of the different constraints and their relationships, as proposed by Anderson & Lebiere (A&L).

Anderson & Lebiere (A&L) make the important step to resurrect a number of benchmarks, originally proposed by Newell, which a theory of cognition should satisfy. One benchmark that is missing from this list is that the proposed architecture must be self-contained. *Self-contained* implies that the knowledge of the cognitive system is acquired through an autonomous learning process; that is, its ontology is derived from the interaction between the system and the world. Both ACT-R and classical connectionism do not score well on this constraint. ACT-R fails because it focuses on the use of predefined knowledge in its productions and its recombination by means of chunking. The implementation of its memory structures using artificial neural networks and the inclusion of a subsymbolic/symbolic nomenclature does not address this problem. Classical connectionism fails because it relies on learning rules, for example, backpropagation, that allow the user to compile a predefined input-output mapping into the model (Verschure 1990; 1992). In both cases the models do not tell us how knowledge is acquired in the first place. One could argue that solving this problem of priors is the most fundamental challenge to any candidate theory of cognition (Verschure 1998).

In order to challenge the authors to define more precisely what it takes to satisfy the Newell Test, I present an alternative proposal for a cognitive architecture, called Distributed Adaptive Control (DAC). DAC describes an embodied cognitive architecture implemented by a neuronal system in the context of real-time, real-world behavior. DAC assumes that behavior is organized around three tightly coupled layers of control: reactive, adaptive, and contextual (Fig. 1A). The typical paradigms in which we have developed this architecture are robot equivalents of random foraging tasks (Fig. 1B). It should be emphasized that DAC develops its own domain ontology out of its continuous interaction with the world. Hence, as opposed to ACT-R, DAC is self-contained.

Flexible behavior ("better"). DAC has been shown to organize landmark-based foraging behavior in different types of robots (Verschure et al. 1992; 1996; Verschure & Voegtlin 1998), has been applied to simple games such as tic-tac-toe (Bouvet 2001), has controlled a large scale public exhibit (Eng et al. 2003), and has been shown to be equivalent to an optimal Bayesian interpretation of goal-oriented problem solving (Verschure & Althaus 2003). By satisfying this last constraint, DAC implicitly addresses a wide range of cognitive phenomena (Massaro 1998). This latter constraint argues that our models should attack abstract models describing large repertoires of performance as opposed to single instances of particular behaviors.

Real-time performance ("better"). As opposed to ACT-R, DAC takes real time literally as the time it takes to control real-world behavior. In biologically detailed models, derived from the DAC architecture, of both the sensory (i.e., the learning-dependent changes in receptive field properties of the primary auditory cortex, as reported by Kilgard & Merzenich 1998) and motor aspects (focusing on the cerebellum) of classical conditioning, we have shown that these principles can account for learning performance both in terms of number of trials and in terms of the relevant real-time interstimulus intervals (Sanchez-Montanez et al. 2002; Hofstötter et al. 2002). Hence, these models generalize the hypothesis of DAC towards the neuronal substrate and can account for properties of performance in terms of the underlying neuronal mechanisms. Important here is that temporal properties of behavior are not redescribed in functional terms, which is an under-constrained problem, but directly interpreted in terms of neuronal mechanisms. This illustrates that the benchmarks cannot be interpreted as independent constraints.

Adaptive behavior ("best"). The DAC architecture has been designed in the context of real-world embodied cognition (see also *flexible behavior*). The claim is that only such an approach can account for this constraint. ACT-R is not embodied.

Vast knowledge base (mixed). DAC shows how task-dependent knowledge can be acquired and used to organize behavior and has been applied to a range of tasks (see *flexible behavior*). However, the full neuronal implementation of its structures for short- and long-term memory is not mature enough to make strong statements on its capacity and flexibility (Voegtlin & Verschure 1999). Hence, DAC takes satisfying neuronal constraints as a fundamental benchmark in answering functional challenges. ACT-R seems to stop at a functional interpretation.

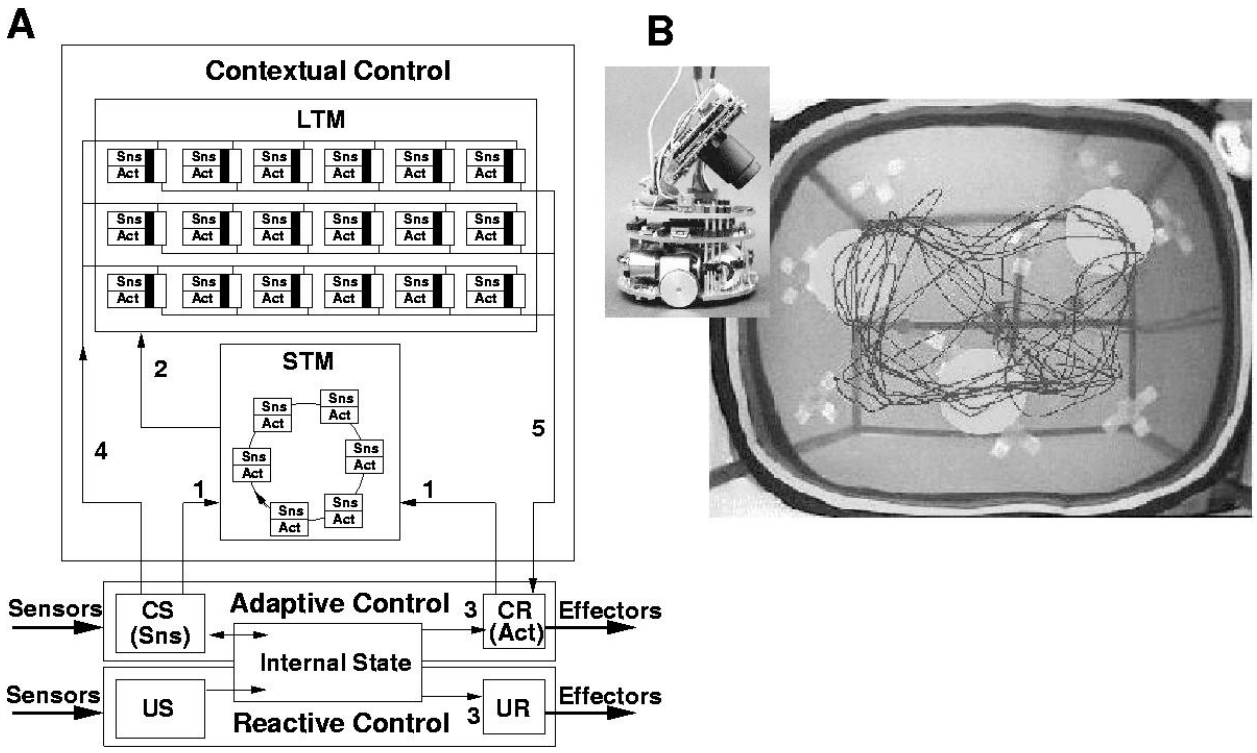

Figure 1 (Verschure). **A.** The DAC architecture. **B.** One example of the application of DAC to robot random foraging using a Khepera micro-robot (K-team, Lausanne).

Dynamic behavior ("best"). DAC has been applied to real-world tasks that include goal conflicts, changing motivational states, and dynamically changing environments, for example, the large-scale exhibition Ada (see *flexible behavior*). In contrast, ACT-R has only been tested on closed problem domains and has not considered the motivational components underlying the organization of dynamic behavior.

Knowledge integration ("better"). DAC has been shown to both acquire the content of its memory structures and to form goal-related recombinations of these representations. Given its Bayesian equivalence, DAC satisfies properties of inference making and induction. However, what is required is a more explicit specification of the experimental data that should be accounted for.

Natural language ("worse"). DAC has not been applied to any form of language acquisition or expression. However, DAC claims that its general learning properties will generalize to language; that is, an explanation of language should emerge from the general principles that underlie the organization of adaptive behavior and not require yet another a priori functional module. In contrast, ACT-R appears to develop in terms of a collection of functionally distinct and independent modules.

Consciousness ("worse"). For now, there is no ambition in the DAC project to attack this phenomenon.

Learning ("best"). DAC was initially conceived to address the behavioral paradigms of classical and operant conditioning. These forms of learning, as opposed to the ones the authors focus on, deal with the problem of autonomous acquisition and expression of knowledge. The biologically detailed models derived from DAC, described above, for instance, account for the phenomenon of blocking central to the Rescorla-Wagner rule of classical conditioning in terms of neuronal mechanisms and not only in functional terms (Hofstötter et al. 2002). This again emphasizes that functional and structural constraints must be satisfied simultaneously and that constraints should be defined around general models, such as the Rescorla-Wagner laws. Moreover, this approach illustrates that a theory of a cognitive architecture will probably be accompanied with a large set of specific derived models that validate a specific subset of its assumptions.

Development ("better"). The DAC architecture interprets development as the progressive involvement of its adaptive and contextual control layers. We have shown that this progression can display stage transitions characteristic for cognitive development (Verschure & Voegtlin 1998). However, the authors should be more precise in specifying what the exact datasets are that should be explained to satisfy this benchmark.

Evolution ("mixed"). Following classic examples of, for example, Pavlov (1928), DAC assumes that cognition arises out of a multilayered architecture that requires a minimum of prior specification. Because the phenomenon of classical conditioning has also been observed in insects (Menzel & Muller 1996), we are currently investigating whether the DAC principles do generalize to insects. Hence, although the results are not in, the claim is that phylogenetic continuity of principles underlying cognition should be evaluated following this comparative approach.

Brain ("better"). As mentioned earlier, the basic principles underlying the adaptive and reactive layers of DAC have been implemented and tested using biophysically and anatomically constrained models. Although the contextual layer makes predictions about the functional properties of neuronal organization, in particular, in relation to the hippocampus, basal ganglia, and prefrontal cortex, these predictions still need to be verified by developing biologically constrained models of these structures. ACT-R seems to stop at finding a correlation between neuronal responses obtained with fMRI measurements and its functional decomposition of cognition. This might not be sufficient. A&L should be congratulated for proposing a common test for theories of cognition and exposing ACT-R to it. The Newell Test in its current form, however, is not mature enough to use it as a gold standard for theories of cognition. This step should be taken in order to advance our understanding of mind, brain, and behavior.

In Figure 1, panel A, the reactive control layer provides a be-

having system with a prewired repertoire of reflexes (unconditioned stimuli and responses – US, UR) that enable it to interact with its environment and accomplish simple automatic behaviors. The activation of any reflex, however, also provides cues for learning that are used by the adaptive control layer via representations of internal states. Adaptive control provides the mechanisms for the adaptive classification of sensory events (conditioned stimulus – CS) and the reshaping of responses (conditioned responses – CR) supporting simple tasks, and can be seen as a model of classical conditioning. The sensory and motor representations formed at the level of adaptive control provide the inputs to the contextual control layer that acquires, retains, and expresses sequential representations using systems for short- and long-term memory. The contextual layer describes goal-oriented learning as observed in operant conditioning. Central-processing steps at this level in the architecture are the following: (1) The representations of sensory cues (Sns) and associated motor states (Act) acquired by the adaptive layer are stored in short-term memory (STM) as a segment. (2) If a goal state is reached, that is, a target found or a collision suffered, the contents of STM are retained in long-term memory (LTM) as a sequence. Each segment of LTM consists of a sensori-motor representation (Sns, Act) a trigger unit (black) and a collector unit (white). (3) The reactive and adaptive control layers can still trigger actions and stand in a competitive relation to the contextual control system. (4) Each Sns state of the adaptive layer is matched against those stored in LTM. (5) The collector units of LTM can trigger actions dependent on the biased competition between LTM segments. By modulating dynamic thresholds of each LTM segment, different chaining rules can be implemented.

In panel B of Figure 1, the robot learns to use the color information in the environment, the patches on the floor and the walls, in order to acquire the shortest route between goal locations, that is, light sources (grey circles). The trajectory visualized is generated during a recall task where the light sources are switched off, after learning for about 30 min. The environment measures about 1.5 by 0.8 m; and the robot, about 55 by 30 mm.

A multilevel approach to modeling human cognition

Hongbin Wang,[1] Todd R. Johnson, and Jiajie Zhang

School of Health Information Sciences, University of Texas Health Science Center at Houston, Houston, TX 77030. **hongbin.wang@uth.tmc.edu**
todd.r.johnson@uth.tmc.edu **jiajie.zhang@uth.tmc.edu**
http://www.shis.uth.tmc.edu

Abstract: Although we agree with Newell and Anderson & Lebiere (A&L) that a unified theory of cognition is needed to advance cognitive science, we disagree on how to achieve it. A hybrid system can score high in the Newell Test but may not offer a veridical and coherent theory of cognition. A multilevel approach, involving theories at both psychological and brain levels, is suggested.

Newell certainly had a very good reason for being frustrated over the progress towards a scientific understanding of the human mind. The human mind is undoubtedly one of most complex entities in the world. It is systematically shaped by genetic and evolutionary forces; fundamentally constrained by physical and biochemical laws; influenced by cultural, social, and environmental factors; and manifests itself both psychologically and neurophysiologically. Given its inherent complexity and our limited knowledge in each of these aspects, it is conceivable that we may not be able to achieve a thorough understanding of the mind's work for a long time.

While we share Newell's frustration, we doubt that the Newell Test, as proposed in the target article, would offer us relief. On the one hand, the attainability of the test is theoretically questionable. It remains controversial, for example, whether self-awareness and consciousness are computationally implementable (e.g., Penrose 1989; 1996; 1997). This controversy helps to explain why both connectionism and ACT-R were graded "worse" on criterion 8 (self-awareness and consciousness) in the target article. On the other hand, even if we ignore the possible theoretical difficulties, we may still encounter practical problems in developing theories of mind that can pass the test, as we elaborate later.

