

To be published in *Behavioral and Brain Sciences* (in press)
© Cambridge University Press 2002

Below is the unedited, uncorrected final draft of a BBS target article that has been accepted for publication. This preprint has been prepared for potential commentators who wish to nominate themselves for formal commentary invitation. Please DO NOT write a commentary until you receive a formal invitation. If you are invited to submit a commentary, a copyedited, corrected version of this paper will be posted.

The Newell Test for a Theory of Mind

-
-
-

John R. Anderson and Christian Lebiere
Carnegie Mellon University

-
-
-
-
-
-
-

Word Count:

Short Abstract: 105
Abstract: 207
Main Text: 13,058
References: 2,961
Entire Text: 16,703

Key Words:

Cognitive Architecture
Connectionism
Hybrid Systems
Language
Learning
Symbolic Systems

-
-

Address Correspondence to:

John R. Anderson
Department of Psychology – BH345D
Carnegie Mellon University
Pittsburgh, PA 15213-3890
Email: 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
Email: cl@cmu.edu
<http://www.andrew.cmu.edu/~cl>

Short Abstract

This paper attempts to advance the issue, raised by Newell, of how cognitive science can avoid being trapped in the study of

disconnected paradigms and mature to provide “the kind of encompassing of its subject matter – the behavior of man – that we all posit as characteristic of a mature science”. To this end we propose the Newell Test that involves measuring theories by how well they do on 12 diverse criteria from his 1980 paper. To illustrate, we evaluate classical connectionism and the ACT-R theory on the basis of these criteria and show how the criteria provide the direction for further development of each theory.

Abstract

Newell (1980, 1990) proposed that cognitive theories be developed trying to satisfy multiple criteria to avoid theoretical myopia. He provided two overlapping lists of 13 criteria that the human cognitive architecture would have to satisfy to be functional. We have distilled these into 12: 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 them to both classical connectionism (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986) 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 derive from its tight integration of the symbolic with the subsymbolic. 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.

1. Introduction

Allen Newell, typically a cheery and optimistic man, often expressed frustration over the 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 became too focused on specific issues and 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. While 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 instance, 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 correspond to the details of human cognition. In addition to behavioral adequacy we would 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 theory of human mind, 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—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 is simply wrong that predicts the correct serial position curve in a particular experiment but also that humans cannot keep track of the situation model implied by a text that they are reading (Ericsson & Kintsch, 1995).

□

So to repeat, we are not proposing that the criteria in Table 1 be the only ones by which a cognitive theory be judged. However, such functional criteria need to be 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 paper will review Newell’s criteria and then consider how they would apply to evaluating various approaches that have been taken to the study of human cognition. This paper will focus on evaluating in detail two approaches. One is classical connectionism as exemplified in publications like McClelland and Rumelhart (1986), Rumelhart and McClelland, (1986) and Elman, Bates, Johnson, Karmiloff-Smith, Parisi, and Plunkett, (1996). The other is our own ACT-R theory. Just to be concrete we will 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 2 pages to describing them and he devoted no more space to them when he redescribed them in his 1990 book. He must have thought that these 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. This section will try to accomplish two things. The first is to make the case that each is a criterion by which all scientific theories of mind 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.

□

2.1 Flexible Behavior.

□

In his 1990 book 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 universality and that it was the most important criterion. He devoted the major portion of that 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 shows anything like this cognitive plasticity.

□

Newell recognized the difficulties he was creating in identifying this capability with formal notions of universal computability. For instance, memory limitations prevent humans from being equivalent to Turing machines (with their infinite tapes) and their frequent slips

prevent people from 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 as 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 to 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 non-natural 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. While there are many complaints about how poorly designed some of these systems are, the artifacts that get into 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 to perform.

□

Grading: If a theory is well specified, it should be relatively straightforward to determine whether it is computational 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 is facing 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 and 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 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 (1990) Soar one can explore whether the mechanism enables behavior that would be functional in the real world. While 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 upon. 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 can 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 criteria 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 versus 5.

□

2.5 Dynamic Behavior.

□

Living in the real world is not like solving a puzzle such as 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 upon 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 test. An appropriate test requires inserting these systems into uncontrolled environments. In this regard, a promising class of tests is to look at cognitive agents, built in these systems, inserted onto real or synthetic environments. For instance, Newell's Soar system successfully simulated pilots in an Air Force mission simulation that involved 5000 agents including human pilots (Jones, Laird, Nielsen, Coulter, Kenny, Koss, 1999).