After evaluating connectionism and ACT-R based on the Newell Test and suggesting that neither was satisfactory on all criteria, the authors Anderson & Lebiere (A&L) go on to recommend some remedies. One major remedy suggested is that we should somehow dissolve the distinctions and join the two approaches close together. Specifically, ACT-R needs to be "more compatible with connectionism," and connectionism needs to be concerned "with more complex tasks and symbolic processing" (sect. 6, para. 3). The authors note that building hybrid systems that can integrate the two approaches is particularly promising (ACT-R itself is already a form of hybrid system). By combining the advantages of different sub-approaches, the authors seem to suggest that hybrid systems would bring us one step closer to a Theory of Mind (ToM) that can score high in the Newell Test.

Unfortunately, there are at least three problems with this hybrid system approach. First, it should be noted that there are two (out of 12) criteria on which both connectionism and ACT-R score worse or worst. They are criterion 8 (self-awareness and consciousness) and criterion 11 (evolution). The simultaneous failure of both approaches on both criteria suggests that simply hybridizing the two approaches might not provide a solution.

Second, what if we develop a theory of self-awareness and an evolutionary ToM, and then hybridize these two theories with the hybrid system we constructed earlier? Does this give us a better ToM? Well, maybe. If doable, it will certainly boost the Newell Test score! But it also induces a paradox. Focusing on isolated and segmented subtheories of mind is what frustrated Newell and motivated the creation of the Newell criteria in the first place. If we first need to develop subtheories to develop high-scoring hybrid systems, we then lose the very point of the Newell Test.

Third, and most important, hybrid systems are artificially assembled systems and thus bear little true psychological and neurophysiological significance. Although we all agree that the human mind is a complex, multilevel construct and involves mechanisms and operations at, among others, both psychological and neuronal networks levels, simply piecing them together is ad hoc and trivializes the problem. A ToM that explains one phenomenon using a neural-network-level mechanism and explains another phenomenon using a rule-based, symbolic-level mechanism may be a convenient hybrid ToM, but is certainly not the *unified* ToM that Newell had wished for (cf. Newell 1990).

In our view, any principled ToM must recognize that the human mind may adopt different mechanisms and follow different laws at different levels. In addition, it is highly unlikely that there exists any simple and linear one-to-one mapping across levels. Penrose, for example, went so far as to hypothesize that there is a noncomputational and nonlocal process called "objective reduction" that connects physics and consciousness (see also Woolf & Hameroff 2001). We would argue that a similar nonlinear relationship exists between the neuronal-network-level and the psychological level, and that each level tells a veracious but adequately distinct story of mind. Such a multilevel view is also consistent with both Marr's (1982) and Newell's (1990) conception of multiple-level description of human cognition. Consequently, we should not expect a single architecture, even a hybrid one, to explain all of the phenomena of mind.

We regard both ACT-R and connectionism as celebratory candidates for a ToM, but at different levels. Whereas ACT-R focuses on the symbolic mental structures and processes and offers a psychologically plausible explanation that closely links to empirical behaviors, connectionism adopts subsymbolic neural-based mechanisms and permits a biologically realistic explanation

of mind that closely links to brain functions (e.g., O'Reilly & Munakata 2000). The two approaches are distinct in that symbols simply do not exist at a subsymbolic level. A unified ToM needs to encompass both levels of description, though each may be embodied in separate cognitive architectures. We regard the attempt to vertically stretch one level of analysis to linearly map to another as problematic. For example, we doubt that there is such a simple one-to-one mapping between ACT-R components and brain structures, as suggested in Figure 1 of the target article. It is hard to imagine (and not supported by neuroscience evidence) that the damage to the basal ganglia would completely destroy the work of mind given the fundamental role that production rules play in ACT-R.

In summary, although we agree with Newell and A&L that a unified ToM is needed to advance cognitive science, we have different opinions regarding how to achieve such a unified theory. Our position is that, instead of hybridizing different approaches or linearly mapping them to boost the Newell Test score, we need to recognize the multilevel nature of the human mind and develop complementary theories at both psychological and connectionist levels, and cross-validate them.

NOTE

1. Hongbin Wang is the corresponding author for this commentary.

Newell's program, like Hilbert's, is dead; let's move on

Yingrui Yang[a] and Selmer Bringsjord[b]

[a]*Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY 12180;* [b]*Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180.* **yangyri@rpi.edu** **selmer@rpi.edu** **http://www.rpi.edu/~brings**

Abstract: We draw an analogy between Hilbert's program (HP) for mathematics and Newell's program (NP) for cognitive modeling. The analogy reveals that NP, like HP before it, is fundamentally flawed. The only alternative is a program anchored by an admission that cognition is more than computation.

As most readers will know, Hilbert's program (HP) was devoted to building a system at or below the level of Turing machines (and their equivalents) to definitively settle all mathematical questions. Most readers will also know that in 1931, a young Viennese logician, Kurt Gödel, proved two incompleteness theorems – and Hilbert's program (HP) instantly died. Out of this death was born a much more sophisticated approach: In a word, true *meta* mathematics arose. One specific innovation was to devise and study infinitary logics not bound by Gödelian incompleteness, because, from an information-processing perspective, these logics are *beyond* Turing machines (Barwise 1980). Newell's program (NP) – the attempt to build a system at or below the level of Turing machines able to satisfy the criteria that Anderson & Lebiere (A&L) distill for us – has likewise expired. The difference is that apparently many influential cognitive scientists want to pretend the funeral never happened.

A&L, and in fact all those who see themselves as advancing NP, presuppose an exceedingly convenient sense of *universality.* According to this sense, a system is "universal" if and only if it can compute all Turing-computable functions. The construal is convenient because the vast majority of functions (in the relevant classes; e.g., functions from N to N) *aren't* Turing-computable (Boolos & Jeffrey 1989). A&L, the classical connectionists they evaluate, those they count as fans (e.g., Dennett), and so forth – all *assume* that human cognition can be nicely packaged beneath the Turing Limit. Yet, after decades of work, no system at or below the Turing Limit has the conversational power of a toddler (to pick just one criterion: 7). You would think the notion Newell (and his similarly sanguine partner, Simon) so confidently preached at the dawn of cognitive science (that thinking is computing at or below the Turing Limit, and computers operating at or below this limit with human-level intelligence will soon arrive) would, like Hilbert's dream, be a carcass at this point, but yet here is a *BBS* target article still stubbornly clinging to the religion (by cheerfully acting as if everyone is a believer). What arguments support the doctrine that cognition can be captured by standard computation? Surely no cogent argument is to be found on the basis of what has been built. LOGIC THEORIST, at the 1956 Dartmouth conference that kicked off AI, was able to prove the marvelously subtle theorem that *if p then q* implies *if not-q then not-p,* and this prompted Simon to declare that thinking machines would soon be among us. The situation is no different now: Here are A&L confidently pressing on to capture cognition in simple computation – but on the strength of what impressive artifact? Since seeing is believing, you will pardon us for not believing.

The problem isn't just criterion 7. Faced with consciousness, NP irremediably fails. Yet A&L once again cook up the convenient: They construe *consciousness* (in criterion 8) so that it simply leaves out the concept that threatens Newell's enterprise: namely, *phenomenal* consciousness. Block (1995) has recently explained the issue in this very journal. ACT-R and all forms of connectionism, and indeed every approach to sustaining NP, can't even take the first step toward expressing, in a third-person scheme, what it feels like to taste deep chocolate ice cream. ACT-R will be used to at most create what one of us (Bringsjord 2000) has called "zombanimals," that is, artificial animals with no inner lives. A robot with the behavioral repertoire of a dog, but with the inner life of a rock, might well be something NP, fueled by ACT-R, can produce.

That NP, as driven specifically by ACT-R, is dead, can be seen with help from concrete, not just philosophical, challenges. ACT-R is wholly incapable of modeling beliefs of the sort that human readers have when reading murder mysteries. For example, as detailed in Bringsjord (2000), readers have n-order beliefs about villains and detectives, and they make inferences based on these beliefs. For example, the reader of a murder mystery often believes that the villain believes that the detective believes that the villain believes the villain will never be caught. You can't represent this in ACT-R, period, because ACT-R is at best part of first-order logic devoid of doxastic operators. Of course, one could hook up a robust theory of human and machine reasoning (e.g., see Yang & Bringsjord, forthcoming) to ACT-R, but then in what sense *is* that new composite system ACT-R? If ACT-R is to be genuine science, it must be falsifiable. Yet A&L seem to describe an evolution in which serious challenges are handled by simply augmenting the system.

Just as the death of HP gave birth to infinitary logic, so should the death of NP give rise to cognitive modeling untrammeled by standard computation. Cognitive modelers need to step outside the notion that mere computation will suffice. They must face up to the fact, first, that the human mind encompasses not just the ordinary, humble computation that Newell and all his followers can't see beyond, but also *hyper*computation: information processing at a level *above* Turing machines, a level that can be formalized with help from analog chaotic neural nets, trial-and-error machines, Zeus machines, and the like (Bringsjord & Zenzen 2003; Siegelmann 1999).

In his famous "twenty questions" paper, Newell (1973) tells us that a sound science of the mind should not be steered by the willy-nilly dictates of experiment-driven empiricism. Instead, we are to do computational cognitive modeling. But such modeling, if limited by NP, fails to let cold hard reality lead the way. Instead, it lets speculative assumptions (e.g., that the mind processes information at or below the Turing Limit) prevent nature from proclaiming that we are more than ordinary machines.

Cognitive architectures need compliancy, not universality

Richard M. Young

Psychology Department, University of Hertfordshire, Hatfield, Herts AL10 9AB, United Kingdom. **r.m.young@herts.ac.uk**
http://www.psy.herts.ac.uk/pub/r.m.young/

Abstract: The criterion of computational universality for an architecture should be replaced by the notion of compliancy, where a model built within an architecture is compliant to the extent that the model allows the architecture to determine the processing. The test should be that the architecture *does easily* – that is, enables a compliant model to do – what people do easily.

Anderson & Lebiere (A&L) are to be congratulated on advancing the agenda of assessing cognitive architectures (or other cognitive theories of broad scope) as a whole. The inspiration is clearly Newell's, but the authors take a major step towards bringing Newell's criteria down to earth by operationalising them and bringing them closer to objective criteria and tests. This present commentary is offered as a minor contribution to that same goal.

In section 2.1, A&L follow Newell in identifying the criterion of *flexible behavior* with computational universality, that is, equivalence to a Turing machine. But Turing computability is inappropriate as a criterion for cognitive architectures. It is by nature an all-or-nothing test: Can, or cannot, the architecture be programmed to compute any Turing-computable function, yes or no? The authors certainly do themselves no favours by adopting Turing universality as the touchstone for flexible behaviour. Indeed, it forces them into a contradiction. Although in section 4.5 they deny that "ACT-R is a general computational system than can be programmed to do anything," that is indeed what being Turing universal means, that the architecture can be "programmed to do anything." What is needed instead is a graded measure, reflecting the reality that, as A&L acknowledge, "some things are much easier for people to learn and do than others." Ideally, the architecture should learn and do easily the things that people learn and do easily and, similarly, find the same things difficult.

Of course, what is meant by an architecture doing or learning something *easily* itself needs careful definition and explication. It is no trivial matter to replace the all-or-nothing concept of Turing computability by a more appropriate measure that both captures and makes precise these important but rather vague ideas about "doing something easily" or doing it by means "in keeping with the spirit of an architecture." However, a start has been made, with the concept of the *compliancy* of models constructed within a cognitive architecture. The idea has been worked through most thoroughly for SOAR, but is applicable in principle to any cognitive architecture.

In Howes and Young (1997), we approach the issue by considering how in practice architectures are used by cognitive modellers, and how credit and blame for the resulting models get assigned in the light of agreement with empirical data (or other evaluative criteria). We note how, in applying an architecture to a particular domain or task, the modeller inherits all the theoretical commitments of the architecture and then adds further commitments, often expressed in the form of production rules, which are specific to the domain or task being modelled. We refer to these additions as *model increments*, by analogy with the *method increments* which Laird (1986) identifies as giving rise to the "weak methods" of problem solving. We are led thereby to pose a methodological question: Given that a model (of this kind) consists of a cognitive architecture together with a specific model increment, in cases where the model does well, that is, provides a good prediction and explanation of the data, where does the credit belong: to the architecture, to the model increment, or somehow to both? And likewise if the model does badly, where lies the blame?

We note too that the extent to which cognitive architectures constrain and shape the models constructed within them, and thereby contribute to their predictions, is not widely recognised by those without first-hand experience of such architectures. Building model increments is not at all like writing programs in a theoretically neutral programming language. An architecture is not simply a programming environment for constructing cognitive models according to the modeller's fancy. Indeed, some architectures, of which SOAR (Newell 1990) is an example, are themselves capable of generating behaviour once they are given a specification of the task to be performed, even without further information about *how* it is to be performed. In such cases, the role of the model increment becomes not so much to generate behaviour, as to modulate or modify the behaviour which the architecture is already advocating.

That observation leads us to introduce the idea of compliancy. A model increment is compliant to the extent that it follows the architecture's lead, that is, takes advantage of the architecture's own tendency, allowing it mostly to do what it wants, intervening just occasionally to keep it on track. A model increment with low compliance, by contrast, disregards or overrules the architecture's own tendencies and simply forces the architecture to do what the model increment wants. (If the architecture is indeed Turing universal, then a model increment can always be written to produce any specified behaviour, but the increment may have to fight against the architecture in order to achieve that behaviour.)

The notion of compliancy allows us to answer the question about credit assignment. To the extent that the model increment is compliant with the architecture, much of the credit or blame attaches to the architecture itself. But to the extent that the model increment is noncompliant, responsibility for the resulting behaviour, whether good or ill, rests mostly with the model increment.

My suggestion is that compliancy also offers promise as a route for explicating what it means for an architecture to perform a task easily or with difficulty. An architecture can be said to find a task easy if a compliant model increment suffices to build a model to do it. Contrariwise, the architecture finds a task difficult if a noncompliant model increment is required, which therefore has to "force" the architecture to do it in a way "not in keeping with its spirit." By utilising compliancy, Newell's criterion of flexible behaviour can be interpreted as a requirement that the architecture does or learns easily (in other words, enables a compliant model to do or learn) what people find easy to do or learn, and finds difficult (in other words, requires a noncompliant model to do or learn) what people find difficult.

Authors' Response

Optimism for the future of unified theories

John R. Anderson and Christian Lebiere

Department of Psychology, Carnegie Mellon University, Pittsburgh, Pa 15213.
ja@cmu.edu cl@andrew.cmu.edu

Abstract: The commentaries on our article encourage us to believe that researchers are beginning to take seriously the goal of achieving the broad adequacy that Newell aspired to. The commentators offer useful elaborations to the criteria we suggested for the Newell Test. We agree with many of the commentators that classical connectionism is too restrictive to achieve this broad adequacy, and that other connectionist approaches are not so limited and can deal with the symbolic components of thought. All these approaches, including ACT-R, need to accept the idea that progress in science is a matter of better approximating these goals, and it is premature to be making judgments of true or false.

We begin by noting how pleased we were with the commentaries. Most commentators found something to disagree with, but these disagreements were by and large constructive and advanced the goals of defining criteria by which cognitive theories should be evaluated and using these criteria to evaluate many theories. In reading the commentaries and devising our responses we both increased our appreciation of alternative theories and refined our goals in pursuing ACT-R. We also increased our appreciation of the current state of theory development. The space of cognitive theories is indeed much more complex than our use of only two candidates could have suggested, with theories sharing some features and mechanisms while differing on others. As Herb Simon was advocating, one needs to go beyond evaluating theories as brands and consider them as a collection of mechanisms and evaluate them as such.

R1. Correcting the misconceptions

Before addressing specific points in the commentaries, we will correct a pair of misconceptions that were found with varying degrees of explicitness in some of the commentaries, reflecting a natural misreading of the paper. We were *not* using the criteria on the Newell Test as a basis for comparing classical connectionism and ACT-R, and we were *not* proposing them as a way to judge whether a theory should be deemed a success or a failure. We were not grading connectionism relative to ACT-R because it would not be credible for us to serve as either judge (in specifying the tests) or jury (in deciding which was best) in a contest between our theory and another one. However, it is perfectly reasonable for others to take these criteria and make judgments about the relative merits of the theories, as indeed some of the commentators have done.

Although it is fine for others to use these criteria for comparing theories, it is at least premature to be in the business of judging any theory an overall success or failure. All theories need a lot more development, and the point of such a set of criteria is to identify places where more work is needed. Therefore, we used a zero-sum grading scheme that forces one to identify where a theory needs the most work. Such a grading scheme forces a humbleness and self-criticism that the field could use.

With respect to the issue of falsification, a number of commentators (e.g., **Agassi**, **Taatgen**) speak with some fondness about the Turing Test in that it provides a criterion for rejecting theories. We too have some fondness for the Turing Test and frequently apply it to ACT-R simulations, not to provide an ultimate test of the theory but to force ourselves to see where ACT-R needs development. To try to repeat one of Herb Simon's frequent rejoinders as exactly as we can remember it: "If you just want to know whether the theory is wrong, then we can go home now. What I want to find out is how it is wrong and how it can be improved." The reason we formulated the Newell Test when the Turing Test was already available is because we wanted to provide some structure in this search for improvement.

The commentary by **Yang & Bringsjord** is surely the strongest in arguing for a yes-no judgment on theories. They argue that the whole class of computational theories, including ACT-R and classical connectionism, is dead. Their choice of the word "dead" rather than "false" gives away a lot. Unlike Gödel, whom they hold up as the ideal, Yang & Bringsjord provide nothing approaching a proof in their claims. As they should know from that important moment in the history of thought, the standards for making such sweeping negative pronouncements should be high. Gödel is such an important figure because he achieved those standards in his proofs.

We would like to particularly commend **Gray, Schoelles, and Myers** (**Gray et al.**) for bringing attention to cognitive engineering as a factor to shape these criteria. As they note, Newell thought cognitive engineering was an extremely important criterion for evaluating theories and much more than "just an application." Cognitive engineering gets at extremely profound issues about the nature of our science and the richly textured considerations that have to be brought to bear in evaluating cognitive theories and why simple yes-no, true-false judgments are typically inappropriate. This is a matter that deserves an elaborate community discussion. Such a discussion would reveal that the individual Newell tests are just the tips of a great iceberg.

R2. Developing the criteria

Agassi is correct that it is not always clear how to fully evaluate some of the criteria. In such cases the criteria should be stimuli for further thought and investigation so that they can be more fully applied. Indeed, many of the commentators have already proposed improvements and elaborations to the criteria. We particularly want to recommend the elaborations offered by **Taatgen**.

Gelepithis does a service in raising the issue of the exact relationship between the criteria we propose and those in Newell. As we think he creates too negative an impression of our scholarship, we will add some elaborations on this point. Ten of our criteria are verbatim from Newell (1980) and in the same order. We discuss at length in the target article the need to reformulate Newell's criterion 6 (symbols) as our criterion 6 (knowledge integration). Our criterion 12 ("be realizable within the brain") merges his criteria 12 ("be realizable within the brain as a physical system") and 13 ("be realizable as a physical system") because his distinction is not important to our paper nor is it a distinction that survived in his 1990 list. It is true that our list bears a less exact relationship to the 1990 list but at just three points: As can be seen from Gelepithis's Table 2, Newell in 1990 merged vast knowledge and robust behavior (criteria 4 and 5 in our table and in his 1980 table) into a single criterion (number 4 in the 1990 list), broke the developmental criterion (number 10 in our Table 1 and his list) into two criteria (8 and 12 in the 1990 list), and introduced a new criterion (social).

Criterion 4 in Newell's 1990 list covers our criteria 4 and 5 plus more. It is close to the embodiment criterion that **Spurrett** advocates, and Newell's reasons for reorganizing his list here may have been close to the arguments given by Spurrett. We think Spurrett's judgment of the relative performance of ACT-R versus connectionism on the criterion of embodiment is rather ungenerous. As **Gray et al.** note, ACT-R does well in addressing a range of HCI issues where connectionism has been almost totally silent. Nonetheless, robots are compelling demonstrations and hopefully someone in the ACT-R community will take up robots to satisfy Spurrett (and, we are sure, others).

Something may have been lost in collapsing Newell's

1990 developmental and embryological growth criteria into just the developmental criteria. **Sirois** offers a somewhat similar distinction between maturation, which he sees as a functional constraint, and development, which he sees as a functional ability to be explained.

Gelepithis offers a number of additional potential criteria from his 1999 paper. We agree that his suggestion of emotion and Newell's 1990 social behavior are two well-justified criteria. We made no claim to be exhaustive in choosing Newell's original 12. Our goal in working with those 12 was to have a manageable number for a *BBS* target article and to have criteria that came from some authoritative external source (to avoid the circularity that **Overgaard & Willert** mention).

As noted in the target article, the big missing criterion was accuracy of empirical predictions – having one's theory correspond to the details of empirical data. The criterion was missing only because it was not on Newell's lists, and it was not on Newell's lists only because he was talking about functionality at the points he introduced the lists, not because he did not strongly believe in its importance. Having a list that excludes predictive accuracy serves as something of a counterweight to the predominant tendency to consider only empirical adequacy and thus produce theories that are narrowly accurate in their predictions but incapable of being integrated into a complete functional theory of human cognition. However, in any final list that will serve as a "gold standard" (**Verschure**) the accuracy of empirical predictions needs to be given first place. It might make sense to give empirical adequacy half the weight in evaluating a theory and give the other half of the weight to functional criteria like those in Newell's list. As **Altmann** notes, such functional criteria can be crucial in deciding among theoretical accounts of particular phenomena that are hard to distinguish on the basis of their predictions. Functional criteria force the theories to consider difficult real-world problems rather than split hairs on tiny tasks that might not provide stringent enough tests to differentiate theories. One lesson that could be learned from the history of AI is the danger of focusing on abstract toy problems and the benefits of tackling the hard real-world problems.

One of the criteria that created some distress among several commentators (e.g., **Pyysiäinen**, **Young**), and for the reasons anticipated in the target article, is our attempt to operationalize flexible behavior as universality. Young has produced a superior version of this criterion in terms on what he calls "compliancy." It includes the test of universality as a component but connects differential difficulty of the models with the characteristics of the architecture. His development falls short of an explicit definition of what it means for one model to be more compliant than another. However, as is the case with other Newell criteria, that is a stimulus for further thought. Even in its current form it is better than the answer we could have composed to respond to **Tadepalli**'s worries about the range of models one can develop in an architecture.

Some commentators (**Wang et al.**; **Yang & Bringsjord**, **Overgaard & Willert**; **Sirois**) wonder whether it is possible to satisfy the consciousness constraint within any such framework. As both Overgaard & Willert and Yang & Bringsjord note, the answer to this question depends on what one takes to be consciousness. If we take it to be those aspects of consciousness that are amenable to scientific investigation, then we think the answer is yes. That may not include Block's (1995) phenomenal consciousness under some construals.

Wang asserts it is not possible to achieve all 12 criteria at the same level of explanation. For instance, he contends that ACT-R is too high-level to map onto brain structure. We disagree and offer the papers by Anderson et al. (2003), Qin et al. (2003), and Sohn et al. (2003) as emerging counterevidence. It is precisely because ACT-R is targeted at the architectural level of cognition that it is relevant to explaining the type of data generated by experimental neuroscience techniques such as fMRI. We think the mappings we proposed in Figure 1 of the target article have a lot of merit, but we agree with Wang that the connections displayed are not complete and that neuroscience evidence indicates that there are direct connections between some modules that do not go through the basal ganglia. Rather than be discouraged by this shortcoming, in the spirit of the Newell Test we take it as stimulus for further theoretical work.

Despite the fact that the list contains the two overlapping criteria of learning and development, a number of the commentators (**Commons & White**, **Prudkov**, **Roy**, and **Verschure**) argue that we did not give enough credit to self-organization. What they want is more emphasis on having a system that really constructed itself from experience without the guiding hand of the theorist. Though advocates of this criterion may not be giving adequate credit to what the system brings to the task as part of its genetic endowment, it is one of the holy grails of functionality. Perhaps it should be raised to a more prominent position. In addition to satisfying the subliminal "mad scientist" desire to see a being grow *de novo* in a computer program, achieving this serves an important role in constraining the degrees of freedom in proposing models within an architecture. Roy and Verschure are quite correct in noting that classical connectionism does not achieve this criterion even in its learning simulations, but we think this criterion is the dimension on which ACT-R suffers most in comparison to classical connectionism. As Prudkov notes, more has to be specified in typical ACT-R models before ACT-R learning can take over, than needs to be specified in connectionist models before connectionist learning can take over. We think this is because ACT-R models address more complex cognition, but the consequence is that it is more difficult to teach ACT-R aspirants what they need to know to become competent ACT-R modelers. One of our major goals in the future development of ACT-R is to move closer to achieving this holy grail.

Clancey challenges us to account for dysfunctional behavior as well as the functional. Of course, one cannot really have a theory of what is dysfunctional without first defining and accounting for functionality. This may not be another criterion to add to the list; rather it seems a different emphasis in evaluating the criteria that Newell has already given. However, we certainly agree with the importance of accounting for dysfunctionalities. Accounting for the full range of functional and dysfunctional behavior would also constitute a response by cognitive modeling to those who suggest that it is merely a parameters tuning game (since specific parameter values may map onto specific dysfunctionalities).

R3. Theories to which the criteria can be applied

An issue in applying the Newell criteria to classical connectionism or ACT-R is the degree to which these are re-

ally theories that can be so evaluated. **O'Loughlin & Karmiloff-Smith** argue that connectionism is a collection of tools that are useful. A similar point is often raised about ACT-R (e.g., by **Tadepalli**), and often elaborated in discussions of the distinctions between models and the architecture. We are certainly aware that connectionism in its largest sense is too broad for such an evaluation but we tried to focus on what we chose to call classical connectionism. **McClelland et al.** believe they have something to be evaluated, although they prefer to call it a framework in contrast to ACT-R, which they correctly call "an architecture." Nonetheless, they do regard themselves as having a "theoretical commitment to a common set of principles" that can serve as a basis for evaluation.

It is true that from among these theories one can define many models for performing tasks and that different models may differ in their predictions. However, it is just because of this fact that one needs to take the broader perspective of the overall functionality of the architecture. In part, this is so one can judge which models are in the spirit of the architecture, or "compliant" in **Young**'s term.

Many commentators (**Commons & White**, **Garzón**, **Gelepithis**, **Grossberg**, **Sun**, and **Verschure**) think that we unnecessarily restricted the kinds of neural networks considered by focusing on classical connectionism. Grossberg refers to classical connectionism as "Carnegie Mellon connectionism," implying that we were provincial in our outlook. Sun reminds us that we wrote that classical connectionism reflects "the core and the bulk" of existing neural network models (cf. target article, last para. of sect. 3). We clearly misspoke when we said "bulk" but we think we can still defend the claim that it is "the core" and not just a reflection of our provincialism. However, such a defense would be a digression here and we will just note our point of agreement with these commentators: They believe that classical connectionism is too restrictive and suffers weaknesses that more liberal uses of neural net ideas do not suffer. In particular, other forms of neural networks need have no problem with symbols. We agree and indeed view ACT-R as just a higher-level description of such a nonclassical connectionist theory. But there is a trade-off between assuming an overly broad definition of a framework that can account for anything (and its opposite) and an overly narrow one that leaves out many related efforts. We tried to strike the best balance possible in our definition of classical connectionism, expressing a common set of principles that are significantly constraining but broad enough to encompass a substantial part of connectionist efforts.