□

2.6 Knowledge Integration

□

We have chosen to re-title this criterion. Newell rather referred to it as "Symbols and Abstractions" and his only comment on this criterion appeared in his 1990 book "Mind 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 in 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 above 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 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 normal capacity for intellectual combination or else great intellectual discoveries would not be so rare. The system should be able to reproduce the intellectual combinations that people display on a day-to-day basis.

□

2.7 Natural Language.

While most criteria on Newell’s list might be questioned by some, it is hard to imagine anyone arguing that a complete theory of mind need not address natural language. Newell and others have wondered about the degree to which natural language might be the basis of human symbol manipulation versus the degree to which symbol manipulation is the basis for natural language. Newell took the view that it was language that depended 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 labeled aptly enough "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 be its ability to produce these phenomena in a way that explains why they are functional aspects of human cognition.

□

2.9 Learning.

□

Learning seems another uncontroversial criterion for a theory of human cognition. A satisfactory theory of cognition must account for humans' ability to acquire their competences.

□

Grading: It seems rather insufficient to grade a theory simply by asking whether it can learn since 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. These 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 on a cognitive architecture. While in some hypothetical world one might imagine the capabilities associated with cognition emerging full blown, human cognition 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, since human development is not a capability but rather a constraint, there are not common tests of whether 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, 2000) or certain constraints on natural language (e.g., Pinker & Bloom, 1990; Pinker, 1994), 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 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.” (p. 460) This paper 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 modern 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 in their theory of the mind (Newell's original criterion 6 but we have revamped it for evaluation) except as this is just a matter of verbal behavior (Criterion 8). Being cognitively modern, connectionism on the other hand 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 where 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 which can be shown to be computationally equivalent to a Turing machine (Hartley, 2000; Hartley & Szu, 1987; Hornik, Stinchcombe, & White, 1989; Siegelman & Sontag, 1992) and endowed it with learning algorithms that could be shown to be universal function approximators (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 while the Newell and Simon approach emphasized the mental steps directly underlying the performance of a cognitive task. However, both approaches for their different reasons de-emphasized 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 largely focus 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 functionality which some had not credited it with. Thus, the issues were very much a matter of functionality in the spirit of the Newell test.

□

Rumelhart and McClelland's (1986) 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 backpropagation 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, Pearlmutter, and Seidenberg (1996) 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 had concerned their inability to handle the computational complexity of the syntax of natural language (e.g., Bever, Fodor, & Garret, 1968). Elman (1995) used a recurrent network to predict plausible continuations for sentence fragments like *boys who chase dogs see girls* which contained multiple embeddings. This was achieved by essentially having hidden units that encoded states reflecting the past words in the sentence.

□

The discussion above has focused on connectionism's account of natural language because that is where the issues 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 connectionistic 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: feed-forward 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 definition 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 involve assumptions not in the spirit of classical connectionism and often involving non-standard assumptions.