One of the things that encouraged us most was that some of commentators (**Clancey**, **Garzón**, **Grossberg**, **Verschure**) took many or all of the Newell criteria seriously and evaluated their theories on the basis of these criteria. Reading their short descriptions helped us appreciate those theories and caused us to read some of the sources they cited. Having done so, we do not want to take issue with their self-evaluations, and we hope the exercise helped them to see how they could improve their architectures.

R4. Past-tense issues

The target article held up the Taatgen and Anderson past-tense model as a paradigm of what could be accomplished in current ACT-R (cf. sect. 4.4; Taatgen & Anderson 2002), and the claims of that model came in for some analysis. One of the reasons for highlighting this model is that it depends so much on ACT-R learning mechanisms and so little on the initial structuring of the problem. As such it comes closest to achieving the *de novo* test that others want. Still, **Tadepalli** wonders to what degree its behavior reflects characteristics of the problem rather than ACT-R. This is an important question that needs to be asked more often. However, we do list things this model achieves that most other models facing the same problem do not achieve.

A number of commentators correctly point out shortcomings of the current model. **Ter Meulen** points out the inadequate conception of the semantics of past tense and failure to embed the model in a system that generates full utterances. **McClelland et al.** point out the impoverished conception of phonology, which limits the ability to extend the model because it relies on measures of phonological cost. One of the virtues of taking the Newell Test seriously is that one cannot just circle the wagons in response to criticisms like these and say that they are beyond the scope of the model. These are valid criticisms and point to directions for future work. Indeed, some steps have already been taken to enrich the treatment of the phonology (Misker & Anderson 2003; Taatgen & Dijkstra 2003). Taatgen and Dijkstra show how the approach can be used to produce "irregular generalizations like bring-brang." The Misker and Anderson analysis shows how complex phonological constraints like those in optimality theory (Prince & Smolensky 1993) can be represented and computed within ACT-R. Although it has not yet been done, we believe that if the Taatgen and Anderson (2002) learning approach were embedded on top of the Misker and Anderson approach, we would be able to account for such things as the distributional evidence that McClelland et al. cite with respect to the phonological characteristics of past tense exceptions.

R5. McClelland, Plaut, Gotts, and Maia (McClelland et al.)

We tried to define classical connectionism somewhat more broadly, but it is worthwhile to follow the lead of **McClelland et al.** and consider parallel distributed processing (PDP) specifically. The similarities between the broad goals of ACT-R and PDP and between some of their mechanisms can appear quite striking. From the perspective of a commentary like that of **Yang & Bringsjord**, our disagreements might seem like disputes between Baptists and Methodists. Aspects of ACT-R have been strongly influenced by connectionist ideas (frequently specifically PDP ideas) as described in the target article. Indeed, we think one of the major reasons for the success of the ACT-R effort is our willingness to incorporate good ideas – whether they come from EPIC (Meyer & Kieras 1997) or PDP.

The **McClelland et al.** commentary brings out three issues between ACT-R and PDP that need discussion. One has to do with the word "approximate," the second with the word "unified," and the third with the word "symbolic."

With respect to the word "approximate" one cannot help but read the commentary as using it a little bit as a dirty word (presumably in contrast to a good word like "exact"). In fact, to avoid any sense of not being collegial in their commentary, **McClelland et al.** hasten to note that they do not mean to suggest that we advocate approximation al-

though they wonder if Newell would. We cannot find where he said it in print, but one of the remarks we remember from our interactions with Newell is his assertion that the development of scientific theories is like an "approximating sequence." We agree with Newell on this score. Presumably no one can lay claim to having the true theory. **McClelland et al.** describe the symbolic level as "sometimes useful as high-level approximations to the underlying mechanisms of thought" (see their commentary Abstract). However, surely the units in a PDP model are only approximations to any neural processing which can at most claim to be useful as well. Their own recounting of the history of the development of their ideas is surely well described as an approximating sequence.

If one acknowledges that one's theory is an approximation that one is trying to make closer to the truth, then it becomes a strategic decision where one wants to work on improving the approximation. **McClelland et al.** advocate sticking within a well-circumscribed domain and working at getting their account closer and closer. Certainly we have done this, trying for more than 25 years (Anderson 1974; Anderson & Reder 1999b) to get an account of associative interference or the fan effect correct because we view this as central to the ACT theory. However, we do agree that we have put more attention in getting the approximations to work reasonably well across domains. This is even true in our work on the fan effect where we have tried to study it over a wide range of tasks. It is a strategic decision whether to try get some things really well, narrowly, and then go on to other topics, or whether to try to get a broad range of topics relatively well and then seek better approximations everywhere. The jury is surely still out on which is the better strategy. If the field of operations research offers any lesson in this regard, it is that the number and distribution of points that one is trying to fit is a stronger constraint than how closely they are fitted.

The second word, "unified," comes from the title of Newell's book, but thinking about it helps us understand the differences and similarities between the ACT-R and the PDP research strategies. Unified can mean two things: (1) that the theory tries to explain everything from the same few basic principles and (2) that the theory tries to explain how the broad range of intellectual functions is achieved in a single brain. We will refer to the first sense as "unitary" and the second sense as "integrated." Theoretical efforts can be cross-classified as to where they stand on these two dimensions. As **McClelland et al.** note, most theoretical accounts are neither unitary nor integrated, and PDP efforts share with Newell's SOAR and the ACT-R effort the aspiration to achieve more. However, it turns out that ACT-R, SOAR, and PDP each occupy a different cell of the remaining three in the two-by-two cross-classification. PDP shares with SOAR and differs from ACT-R in the desire to find a unitary theory – a small set of domain-general principles. ACT-R's predecessor, ACT* (Anderson 1983), did aspire to the same sort of unitary theory as SOAR and PDP. However, in response to the need to make progress on the Newell criteria we found this constraint to be an obstacle. Also, our understanding of biology "inspires" us to take the modular view described in Figure 1 of the target article. Imagine trying to account for respiration and digestion from a unitary set of principles! We see no more reason in our understanding of the brain to have that attitude about, for example, audition and manual control. (However, when possible we do try to exploit common principles in accounting for different modules – for we too like generalizations that work.)

On the other hand, we share with Newell and differ from the described PDP goals in having the aspiration to produce an integrated theory that explains how diverse and complex behaviors arise from one brain that has one set of mechanisms. This naturally leads to a focus on more complex behaviors such as mathematical problem solving or driving. We suspect we are more sympathetic to **ter Meulen**'s argument that the past tense model should be extended to deal with more complex constructions. If one believes that it is the same few principles working out the same way in domain after domain, then it makes sense to look at relatively simple tasks and model them intensely. If one believes that it is many modules interacting to produce complex adaptations, then it makes sense to look at a number of complex tasks.

Of course, there is the danger of becoming a jack of many trades and a master of none. This is why Anderson in his work on tutoring (Anderson et al. 1995; Koedinger et al. 1997) has focused almost exclusively on mathematical problem solving (and of a high school variety at that) because one has to understand that domain deeply. Newell (1973) himself saw the need to focus in depth on topics like chess to properly treat their richness. Fortunately, others in the ACT-R community have taken up other topics such as driving (Salvucci 2001) or past tense. Therefore, we certainly respect the decision of PDP researchers to focus on certain domains such as reading of words. One enviable feature of connectionism is the number of researchers who have taken up applying it to different domains. However, our bet is that the lack of concern with integration will lead to systems that cannot be put together – all the king's horses and all the king's men won't be able to put Humpty Dumpty together.

Finally, there is the word "symbolic." We changed Newell's criterion 6 to avoid the use of that word because it seemed too hopelessly loaded to ever serve as a useful criterion (and because his specification of this criterion really did not fit the functional character of the other criteria). Despite the frequency with which "symbolic" is used in Cognitive Science it seems to be more often a hindrance to communication than a help. A case in point is our claims about **McClelland et al.**'s attitude toward the symbolic level. McClelland et al. deny that the symbolic level is "the appropriate level at which principles of processing and learning should be formulated." That is what we meant when we said they did not "acknowledge a symbolic level to thought" (target article, Abstract), but apparently for them treating the symbolic level as sometimes a "fairly good approximation" amounts to acknowledging it. We did understand this about the PDP account (e.g., see Anderson 1990, pp 11–14). So we agree on what the PDP position does and does not say about the symbolic level, even if we cannot agree on the words to describe it.

For better or worse, we cannot entirely abandon using the word "symbolic" because we have long since committed to describing certain of ACT-R's principles as being at the symbolic level and others being at the subsymbolic level. Presumably, **McClelland et al.** would deny the appropriateness of the ACT-R principles that we describe as being at the symbolic level. We and others believe that it is this failure to incorporate such principles that produces the

limitations in their accounts. As we described in the target article, ACT-R's symbolic account cashes out at a connectionist level as prior constraints on the communication among the modules. Although McClelland et al. may not want to acknowledge such constraints, other connectionists have done so in terms of things like architectural constraints (Elman et al. 1996).

R6. Conclusion

Ours was a different target article than Newell (1992) and so naturally provoked a different set of commentaries. Still, we think that if he were to compare the commentaries in 2003 with those in 1992 he would see growth in the attitudes in Cognitive Science, maturation in the theories, and hope for the future.

References

Letters "a" and "r" appearing before authors' initials refer to target article and response respectively.

Ackley, D. H., Hinton, G. E. & Sejnowsky, T. J. (1985) A learning algorithm for Boltzmann machines. *Cognitive Science* 9:147–69. [aJRA]

Agassi, J. (1987) The wisdom of the eye. *Journal of Social and Biological Structures* 10:408–13. [JA]

(1988/2003) Winter 1988 Daedalus. *SIGArt Newsletter* 105:15–22; reprinted in Agassi (2003). [JA]

(1992) Heuristic computer-assisted, not computerized: Comments on Simon's project. *Journal of Epistemological and Social Studies on Science and Technology* 6:15–18. [JA]

(2003) *Science and culture. Boston Studies in the Philosophy of Science, vol. 231.* [JA]

Agassi, J. & Laor, N. (2000) How ignoring repeatability leads to magic. *Philosophy of the Social Sciences* 30:528–86. [JA]

Albright, A. & Hayes, B. (2001) *Rules vs. analogy in English past tenses: A computational/experimental study.* Department of Linguistics, UCLA. [JLM]

Altmann, E. M. (2002) Functional decay of memory for tasks. *Psychological Research* 66:287–97. [WDG]

Altmann, E. M. & Gray, W. D. (2000) An integrated model of set shifting and maintenance. In: *Proceedings of the third international conference on cognitive modeling,* pp. 17–24, ed. N. Taatgen & J. Aasman. Universal Press. [EMA]

(2002) Forgetting to remember: The functional relationship of decay and interference. *Psychological Science* 13(1):27–33. [WDG]

Altmann, E. M. & Trafton, J. G. (2002) Memory for goals: An activation-based model. *Cognitive Science* 26:39–83. [aJRA]

Anderson, J. R. (1974) Retrieval of propositional information from long-term memory. *Cognitive Psychology* 5:451–74. [rJRA]

(1976) *Language, memory, and thought.* Erlbaum. [aJRA]

(1983) *The architecture of cognition.* Harvard University Press. [arJRA]

(1990) *The adaptive character of thought.* Erlbaum. [arJRA]

(1991) Is human cognition adaptive? *Behavioral and Brain Sciences* 14:471–84. [aJRA]

(1993) *Rules of the mind.* Erlbaum. [aJRA, PAMG]

(2000) *Learning and memory,* 2nd edition. Wiley. [aJRA]

Anderson, J. R. & Betz, J. (2001) A hybrid model of categorization. *Psychonomic Bulletin and Review* 8:629–47. [aJRA]

Anderson, J. R., Bothell, D., Lebiere, C. & Matessa, M. (1998a) An integrated theory of list memory. *Journal of Memory and Language* 38:341–80. [aJRA]

Anderson, J. R., Boyle, C. F., Corbett, A. T. & Lewis, M. W. (1990) Cognitive modeling and intelligent tutoring. In: *Artificial intelligence and learning environments,* ed. W. J. Clancey & E. Soloway. Elsevier. [WJC]

Anderson, J. R., Budiu, R. & Reder, L. M. (2001) A theory of sentence memory as part of a general theory of memory. *Journal of Memory and Language* 45:337–67. [aJRA]

Anderson, J. R., Corbett, A. T., Koedinger, K. & Pelletier, R. (1995) Cognitive tutors: Lessons learned. *The Journal of Learning Sciences* 4:167–207. [rJRA]

Anderson, J. R. & Douglass, S. (2001) Tower of Hanoi: Evidence for the cost of goal retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27:1331–46. [aJRA]

Anderson, J. R. & Lebiere, C. (1998) *The atomic components of thought.* Erlbaum. [aJRA, PAMG]

Anderson, J. R., Lebiere, C., Lovett, M. C. & Reder, L. M. (1998b) ACT-R: A higher-level account of processing capacity. *Behavioral and Brain Sciences* 21:831–32. [aJRA]

Anderson, J. R., Qin, Y., Sohn, M-H., Stenger, V. A. & Carter, C. S. (2003) An information-processing model of the BOLD response in symbol manipulation tasks. *Psychonomic Bulletin and Review* 10:241–61. [arJRA]

Anderson, J. R. & Reder, L. M. (1999a) The fan effect: New results and new theories. *Journal of Experimental Psychology: General* 128:186–197. [aJRA]

Anderson, J. R. & Reder, L. M. (1999b) The size of the fan effect: Process not representation. *Journal of Experimental Psychology: General* 128:207–10. [rJRA]

Asher, N., Aurnague, M., Bras, M., Sblayrolles, P. & Vieu. L. (1994) Computing the spatiotemporal structure of discourse. *1st International Workshop on Computational Semantics,* Dept. of Computational Linguistics, University of Tilburg, NL. [AGBtM]

Atran, S. (2002) Modes of adaptationism: Muddling through cognition and language. Commentary on Andrews et al. *Behavioral and Brain Sciences* 25(4):504–06. [IP]

Baddeley, A. D. (1986) *Working memory.* Oxford University Press. [aJRA]

Ballard, D. H., Hayhoe, M. M., Pook, P. K. & Rao, R. P. N. (1997) Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences* 20(4):723–42. [WDG]

Baron-Cohen, S. (1996) Can children with autism integrate first and third person representations? *Behavioral and Brain Sciences* 19(1):123–24. [WJC]

Barresi, J. & Moore, C. (1996) Intentional relations and social understanding. *Behavioral and Brain Sciences* 19(1)107–54. [WJC]

Barrett, J. L. (1998) Cognitive constraints on Hindu concepts of the divine. *Journal for the Scientific Study of Religion* 37:608–19. [IP]

(1999) Theological correctness: Cognitive constraint and the study of religion. *Method and Theory in the Study of Religion* 11:325–39. [IP]

Barrett, J. L. & Keil, F. E. (1996) Conceptualizing a nonnatural entity: Anthropomorphism in God concepts. *Cognitive Psychology* 31:219–47. [IP]

Barwise, J. (1980) Infinitary logics. In: *Modern logic: A survey,* ed. E. Agazzi. Reidel. [YY]

Besner, D., Twilley, L., McCann, R. S. & Seergobin, K. (1990) On the connection between connectionism and data: Are a few words necessary? *Psychological Review* 97(3):432–46. [JLM]

Bever, T. G., Fodor, J. A. & Garret, M. (1968) A formal limitation of association. In: *Verbal behavior and general behavior theory,* ed. T. R. Dixon & D. L. Horton. Prentice Hall. [aJRA]

Bird, H., Lambon Ralph, M. A., Seidenberg, M. S., McClelland, J. L. & Patterson, K. (2003) Deficits in phonology and past-tense morphology: What's the connection? *Neuropsychologia* 48:502–26. [JLM]

Block, N. (1995) On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18:227–87. [rJRA, MO, YY]

Boardman, I., Grossberg, S., Myers, C. & Cohen, M. (1999) Neural dynamics of perceptual order and context effects for variable-rate speech syllables. *Perception and Psychophysics* 61:1477–1500. [SG]

Bock, K. (1986) Syntactic persistence in language production. *Cognitive Psychology* 18:355–87. [aJRA]

Bock, K. & Griffin, Z. M. (2000) The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental psychology: General* 129:177–92. [aJRA]

Boolos, G. & Jeffrey, R. (1989) *Computability and logic.* Cambridge University Press. [YY]

Botvinick, M. & Plaut, D. C. (submitted) Doing without schema hierarchies: A recurrent connectionist approach to routine sequential action and its pathologies. [aJRA]

Bouvet, S. (2001) Learning an ideal strategy in tic-tac-toe with DAC5. *Technical Report: Institute of Neuroinformatics* 2001–12. [PFMJV]

Boyer, P. (2001) *Religion explained: The evolutionary origins of religious thought.* Basic Books. [IP]

Bradski, G., Carpenter, G. A. & Grossberg, S. (1994) STORE working memory networks for storage and recall of arbitrary temporal sequences. *Biological Cybernetics* 71:469–80. [SG]

Bringsjord, S. (2000) Animals, zombanimals, and the total Turing Test: The essence of artificial intelligence. *Journal of Logic, Language, and Information* 9:397–18. [YY]

(2001) Are we evolved computers?: A critical review of Steven Pinker's *How the mind works. Philosophical Psychology* 14(2):227–43. [IP]

Bringsjord, S. & Zenzen, M. (2003) *Superminds: People harness hypercomputation, and more.* Kluwer. [YY]

Brown, J., Bullock, D. & Grossberg, S. (1999) How the basal ganglia use parallel

excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience* 19:10502–511. [SG]
Brown, R. (1973) *A first language.* Harvard University Press. [aJRA,JLM]
Browne, A. & Sun, R. (2001) Connectionist inference models. *Neural Networks* 14:1331–55. [aJRA]
Budiu, R. (2001) *The role of background knowledge in sentence processing.* Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA. [aJRA]
Budiu, R. & Anderson, J. R. (submitted) Interpretation-based processing: A unified theory of semantic processing. *Cognitive Science.* [aJRA]
Bullock, D., Cisek, P. & Grossberg, S. (1998) Cortical networks for control of voluntary arm movements under variable force conditions. *Cerebral Cortex* 8:48–62. [SG]
Bullock, D., Grossberg, S. & Guenther, F. (1993a) A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm. *Journal of Cognitive Neuroscience* 5:408–35. [SG]
Bullock, D., Grossberg, S. & Mannes, C. (1993b) A neural network model for cursive script production. *Biological Cybernetics* 70:15–28. [SG]
Bunge, M. (1980) *The mind-body problem: A psychobiological approach.* Pergamon Press. [PAMG]
Burzio, L. (1999) Missing players: Phonology and the past-tense debate. Unpublished manuscript. Johns Hopkins University, Baltimore, MD. [aJRA]
Bybee, J. L. (1995) Regular morphology and the lexicon. *Language and Cognitive Processes* 10:425–55. [JLM]
Byrne, M. D. & Anderson, J. R. (1998) Perception and action. In: *The atomic components of thought,* ed. J. R. Anderson & C. Lebiere. Erlbaum. [aJRA]
(2001) Serial modules in parallel: The psychological refractory period and perfect time-sharing. *Psychological Review* 108:847 – 69. [aJRA]
Callan, D. E., Kent, R. D., Guenther, F. H. & Vorperian, H. K. (2000) An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech, Language, and Hearing Research* 43:721–36. [SG]
Calvo, F. & Colunga, E. (2003) The statistical brain: Reply to Marcus' *The Algebraic Mind.* In: *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society,* ed. R. Alterman & D. Kirsh. Erlbaum. [FCG]
(in preparation) Transfer of learning in infants: Combined Hebbian and error-driven learning. [FCG]
Carpenter, G. A., Gopal, S., Macomber, S., Martens, S., Woodcock, C. E. & Franklin, J. (1999) A neural network method for efficient vegetation mapping. *Remote Sensing of Environment* 70:326–38. [SG]
Carpenter, G. A. & Grossberg, S., eds. (1991) *Pattern recognition by self-organizing neural networks.* MIT Press. [SG]
Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H. & Rosen, D. (1992) Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks* 3(5):698–713. [SG]
Carpenter, G. A. & Milenova, B. L. (2000) ART neural networks for medical data analysis and fast distributed learning. In: *Artificial neural networks in medicine and biology. Proceedings of the ANNIMAB-1 Conference, Göteborg, Sweden, 13–16 May 2000,* ed. H. Malmgren, M. Borga & L. Niklasson. Springer-Verlag. (Springer series, *Perspectives in Neural Computing.*) [SG]
Chaiken, S. & Trope, Y., eds. (1999) *Dual-process theories in social psychology.* Guilford Press. [IP]
Chalmers, D. J. (1996) *The conscious mind.* Oxford University Press. [MO]
Chase, W. G. & Ericsson, K. A. (1982) Skill and working memory. In: *The psychology of learning and motivation, vol. 16,* ed G. H. Bower. Academic Press. [aJRA]
Chomsky, N. A. (1965) *Aspects of a theory of syntax.* MIT Press. [aJRA]
Clancey, W. J. (1999a) *Conceptual coordination: How the mind orders experience in time.* Erlbaum. [WJC]
(1999b) Studying the varieties of consciousness: Stories about zombies or the life about us? *Journal of the Learning Sciences* 8(3–4):525–40. [WJC]
(2000) Conceptual coordination bridges information processing and neuropsychology. *Behavioral and Brain Sciences* 23(6):919–22. [WJC]
Clark, A. (1997) *Being there.* MIT Press. [DS]
(1998) The dynamic challenge. *Cognitive Science* 21(4):461–81. [aJRA]
(1999) Where brain, body, and world collide. *Journal of Cognitive Systems Research* 1:5–17. [aJRA]
Cleeremans, A. (1993) *Mechanisms of implicit learning: Connectionist models of sequence processing.* MIT Press. [aJRA]
Cohen, J. D. & Schooler, J. W., eds. (1997) *Scientific approaches to consciousness: 25th Carnegie Symposium on Cognition.* Erlbaum. [aJRA]
Cohen, M. A. & Grossberg, S. (1986) Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short-term memory. *Human Neurobiology* 5:1–22. [SG]
Collins, M. (1999) *Head-driven statistical models for nature language parsing.* Doctoral dissertation, University of Pennsylvania. [aJRA]
Coltheart, M., Curtis, B., Atkins, P. & Haller, M. (1993) Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review* 100(4):589–608. [JLM]
Commons, M. L. & Richards, F. A. (2002) Organizing components into combinations: How stage transition works. *Journal of Adult Development* 9(3):159–77. [MLC]
Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, F. A. & Krause, S. R. (1998) The existence of developmental stages as shown by the hierarchical complexity of tasks. *Developmental Review* 8(3):237–78. [MLC]
Contreras-Vidal, J. L., Grossberg, S. & Bullock, D. (1997) A neural model of cerebellar learning for arm movement control: Cortico-spino-cerebellar dynamics. *Learning and Memory* 3:475–502. [SG]
Cosmides, L. & Tooby, J. (2000a) Consider the source: The evolution of adaptations for decoupling and metarepresentation. In: *Metarepresentations: A multidisciplinary perspective,* ed. D. Sperber. Oxford University Press. [IP]
(2000b) The cognitive neuroscience of social reasoning. In: *The new cognitive neurosciences,* 2nd edition, ed. M. S. Gazzaniga. MIT Press. [aJRA]
Dawson, T. L. (2002) A comparison of three developmental stage scoring systems. *Journal of Applied Measurement* 3(2):146–89. [MLC]
Denes-Raj, V. & Epstein, S. (1994) Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology* 66(5):819–29. [IP]
Dennett, D. C. (1991) *Consciousness explained.* Little, Brown. [aJRA]
Dennett, D. C. & Kinsbourne, M. (1995) Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences* 15(2):183–247. [aJRA]
Dolan, C. P. & Smolensky, P. (1989) Tensor product production system: A modular architecture and representation. *Connection Science* 1:53–68. [aJRA]
Edelman, G. M. (1989) *The remembered present: A biological theory of consciousness.* BasicBooks. [PAMG]
(1992) *Bright air, brilliant fire: On the matter of the mind.* BasicBooks. [PAMG]
Edelman, G. M. & Tononi, G. (2000) *Consciousness: How matter becomes imagination.* Penguin Press. [PAMG]
Ehret, B. D., Gray, W. D. & Kirschenbaum, S. S. (2000) Contending with complexity: Developing and using a scaled world in applied cognitive research. *Human Factors* 42(1):8–23. [WDG]
Elman, J. L. (1995) Language as a dynamical system. In: *Mind as motion: Explorations in the dynamics of cognition,* ed. R. F. Port & T. V. Gelder. MIT Press. [aJRA]
Elman, J. L., Bates, E. A, Johnson, M. H., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996) *Rethinking innateness: A connectionist perspective on development.* MIT Press. [arJRA, CFO]
Emond, B. (in preparation) ACT-R/WN: Towards an implementation of WordNet in the ACT-R cognitive architecture. [aJRA]
Emond, B. & Ferres, L. (2001) Modeling the false-belief task: An ACT-R implementation of Wimmer & Perner (1983). Paper presented at the Second Bisontine Conference for Conceptual and Linguistic Development in the Child Aged from 1 to 6 Years, Besançon, France, March 21–23, 2001. [aJRA]
Eng, K., Klein, D., Bäbler, A., Bernardet, U., Blanchard, U., Costa, M., Delbrück, T., Douglas, R. J., Hepp, K., Manzolli, J., Mintz, M., Roth, F., Rutishauser, U., Wassermann, K., Whatley, A. M., Wittmann, A., Wyss, R., Verschure, P. F. M. J. (2003) Design for a brain revisited: The neuromorphic design and functionality of the interactive space Ada. *Reviews in the Neurosciences* 1–2:145–80. [PFMJV]
Engel, A. K., Fries, P. & Singer, W. (2001) Dynamic predictions, oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience* 2:704–16. [SG]
Epstein, S., Lipson, A., Holstein, C. & Huh, E. (1992) Irrational reactions to negative outcomes: Evidence for two conceptual systems. *Journal of Personality and Social Psychology* 62(2):328–39. [IP]
Epstein, S. & Pacini, R. (1999) Some basic issues regarding dual-process theories from the perspective of cognitive-experiential self-theory. In: *Dual-process theories in social psychology,* ed. S. Chaiken & Y. Trope. Guilford Press. [IP]
Ericsson, K. A. & Kintsch, W. (1995) Long-term working memory. *Psychological Review* 102:211–45. [aJRA]
Fellbaum, C., ed. (1998) *WordNet: An electronic lexical database.* MIT Press. [aJRA]
Ferreira, F. & Clifton, C. (1986) The independence of syntactic processing. *Journal of Memory and Language* 25:348–68. [aJRA]
Fiala, J. C., Grossberg, S. & Bullock, D. (1996) Metabotropic glutamate receptor activation in cerebellar Purkinje cells as substrate for adaptive timing of the classically conditioned eye-blink response. *Journal of Neuroscience* 16:3760–74. [SG]
Fincham, J. M., VanVeen, V., Carter, C. S., Stenger, V. A. & Anderson, J. R. (2002) Integrating computational cognitive modeling and neuroimaging: An event-