□

4. ACT-R

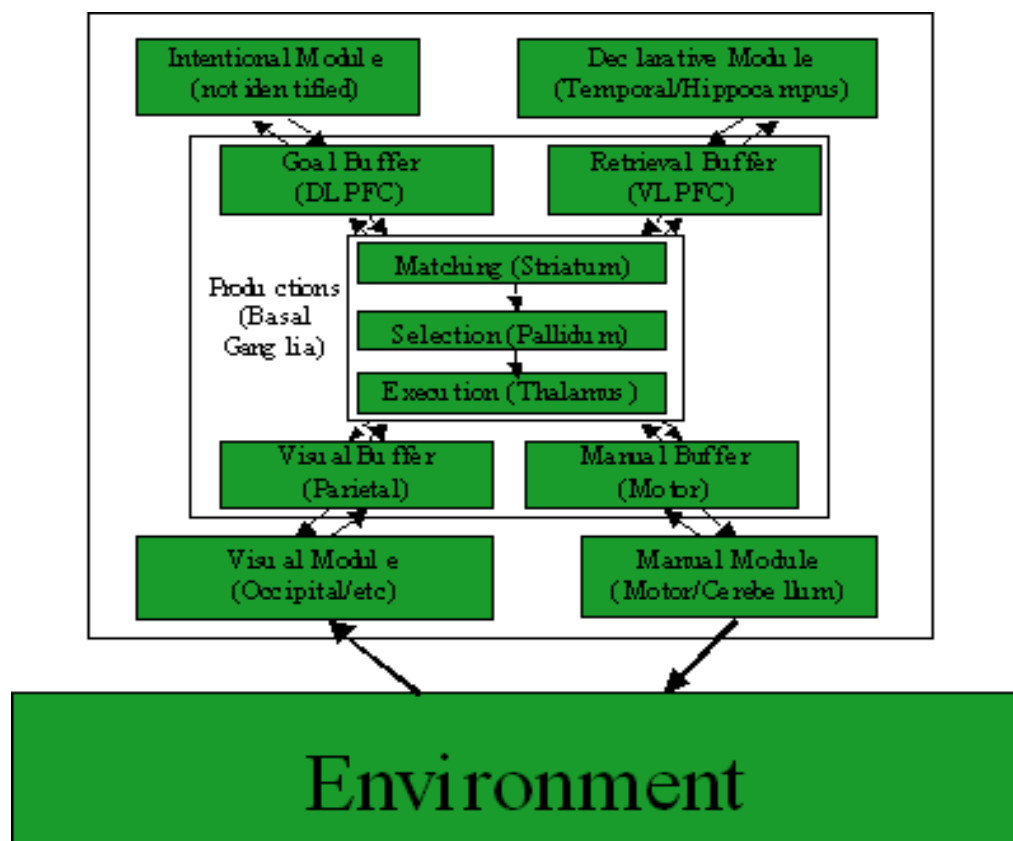
4.1 ACT-R's History of Development

While ACT-R is a theory of cognition rather than a framework of allied efforts, 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 the 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 the 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 second and fifth criteria), 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 (Byrne & Anderson, 1998—PM for perceptual-motor) based in considerable part on the perceptual-motor components of EPIC (Meyer & Kieras, 1997). This paper will focus on what is known as 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 go to Anderson & 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). The components (goal, declarative memory, perceptual, and motor modules) that hold the information ACT-R can access in what are referred to as "buffers" and these buffers



served 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 determining 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 (Paschler, 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 would correspond to associated cortical areas, the current goal to frontal areas, and declarative memory to posterior cortical and hippocampal areas. There is evidence (Wise, Murray, & Gerfen, 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, Loughry, & O'Reilly, 2000) and 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, chapter 4, for examples).
- (2) Added to this base level activation is an associative component that reflects priming 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.
- (3) The activation controlled by factors (1) and (2) is modulated by the degree to which the chunk matches current retrieval specifications. Thus, for instance a chunk that encodes a similar situation to the current one will receive some activation. This partial 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 and 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 is the one selected. This principle combined with variability in activation produces predictions for probability of recall according to the softmax Boltzmann distribution (Ackley, Hinton, Sejnowsky, 1985; Hinton & Sejnowsky, 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, Bothell, Lebiere, & Matessa, 1998; Anderson & Reder, 1999).

□

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 it is added to declarative memory. New production rules are learned on the basis of specializing and merging 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 just a poor approximation to the subsymbolic level as claimed by Rumelhart and McClelland (1986) 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 instance, 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 opposition to classical connectionism except in connectionism's rejection of a symbolic level. While 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 instance, addition facts are represented by one type memory while 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 Figure 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 insure optimal retrieval, it is necessary to "clean" the header. This can be achieved in a number of ways. One would be 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).

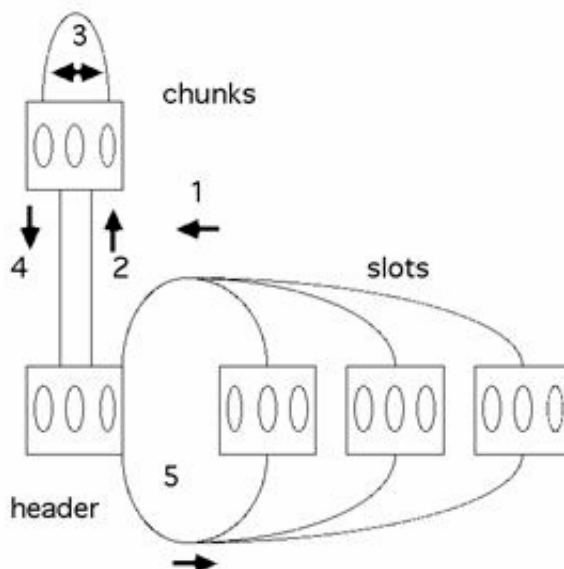


Figure 2: Declarative Memory in ACT-RN

□

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.

ACT-R is a goal-oriented system. To implement this, ACT-RN has a central memory (which probably should be identified with dorsolateral pre-frontal cortex), which at all times contains the current goal chunk (Figure 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 Figure 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 would 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.