related fMRI study of the Tower of Hanoi task. *Proceedings of the National Academy of Science* 99:3346–51. [aJRA]
Fodor, J. A. (1983) *The modularity of mind.* MIT Press/Bradford Books. [aJRA]
(2000) *The mind doesn't work that way.* MIT Press. [aJRA]
Frank, M. J., Loughry, B. & O'Reilly, R. C. (2000) Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Technical Report* (00–01, November), Institute of Cognitive Science, University of Colorado, Boulder, CO. [aJRA]
Gancarz, G. & Grossberg, G. (1999) A neural model of the saccadic eye movement control explains task-specific adaptation. *Vision Research* 39:3123–43. [SG]
Gelepithis, P. A. M. (1984) *On the foundations of artificial intelligence and human cognition.* Ph. D. thesis, Department of Cybernetics, Brunel University, England. [PAMG]
(1991) The possibility of machine intelligence and the impossibility of human-machine communication. *Cybernetica* 34(4):255–68. [PAMG]
(1997) A Rudimentary theory of information: Consequences for information science and information systems. *World Futures* 49:263–74. (Reprinted in: *The quest for a unified theory of information,* ed. W. Hofkirchner. Gordon and Breach.) [PAMG]
(1999) Embodiments of theories of mind: A review and comparison. In: *Computational methods and neural networks,* ed. M. P. Bekakos, M. Sambandham & D. J. Evans. Dynamic. [PAMG]
(2001) A concise comparison of selected studies of consciousness, *Cognitive Systems* 5(4):373–92. [PAMG]
(2002) An axiomatic approach to the study of mind. *Res-Systemica: Proceedings of the Fifth European Systems Science Congress, October 2002, Crete.* http://www.afscet.asso.fr/resSystemica/. [PAMG]
Gelepithis, P. A. M. & Goodfellow, R. (1992) An alternative architecture for intelligent tutoring systems: Theoretical and implementational aspects. *Interactive Learning International* 8(3):171–75. [PAMG]
Gelepithis, P. A. M. & Parillon, N. (2002) Knowledge management: Analysis and some consequences. In: *Knowledge management and business process reengineering,* ed. V. Hlupic. Idea Book. [PAMG]
Gibson, J. J. (1979) *The ecological approach to visual perception.* Houghton Mifflin. [MO]
Giere, R. (1998) *Explaining science.* Chicago University Press. [PAMG]
Gigerenzer, G. (2000) *Adaptive thinking: Rationality in the real world.* Oxford University Press. [aJRA]
Gopnik, M. & Crago, M. B. (1991) Familial aggregation of a developmental language disorder. *Cognition* 39:1–50. [JLM]
Gould, S. J. & Lewontin, R. C. (1979) The Spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist program. *Proceedings of the Royal Society of London* 205:581–98. [aJRA]
Granger, E., Rubin, M., Grossberg, S. & Lavoie, P. (2001) A what-and-where fusion neural network for recognition and tracking of multiple radar emitters. *Neural Networks* 14:325–44. [SG]
Gray, W. D. & Boehm-Davis, D. A. (2000) Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied* 6(4):322–35. [WDG]
Gray, W. D., John, B. E. & Atwood, M. E. (1993) Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world performance. *Human-Computer Interaction* 8(3):237–309. [WDG]
Gray, W. D., Schoelles, M. J. & Fu, W.-T. (2000) Modeling a continuous dynamic task. In: *Third International Conference on Cognitive Modeling,* ed. N. Taatgen & J. Aasman. Universal Press. [WDG]
Gray, W. D., Schoelles, M. J. & Myers, C. W. (2002) Computational cognitive models ISO ecologically optimal strategies. In: *46th Annual Conference of the Human Factors and Ergonomics Society,* pp. 492–96. Human Factors & Ergonomics Society. [WDG]
Green C. D. (1998) Are connectionist models theories of cognition? *Psycoloquy* 9(04). http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?9.04. [PNP]
Greeno, J. G. (1989) Situations, mental models and generative knowledge. In: *Complex information processing: The impact of Herbert A. Simon,* ed. D. Klahr & K. Kotovsky. Erlbaum. [aJRA]
Grossberg, S. (1976) Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics* 23:187–202. [SG]
(1978a) A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In: *Progress in theoretical biology, vol. 5,* ed. R. Rosen & F. Snell. Academic Press. [SG]
(1978b) Behavioral contrast in short-term memory: Serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology* 3:199–219. [SG]
(1980) How does a brain build a cognitive code? *Psychological Review* 87:1–51. [SG]
(1988) Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks* 1:17–61. [SG]
(1999a) How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex. *Spatial Vision* 12:163–86. [SG]
(1999b) The link between brain learning, attention, and consciousness. *Consciousness and Cognition* 8:1–44. [SG]
(2000a) How hallucinations may arise from brain mechanisms of learning, attention, and volition. *Journal of the International Neuropsychological Society* 6:583–92. [SG]
(2000b) The complementary brain: Unifying brain dynamics and modularity. *Trends in Cognitive Sciences* 4:233–46. [SG]
(2000c) The imbalanced brain: From normal behavior to schizophrenia. *Biological Psychiatry* 48:81–98. [SG]
Grossberg, S., Boardman, I. & Cohen, C. (1997a) Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance* 23:418–503. [SG]
Grossberg, S. & Kuperstein, M. (1989) Neural dynamics of adaptive sensory-motor control: Expanded edition. Pergamon Press. [SG]
Grossberg, S. & Merrill, J. W. L. (l992) A neural network model of adaptively timed reinforcement learning and hippocampal dynamics. *Cognitive Brain Research* 1:3–38. [SG]
(l996) The hippocampus and cerebellum in adaptively timed learning, recognition, and movement. *Journal of Cognitive Neuroscience* 8:257–77. [SG]
Grossberg, S. & Myers, C. W. (2000) The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review* 107:735–67. [SG]
Grossberg, S. & Paine, R. W. (2000) A neural model of corticocerebellar interactions during attentive imitation and predictive learning of sequential handwriting movements. *Neural Networks* 13:999–1046. [SG]
Grossberg, S., Roberts, K., Aguilar, M. & Bullock, D. (1997b) A neural model of multimodal adaptive saccadic eye movement control by superior colliculus. *Journal of Neuroscience* 17:9706–25. [SG]
Grossberg, S. & Stone, G. O. (1986a) Neural dynamics of attention switching and temporal order information in short-term memory. *Memory and Cognition* 14:451–68. [SG]
(1986b) Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance. *Psychological Review* 93:46–74. [SG]
Guenther, F. H. (1995) Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review* 102:594–621. [SG]
Hahn, U. & Nakisa, R. C. (2000) German inflection: Single-route or dual-route? *Cognitive Psychology* 41:313–60. [JLM]
Hameroff, S. R., Kaszniak, A. W. & Scott, A. C., eds. (1998) *Toward a science of consciousness II: The second Tucson discussions and debates.* MIT Press. [PAMG]
Harnad, S. (1990) The symbol grounding problem. *Physica D* 42:335–46. [aJRA]
(1994) Computation is just interpretable symbol manipulation; Cognition isn't. *Minds and Machines* 4:379–90. [aJRA]
Hartley, R. F. (2000) Cognition and the computational power of connectionist networks. *Connection Science* 12(2):95–110. [aJRA]
Hartley, R. & Szu, H. (1987) A comparison of the computational power of neural network models. *Proceedings of the First International Conference on Neural Networks* 3:15–22. [aJRA]
Haverty, L. A., Koedinger, K. R., Klahr, D. & Alibali, M. W. (2000) Solving induction problems in mathematics: Not-so-trivial pursuit. *Cognitive Science* 24(2):249–98. [aJRA]
Hebb, D. O. (1949) *The organization of behavior.* Wiley. [PAMG]
Hinrichs, E. (1986) Temporal anaphora in discourses of English. *Linguistics and Philosophy* 9:63–82. [AGBtM]
Hinton, G. E. & Sejnowsky, T. J. (1986) Learning and relearning in Boltzmann machines. In: *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1: Foundations,* ed. D. E. Rumelhart, J. L. McClelland & The PDP Group. MIT Press. [aJRA]
Hoeffner, J. H. (1996) A single mechanism account of the acquisition and processing of regular and irregular inflectional morphology. Unpublished doctoral dissertation, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA. [JLM]
Hofstötter, C., Mintz, M. & Verschure, P. F. M. J. (2002) The cerebellum in action: A simulation and robotics study. *European Journal of Neuroscience* 16:1361–76. [PFMJV]
Holyoak, K. J. & Spellman, B. A. (1993) Thinking. *Annual Review of Psychology* 44:265–315. [IP]
Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA* 79:2554–58. [aJRA]
Hornik, K., Stinchcombe, M. & White, H. (1989) Multilayer feed forward networks are universal approximators. *Neural Computation* 2:210–15. [aJRA]
Howes, A. & Young, R. M. (1997) The role of cognitive architecture in modelling

the user: Soar's learning mechanism. *Human-Computer Interaction* 12:311–43. [RMY]

Hummel, J. E. & Holyoak, K. J. (1998) Distributed representations of structure. A theory of analogical access and mapping. *Psychological Review* 104:427–66. [aJRA]

Hutchins, E. (1995) *Cognition in the wild.* MIT Press. [DS]

Joanisse, M. F. & Seidenberg, M. S. (1999) Impairments in verb morphology following brain injury: A connectionist model. *Proceedings of the National Academy of Sciences* 96:7592–97. [JLM]

Johnson, M. H., Munakata, Y. & Gilmore, R. O., eds. (2002) *Brain development and cognition: A reader.* Blackwell. [PAMG]

Jones, G., Ritter, F. E. & Wood, D. J. (2000) Using a cognitive architecture to examine what develops. *Psychological Science* 11(2):1–8. [aJRA]

Jones, R. M., Laird, J. E., Nielsen, P. E., Coulter, K. J., Kenny, P. & Koss, F. V. (1999) Automated intelligent pilots for combat flight simulation. *AI Magazine* 20:27–41. [aJRA]

Jongman, L. & Taatgen, N. A. (1999) An ACT-R model of individual differences in changes in adaptivity due to mental fatigue. In: *Proceedings of the 21st Annual Conference of the Cognitive Science Society.* Erlbaum. [aJRA]

Kandel, E. R. & O'Dell, T. J. (1992) Are adult learning mechanisms also used for development? *Science* 258:243–45. [SG]

Karmiloff-Smith, A. (1992) *Beyond modularity: A developmental perspective on cognitive science.* MIT Press. [CFO]

(1997) Crucial differences between developmental cognitive neuroscience and adult neuropsychology. *Developmental Neuropsychology* 13(4):513–24. [CFO]

(1998) Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences* 2(10):389–98. [CFO]

Karmiloff-Smith, A., Plunkett, K., Johnson, M. H., Elman, J. L. & Bates, E. A. (1998) What does it mean to claim that something is "innate"? *Mind and Language* 13(4):588–97. [CFO]

Karmiloff-Smith, A., Scerif, G. & Ansari, D. (2003) Double dissociations in developmental disorders? Theoretically misconceived, empirically dubious. *Cortex* 39:161–3. [CFO]

Karmiloff-Smith, A., Scerif, G. & Thomas, M. S. C. (2002) Different approaches to relating genotype to phenotype in developmental disorders. *Developmental Psychobiology* 40:311–22. [CFO]

Kilgard, M. P. & Merzenich, M. M. (1998) Cortical map reorganization enabled by nucleus basalis activity. *Science* 279:1714–18. [PFMJV]

Kimberg, D. Y. & Farah, M. J. (1993) A unified account of cognitive impairments following frontal lobe damage: The role of working memory in complex, organized behavior. *Journal of Experimental Psychology: General* 122:411–28. [aJRA]

Kirsh, D. & Maglio, P. P. (1994) On distinguishing epistemic from pragmatic action. *Cognitive Science* 18(4):513–49. [DS]

Koedinger, K. R., Anderson, J. R., Hadley, W. H. & Mark, M. (1997) Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education* 8:30–43. [rJRA]

Laird, J. E. (1986) Universal subgoaling. In: *Universal subgoaling and chunking: The automatic generation and learning of goal hierarchies,* ed. J. E. Laird, P. S. Rosenbloom & A. Newell. Kluwer. [RMY]

Lakatos, I. (1970) Falsification and the methodology of scientific research programmes. In: *Criticism and the growth of knowledge,* ed. I. Lakatos & A. Musgrave. Cambridge University Press. [CFO]

Langacker, R. W. (1986) An introduction to cognitive grammar. *Cognitive Science* 10(1):1–40. [WJC]

Langley, P. (1999) Concrete and abstract models of category learning. In: *Proceedings of the 21st Annual Conference of the Cognitive Science Society.* Erlbaum. [PT]

Lave, J. (1988) *Cognition in practice: Mind, mathematics, and culture in everyday life.* Cambridge University Press. [aJRA]

Lebiere, C. (1998) *The dynamics of cognition: An ACT-R model of cognitive arithmetic.* Doctoral dissertation. CMU Computer Science Dept. Technical Report CMU-CS-98–186. Pittsburgh, PA. [aJRA]

Lebiere, C. & Anderson, J. R. (1993) A connectionist implementation of the ACT-R production system. In: *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society.* [aJRA]

Lerch, F. J., Gonzalez, C. & Lebiere, C. (1999) Learning under high cognitive workload. In: *Proceedings of the Twenty-first Conference of the Cognitive Science Society.* Erlbaum. [aJRA]

Lewis, R. L. (1999) Attachment without competition: A race-based model of ambiguity resolution in a limited working memory. Presented at the CUNY Sentence Processing Conference, New York. [aJRA]

Liben, L. S. (1987) Approaches to development and learning: Conflict and congruence. In: *Development and learning: Conflict or congruence?* ed. L. S. Liben. Erlbaum. [SS]

Lieberman, M. D. (2000) Intuition: A social cognitive neuroscience approach. *Psychological Bulletin* 126(1):109–37. [IP]

Lieberman, M. D., Gaunt, R., Gilbert, D. T. & Trope, Y. (2002) Reflexion and reflection: A social cognitive neuroscience approach to attributional inference. *Advances in Experimental Social Psychology* 34:200–250. [IP]

Lodge, D. (1984) *Small world.* Penguin. [NAT]

Logan, G. D. (1988) Toward an instance theory of automatization. *Psychological Review* 95:492–527. [aJRA]

Lovett, M. C. (1998) Choice. In: *The atomic components of thought,* ed. J. R. Anderson & C. Lebiere. Erlbaum. [aJRA]

Lovett, M. C., Daily, L. Z. & Reder, L. M. (2000) A source activation theory of working memory: Cross-task prediction of performance in ACT-R. *Cognitive Systems Research* 1:99–118. [aJRA]

Lovett, M. C., Reder, L. & Lebiere, C. (1997) Modeling individual differences in a digit working memory task. In: *Proceedings of the 19th Annual Conference of the Cognitive Science Society,* ed. M. G. Shafto. & P. Langley. Erlbaum. [NAT]

MacDonald, M. C., Pearlmutter, N. J. & Seidenberg, M. S. (1994) The lexical nature of syntactic ambiguity resolution. *Psychological Review* 101:676–703. [aJRA]

Magerman, D. (1995) Statistical decision-tree models for parsing. In: *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics.* [aJRA]

Marcus, G. F. (2001) *The algebraic mind: Integrating connectionism and cognitive science.* MIT Press. [aJRA, FCG, JLM]

Marcus G. F., Brinkmann U., Clahsen H., Wiese R. & Pinker S. (1995) German inflection: The exception that proves the rule. *Cognitive Psychology* 29:189–256. [JLM]

Marcus, G. F., Vijayan, S., Rao, S. B. & Vishton, P. M. (1999) Rule learning in seven-month-old infants. *Science* 283:77–80. [FCG]

Marín, J., Calvo, F. & Valenzuela, J. (2003) The creolization of pidgin: A connectionist exploration. In: *Proceedings of the European Cognitive Science Conference,* ed. F. Schmalhofer, R. M. Young & G. Katz. Erlbaum. [FCG]

Marr, D. (1982) *Vision.* Freeman. [JA, PT, HW]

Massaro, D. W. (1989) Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology* 21:398–421. [JLM]

(1998) *Perceiving talking faces: From speech perception to a behavioral principle.* MIT Press. [PFMJV]

Massaro, D. W. & Cohen, M. M. (1991) Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology* 23:558–614. [JLM]

Matessa, M. (2001) *Simulating adaptive communication.* Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA. [aJRA]

Matessa, M. & Anderson, J. R. (2000) Modeling focused learning in role assignment. *Language and Cognitive Processes* 15:263–92. [aJRA]

Mayr, E. (1988) *Toward a new philosophy of biology: Observations of an evolutionist.* Harvard University Press. [PAMG]

McClelland, J. L. (1979) On time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review* 86:287–330. [aJRA]

(1991) Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology* 23:1–44. [JLM]

(1994) The interaction of nature and nurture in development: A parallel distributed processing perspective. In: *International perspectives on psychological science, vol. 1: Leading themes,* ed. P. Bertelson, P. Eelen & G. D'Ydewalle. Erlbaum. [FCG]

McClelland, J. L. & Chappell, M. (1998) Familiarity breeds differentiation: A Bayesian approach to the effects of experience in recognition memory. *Psychological Review* 105:724–60. [aJRA]

McClelland, J. L. & Elman, J. L. (1986) The TRACE model of speech perception. *Cognitive Psychology* 18:1–86. [JLM]

McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. (1995) Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review* 102:419–57. [aJRA]

McClelland, J. L. & Patterson, K. (2002a) Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences* 6:465–72. [JLM]

(2002b) 'Words or Rules' cannot exploit the regularity in exceptions. *Trends in Cognitive Sciences* 6:464–65. [JLM]

McClelland, J. L. & Plaut, D. C. (1999) Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences* 3:166–68. [aJRA]

McClelland, J. L. & Rumelhart, D. E. (1981) An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review* 88(5):375–407. [aJRA,JLM]

(1986) *Parallel distributed processing. Explorations in the microstructure of cognition, vol. 2.* MIT Press/Bradford. [aJRA]

McCloskey, M. & Cohen, N. J. (1989) Catastrophic interference in connectionist networks: The sequential learning problem. In: *The psychology of learning*

and motivation: Advances in research and theory, vol. 24, ed. G. H. Bower. Academic Press. [aJRA]
Menzel, R. & Muller, U. (1996) Learning and memory in honeybees: From behavior to neural substrate. *Annual Review Neuroscience* 19:379–404. [PFMJV]
Meyer, D. E. & Kieras, D. E. (1997) A computational theory of executive cognitive processes and multiple-task performance. Part 1. Basic mechanisms. *Psychological Review* 104:2–65. [arJRA]
Minsky, M. L. & Papert, S. A. (1969) *Perceptrons.* MIT Press. [aJRA]
Misker, J. M. V. & Anderson, J. R. (2003) Combining optimality theory and a cognitive architecture. In: *Proceedings of the Fifth International Conference on Cognitive Modeling, Bamberg, Germany, April 2003.* [rJRA]
Movellan, J. R. & McClelland, J. L. (2001) The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review* 108(1):113–48. [JLM]
Nagel, T. (1974) What is it like to be a bat? *Philosophical Review* 83:435–51. [MO]
Newcombe, N. (1998) Defining the "radical middle." *Human Development* 41:210–14. [CFO]
Newell, A. (1973) You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In: *Visual information processing,* ed. W. G. Chase. Academic Press. [EMA, arJRA, YY]
(1980) Physical symbol systems. *Cognitive Science* 4:135–83. [arJRA, PAMG, SS]
(1990) *Unified theories of cognition.* Harvard University Press. [arJRA, PAMG, WDG, AGBtM, SS, HW, RMY]
(1992) Précis of *Unified theories of cognition. Behavioral and Brain Sciences* 15:425–92. [aJRA, MLC]
Newell, A. & Card, S. K. (1985) The prospects for psychological science in human-computer interaction. *Human-Computer Interaction* 1(3):209–42. [WDG]
Newell, A. & Simon, H. A. (1963/1995) GPS, a program that simulates human thought. In: *Computers and thought,* ed. E. Feigenbaum, J. Feldman & P. Armer. AAAI Press. [MLC]
(1972) *Human problem solving.* Prentice Hall. [aJRA]
Norman, D. A. & Shallice, T. (1986) Attention to action: Willed and automatic control of behaviour. In: *The design of everyday things,* ed. R. J. Davidson, G. E. Schwartz & D. Shapiro. Doubleday. [MO]
Oaksford, M. & Chater, N., eds. (1998) *Rational models of cognition.* Oxford University Press. [aJRA]
O'Hear, A. (1997) *Beyond evolution: Human nature and the limits of evolutionary explanation.* Clarendon Press. [PAMG]
Ohlsson, S. & Jewett, J. J. (1997) Simulation models and the power law of learning. *Proceedings of the 19th Annual Conference of the Cognitive Science Society.* Erlbaum. [PT]
O'Reilly, R. & Munakata, Y. (2000) *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain.* MIT Press. [FCG, HW]
Overgaard, M. (in press) On the theoretical and methodological foundations for a science of consciousness. *Bulletin from Forum for Antropologisk Psykologi.* [MO]
Palmer-Brown, D., Tepper, J. A. & Powell, H. M. (2002) Connectionist natural language parsing. *Trends in Cognitive Sciences* 6:437–42. [FCG]
Partee, B. (1984) Nominal and temporal anaphora. *Linguistics and Philosophy* 7:243–86. [AGBtM]
(1997) Montague grammar. In: *Handbook of logic and language,* ed. J. van Benthem & A. G. B. ter Meulen. Elsevier Science/MIT Press. [AGBtM]
Pashler, H. (1998) *The psychology of attention.* MIT Press. [aJRA]
Pavlov, I. P. (1928) *Lectures on conditioned reflexes: Twenty-five years of objective study of the higher nervous ability (behavior) of animals.* International Publishers. [PFMJV]
Penrose, R. (1989) *The emperor's new mind: Concerning computers, minds, and the laws of physics.* Oxford University Press. [HW]
(1996) *Shadows of the mind: A search for the missing science of consciousness.* Oxford University Press. [HW]
(1997) *The large, the small and the human mind.* Cambridge University Press. [HW]
Piaget, J. (1967/1971) *Biology and Knowledge.* University of Chicago Press. [PAMG]
Pinker, S. (1991) Rules of language. *Science* 253:530–35. [JLM]
(1994) *The language instinct.* Morrow. [aJRA]
(1997) *How the mind works.* Norton. [IP]
Pinker, S. & Bloom, P. (1990) Natural language and natural selection. *Behavioral and Brain Sciences* 13(4):707–84. [aJRA]
Pinker, S. & Ullman, M. T. (2002a) The past and future of the past tense. *Trends in Cognitive Sciences* 6:456–63. [JLM]
(2002b) Combination and structure, not gradedness, is the issue. *Trends in Cognitive Sciences* 6:472–74. [JLM]
Plaut, D. C. & Booth, J. R. (2000) Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review* 107:786–823. [aJRA]
Plaut, D. C., McClelland, J. L. & Seidenberg, M. S. (1995) Reading exception words and pseudowords: Are two routes really necessary? In: *Connectionist models of memory and language,* ed. J. P. Levy, D. Bairaktaris, J. A. Bullinaria & P. Cairns. UCL Press. [JLM]
Plaut, D. C., McClelland, J. L., Seidenberg, M. S. & Patterson, K. (1996) Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* 103:56–115. [JLM]
Plunkett, K., Karmiloff-Smith, A., Bates, E., Elman, J. L. & Johnson, M. H. (1997) Connectionism and developmental psychology. *Journal of Child Psychology and Psychiatry* 38:53–80. [CFO]
Plunkett, K. & Marchman, V. A. (1991) U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition* 38:43–102. [JLM]
Pollen, D. A. (1999) On the neural correlates of visual perception. *Cerebral Cortex* 9:4–19. [SG]
Pomerleau, D. A. (1991) Efficient training of artificial neural networks for autonomous navigation. *Neural Computation* 3:88–97. [aJRA]
Pomerleau, D. A., Gowdy, J. & Thorpe, C. E. (1991) Combining artificial neural networks and symbolic processing for autonomous robot guidance. *Engineering Applications of Artificial Intelligence* 4:279–85. [aJRA]
Popper, K. R. (1963) *Conjectures and refutations: The growth of scientific knowledge.* Routledge. [NAT]
Prince, A. & Smolensky, P. (1993) Optimality theory: Constraint interaction in generative grammar. *Technical Report CU-CS-696–93,* Department of Computer Science, University of Colorado at Boulder, and *Technical Report TR-2,* Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ. April. [rJRA]
Prudkov, P. (1994) A model of self-organization of cognitive processes. *Cognitive Systems* 4(1):1–19. [PNP]
Pyysiäinen, I. (2003) True fiction: Philosophy and psychology of religious belief. *Philosophical Psychology* 16(1):109–25. [IP]
Qin, Y., Sohn, M-H, Anderson, J. R., Stenger, V. A., Fissell, K., Goode, A. & Carter, C. S. (2003) Predicting the practice effects on the blood oxygenation level-dependent (BOLD) function of fMRI in a symbolic manipulation task. *Proceedings of the National Academy of Sciences USA* 100(8):4951–56. [rJRA]
Quartz, S. R. (1993) Neural networks, nativism, and the plausibility of constructivism. *Cognition* 48:223–42. [SS]
Quinn, R. & Espenschied, K. (1993) Control of a hexapod robot using a biologically inspired neural network. In: *Biological neural networks in invertebrate neuroethology and robotics,* ed. R. Beer et al. Academic Press. [DS]
Raizada, R. & Grossberg, S. (2003) Towards a theory of the laminar architecture of cerebral cortex: Computational clues from the visual system. *Cerebral Cortex* 13:100–13. [SG]
Ramscar, M. (2002) The role of meaning in inflection: Why the past tense does not require a rule. *Cognitive Psychology* 45(1):45–94. [JLM]
Ratcliff, R. (1990) Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review* 97:285–308. [aJRA]
Ratcliff, R., Van Zandt, T & McKoon, G. (1999) Connectionist and diffusion models of reaction time. *Psychological Review* 106:261–300. [aJRA]
Reder, L. M., Nhouyvansivong, A., Schunn, C. D., Ayers, M. S., Angstadt, P. & Hiraki, K. (2000) A mechanistic account of the mirror effect for word frequency: A computational model of remember/know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26:294–320. [aJRA]
Revonsuo, A. & Kampinnen, M., eds. (1994) *Consciousness in philosophy and cognitive neuroscience.* Erlbaum. [PAMG]
Roberts, S. & Pashler, H. (2000) How persuasive is a good fit? A comment on theory testing. *Psychological Review* 107:358–67. [aJRA]
Rogers, R. D. & Monsell, S. (1995) Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General* 124(2):207–31. [EMA, WDG]
Rogers, T. T. & McClelland, J. L. (2003) *Semantic cognition: A parallel distributed processing approach.* MIT Press. [aJRA]
Rolls, E. T. (2000) Memory systems in the brain. *Annual Reviews, Psychology* 51(1):599–630. [IP]
Rolls, E. T. & Treves, A. (1998) *Neural networks and brain function.* Oxford University Press. [FCG]
Roy, A., Govil, S. & Miranda, R. (1997) A neural network learning theory and a polynomial time RBF algorithm. *IEEE Transactions on Neural Networks* 8(6):1301–13. [AR]
Rumelhart, D. E. & McClelland, J. L. (1982) An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect

and some tests and extensions of the model. *Psychological Review* 89:60–94. [aJRA]
(1986a) On learning the past tenses of English verbs. In: *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2,* ed. J. L. McClelland, D. E. Rumelhart & the PDP Research Group. MIT Press. [JLM]
(1986b) PDP models and general issues in cognitive science. In: *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations, vol. 1,* ed. J. L. McClelland, D. E. Rumelhart & The PDP Research Group. MIT Press. [aJRA]
Rumelhart, D. E., McClelland, J. L. & the PDP Research Group (1986) *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. I: Foundations; Vol. II: Psychological and biological models.* MIT Press. [JLM, AR, RS]
Salvucci, D. D. (2001) Predicting the effects of in-car interface use on driver performance: An integrated model approach. *International Journal of Human-Computer Studies* 55:85–107. [rJRA]
Salvucci, D. D. & Anderson, J. R. (2001) Integrating analogical mapping and general problem solving: The path-mapping theory. *Cognitive Science* 25:67–110. [aJRA]
Sanchez-Montanes, M. A., König, P. & Verschure, P. F. M. J. (2002) Learning sensory maps with real-world stimuli in real time using a biophysically realistic learning rule. *IEEE Transactions on Neural Networks* 13:619–32. [PFMJV]
Sanner, S., Anderson J. R., Lebiere C. & Lovett, M. (2000) Achieving efficient and cognitively plausible learning in backgammon. In: *Proceedings of the Seventeenth International Conference on Machine Learning.* Morgan Kaufmann. [aJRA]
Schneider, W. & Oliver, W. L. (1991) An instructable connectionist/control architecture: Using rule-based instructions to accomplish connectionist learning in a human time scale. In: *Architecture for intelligence: The 22nd Carnegie Mellon Symposium on Cognition,* ed. K. VanLehn. Erlbaum. [aJRA]
Schoelles, M. J. (2002) Simulating human users in dynamic environments. Unpublished doctoral dissertation, George Mason University, Fairfax, VA. [WDG]
Schoelles, M. J. & Gray, W. D. (2003) Top-down versus bottom-up control of cognition in a task switching paradigm, *Fifth International Conference on Cognitive Modeling.* Bamberg. [WDG]
Schoppek, W. (2001) The influence of causal interpretation on memory for system states. In: *Proceedings of the 23rd Annual Conference of the Cognitive Science Society,* ed. J. D. Moore & K. Stenning. Erlbaum. [aJRA]
Seidenberg, M. S. & McClelland, J. L. (1989) A distributed, developmental model of word recognition and naming. *Psychological Review* 96:523–68. [JLM]
Sejnowski, T. J. & Rosenberg, C. R. (1987) Parallel networks that learn to pronounce English text. *Complex Systems* 1:145–68. [aJRA]
Shastri, L., Grannes, D., Narayanan, S. & Feldman, J. (2002) A connectionist encoding of parameterized schemas and reactive plans. In: *Hybrid information processing in adaptive autonomous vehicles,* ed. G. K. Kraetzschmar & G. Palm. Springer-Verlag. [RS]
Sherrington, C. (1906) *The integrative action of the nervous system.* Charles Scribner's. [PAMG]
Shirai, Y. & Anderson, R. W. (1995) The acquisition of tense-aspect morphology: A prototype account. *Language* 71:743–62. [JLM]
Siegelmann, H. (1999) *Neural networks and analog computation: Beyond the Turing Limit.* Birkhauser. [YY]
Siegelman, H. T. & Sontag, E. D. (1992) On the computational power of neural nets. In: *Proceedings of the 5th ACM Workshop on Computational Learning Theory.* [aJRA]
Siegler, R. S. (1988) Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General* 117:258–75. [aJRA]
Siegler, R. S. & Lemaire, P. (1997) Older and younger adults' strategy choices in multiplication: Testing predictions of ASCM using the choice/no-choice method. *Journal of Experimental Psychology: General* 126(1):71–92. [WDG]
Siegler, R. S. & Stern, E. (1998) Conscious and unconscious strategy discoveries: A microgenetic analysis. *Journal of Experimental Psychology: General* 127(4):377–97. [WDG]
Simon, L., Greenberg, J., Harmon-Jones, E., Solomon, S., Pyszczynski, T., Arndt, J. & Abend, T. (1997) Terror management and cognitive-experiential self-theory: Evidence that terror management occurs in the experiential system. *Personality and Social Psychology* 72(5):1132–46. [IP]
Simon, T. J. (1998) Computational evidence for the foundations of numerical competence. *Developmental Science* 1:71–78. [aJRA]
(submitted) De-mystifying magical object appearance with a theory of the foundations of numerical competence. *Developmental Science.* [aJRA]
Sirois, S. & Mareschal, D. (2002) Computational approaches to infant habituation. *Trends in Cognitive Sciences* 6:293–98. [SS]
Sirois, S. & Shultz, T. R. (1999) Learning, development, and nativism: Connectionist implications. In: *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society,* ed. M. Hahn & S. C. Stoness. Erlbaum. [SS]
(2003) A connectionist perspective on Piagetian development. In: *Connectionist models of development,* ed. P. Quinlan. Psychology Press. [SS]
Sloman, S. A. (1996) The empirical case for two systems of reasoning. *Psychological Bulletin* 119(1):3–22. [IP]
(1999) Rational versus arational models of thought. In: *The nature of cognition,* ed. R. J. Sternberg. MIT Press. [IP]
Smith, E. R. & DeCoster, J. (2000) Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review* 4(2):108–31. [IP]
Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11:1–74. [aJRA]
Sohn, M.-H. & Anderson, J. R. (2001) Task preparation and task repetition: Two-component model of task switching. *Journal of Experimental Psychology: General* 130:764–78. [EMA]
Sohn, M.-H., Goode, A., Stenger, V. A, Carter, C. S. & Anderson, J. R. (2003). Competition and representation during memory retrieval: Roles of the prefrontal cortex and the posterior parietal cortex, *Proceedings of the National Academy of Sciences USA* 100:7412–17. [rJRA]
Sohn, M.-H., Ursu, S., Anderson, J. R., Stenger, V. A. & Carter, C. S. (2000) The role of prefrontal cortex and posterior parietal cortex in task-switching. *Proceedings of the National Academy of Science* 13:448–53. [aJRA]
Sperber, D. (1997) Intuitive and reflective beliefs. *Mind and Language* 12(1):67–83. [IP]
Squire, L. R. (1992) Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review* 99:195–232. [aJRA]
Suchman, L. A. (1987) *Plans and situated actions: The problem of human-machine communication.* Cambridge University Press. [aJRA]
Sun, R. (1994) *Integrating rules and connectionism for robust commonsense reasoning.* Wiley. [aJRA, IP, RS]
(2002) *Duality of the mind: A bottom-up approach toward cognition.* Erlbaum. [aJRA, IP, RS]
Sun, R. & Bookman, L., eds. (1994) *Computational architectures integrating neural and symbolic processes. A perspective on the state of the art.* Kluwer. [IP, RS]
Sun, R., Merrill, E. & Peterson, T. (2001) From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive Science* 25(2):203–44. [RS]
Taatgen, N. A. (2001) Extending the past-tense debate: A model of the German plural. *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society,* pp. 1018–23. Erlbaum. [aJRA]
(2002) A model of individual differences in skill acquisition in the Kanfer-Ackerman air traffic control task. *Cognitive Systems Research* 3:103–12. [aJRA, NAT]
Taatgen, N. & Anderson, J. R. (2002) Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition* 86:(2)123–55. [arJRA, JLM, AGBtM, PNP]
Taatgen, N. & Dijkstra, M. (2003) Constraints on generalization: Why are past-tense irregularization errors so rare? *Proceedings of the 25th Annual Conference of the Cognitive Science Society.* Erlbaum. [rJRA]
Taatgen, N. A. & Lee, F. J. (2003). Production composition: A simple mechanism to model complex skill acquisition. *Human Factors* 45(1):61–76. [WDG]
ter Meulen, A. (1995) *Representing time in natural language: The dynamic interpretation of tense and aspect.* Bradford Books. [AGBtM]
(2000) Chronoscopes: The dynamic representation of facts and events. In: *Speaking about events,* ed. J. Higginbotham et al. Oxford University Press. [AGBtM]
Tesauro, G. (2002) Programming backgammon using self-teaching neural nets. *Artificial Intelligence* 134:181–99. [aJRA]
Thagard, P. (1992) *Conceptual revolutions.* Princeton University Press. [CFO]
Thelen, E. & Smith, L. B. (1994) *A dynamic systems approach to the development of cognition and action.* MIT Press. [DS]
Thomas, M. S. C. & Karmiloff-Smith, A. (2002) Are developmental disorders like cases of adult brain damage? Implications from connectionist modelling. *Behavioral and Brain Sciences* 25:727–88. [CFO]
(2003) Modelling language acquisition in atypical phenotypes. *Psychological Review* 110. (in press). [CFO]
Tolman, E. C. (1948) Cognitive maps in rats and men. *Psychological Review* 55:189–208. [MO]
Treisman, A. M. & Gelade, G. (1980) A feature integration theory of attention. *Cognitive Psychology* 12:97–136. [MO]
Turing, A. (1950) Computing machinery and intelligence. *Mind* 49:433–60. [NAT]
Ullman, M. T., Corkin, S., Coppola, M., Hicock, G., Growdon, J. H., Koroshetz, W. J. & Pinker, S. (1997) A neural dissociation within language: Evidence that the

mental dictionary is part of declarative memory and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience* 9:266–76. [JLM]

Usher, M. & McClelland, J. L. (2001) On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review* 108:550–92. [JLM]

van Eijck, J. & Kamp, H. (1997) Representing discourse in context. In: *Handbook of logic and language,* ed. J. van Benthem & A. G. B. ter Meulen. Elsevier Science/MIT Press. [AGBtM]

van Rijn, H., Someren, M. & van der Maas, H. (2000) Modeling developmental transitions in ACT-R. Simulating balance scale behavior by symbolic and subsymbolic learning. In: *Proceedings of the 3rd International Conference on Cognitive Modeling,* pp. 226–33. Universal Press. [aJRA]

Vargha-Khadem, F., Watkins, K., Alcock, K., Fletcher, P. & Passingham, R. (1995) Praxic and nonverbal cognitive deficits in a large family with a genetically transmitted speech and language disorder. *Proceedings of the National Academy of Science* 92:930–33. [JLM]

Velmans, M. (1991) Is human information processing conscious? *Behavioral and Brain Sciences* 14(4):651–726. [MO]

Velmans, M., ed. (1996) *The science of consciousness: Psychological, neuropsychological and clinical reviews.* Routledge. [PAMG]

Vere, S. A. (1992) A cognitive process shell. *Behavioral and Brain Sciences* 15:460–61. [aJRA]

Verkuyl, H. (1996) *A theory of aspectuality: The interaction between temporal and atemporal structure.* Cambridge University Press. [AGBtM]

Verschure, P. F. M. J. (1990) Smolensky's theory of mind. *Behavioral and Brain Sciences* 13:407. [PFMJV]

(1992) Taking connectionism seriously: The vague promise of subsymbolism and an alternative. *Proceedings of the 14th Annual Conference of the Cognitive Science Society,* pp. 653–58, Erlbaum. [PFMJV]

(1998) Synthetic epistemology: The acquisition, retention, and expression of knowledge in natural and synthetic systems. In: *Proceedings 1998 IEEE World Conference on Computational Intelligence,* pp. 147–153. IEEE. [PFMJV]

Verschure, P. F. M. J. & Althaus, P. (2003) A real-world rational agent: Unifying old and new AI. *Cognitive Science* 27:561–90. [PFMJV]

Verschure, P. F. M. J., Kröse, B. J. A. & Pfeifer, R. (1992) Distributed adaptive control: The self-organization of structured behavior. *Robotics and Autonomous Systems* 9:181–96.

Verschure, P. F. M. J. & Voegtlin, T. (1998) A bottom-up approach towards the acquisition, retention, and expression of sequential representations: Distributed adaptive control, III. *Neural Networks* 11:1531–49. [PFMJV]

Verschure, P. F. M. J., Wray, J., Sporns, O., Tononi, G. & Edelman, G. M. (1996) Multilevel analysis of classical conditioning in a behaving real world artifact. *Robotics and Autonomous Systems* 16:247–65. [PFMJV]

Voegtlin, T. & Verschure, P. F. M. J. (1999) What can robots tell us about brains? A synthetic approach towards the study of learning and problem solving. *Reviews in the Neurosciences* 10:291–310. [PFMJV]

Waddington, C. H. (1975) *The evolution of an evolutionist.* Edinburgh University Press. [CFO]

Wallach, D. & Lebiere, C. (2000) Learning of event sequences: An architectural approach. In: *Proceedings of the 3rd International Conference on Cognitive Modeling,* ed. N. Taatgen. Universal Press. [aJRA]

(in press) Conscious and unconscious knowledge: Mapping to the symbolic and subsymbolic levels of a hybrid architecture. In: *Attention and implicit learning,* ed. L. Jimenez. John Benjamins. [aJRA]

Weizenbaum, J. (1966) ELIZA – A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery* 9:36–45. [NAT]

Wermter, S. & Sun, R., eds. (2000) *Hybrid neural systems* (Lecture Notes in Artificial Intelligence, LNCS 1778). Springer-Verlag. [RS]

Whitehouse, H. (2002) Modes of religiosity: Towards a cognitive explanation of the sociopolitical dynamics of religion. *Method and Theory in the Study of Religion* 14(3–4):293–315. [IP]

Wise, S. P., Murray, E. A. & Gerfen, C. R. (1996) The frontal cortex–basal ganglia system in primates. *Critical Reviews in Neurobiology* 10:317–56. [aJRA]

Woolf, N. J. & Hameroff, S. R. (2001) A quantum approach to visual consciousness. *Trends in Cognitive Sciences* 5(11):472–78. [HW]

Yang, Y. & Bringsjord, S. (forthcoming) *Mental metalogic: A new, unifying theory of human and machine reasoning.* Erlbaum. [YY]