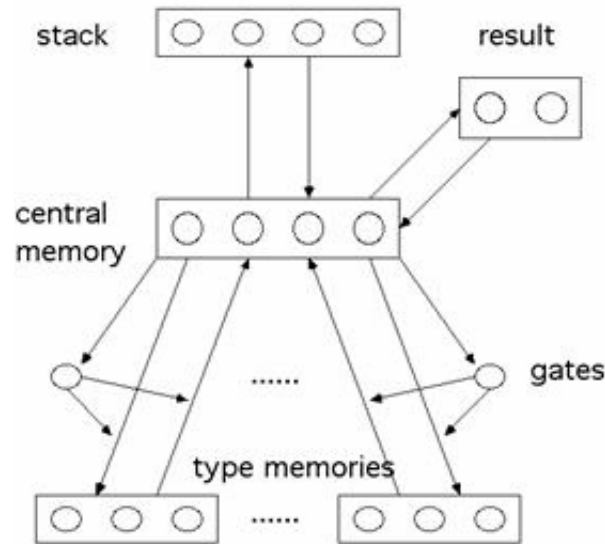


Figure 3: Procedural Memory in ACT-RN

□

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 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-R

Recently, Taatgen (2001; Taatgen & Anderson, submitted) has developed a successful ACT-R model of the 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 since 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 they can 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 has 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 while 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, 1999) and in models for a number of laboratory tasks (Anderson & Betz, 2001; Lerch, Gonzalez, & Lebiere, 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. While 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. Indeed, in another model that basically invents its own past-tense grammar without input from the environment, Taatgen showed that it will develop one or more past-tense rules for low-frequency words but tend 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 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 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, it is also a system with strong limitations. Because of prior constraints on its timing there are strong 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 under 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, Budiu, & Reder, 2001; Budiu and 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 instance, 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 could not be incorporated. For instance, 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 instance, there are no ACT-R models of deductive reasoning tasks. Also within domains that ACT-R has addressed there are important phenomena that ACT-R has not addressed. For instance, while there is an ACT-R model of recognition memory (Anderson, Bothell, Lebiere, & Matessa, 1998) it has not addressed the remember-know distinction (Reder, Nhouyvansivong, Schunn, Ayers, Angstadt, & Hiraki, 2000) or data on latency distributions (Ratcliff, Van Zandt, & McKoon, 1999). It is not clear whether these reflect things that ACT-R researchers just have not addressed or reflect fundamental failings of the theory. For instance, Reder (personal communication) has argued that the failure to address the remember-know distinction reflects the fact that ACT-R cannot address a whole class of metacognitive judgements because it does not have conscious access to its own subsymbolic quantities.

Finally, there is a set of implementation issues acknowledged among researchers in the ACT-R community. We do not want to belabor these because they have an esoteric flavor but just to admit that such things exist we will name a few (and ACT-R researchers will recognize

them): avoiding repeatedly retrieving a chunk since 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 now described Newell's criteria and the two theories, it is 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 is not 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 are 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 are.

□

We had originally proposed a letter-grading scheme for these criteria that we applied to ACT-R. However, we were persuaded in the review process to apply these criteria to classical connectionism by the argument that these 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 along with another and therefore we decided to change these grading into a rough within-theory rank ordering of how well that theory did on that criteria. That is, we will be rating how well that theory has done on that 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 since 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

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 (Botvinck & Plaut, submitted). There is a balance between capability and limitation in classical connectionism but in our opinion and others (e.g., Marcus, 2001) it 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 be 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, Gowdy, & Thorpe, 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 ACT-R. ACT-R, but for its subsymbolic limitations, is Turing equivalent as are most productions systems (and proof for an early version of ACT appears in Anderson, 1976). However, because of variability and memory errors ACT-R will frequently deviate from the prescribed course of its symbolic processing. This shows up, for instance, 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 just at 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 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 main course.

□

5.2. Real-Time Performance

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 instance some relatively arbitrary function of the unit activation is often proposed (e.g., Rumelhart & McClelland, 1981). 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 number of cycle but even when this is adopted it is often arbitrarily scaled (for instance, a linear function of number of cycles with a negative intercept – 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 they 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, in press), 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. They 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 one need not be fitting one experiment or paradigm in isolation.

□

5.3. Adaptive Behavior

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 & Chappell, 1998), to have their underlying learning rules tune the system to the statistical structure of the environment. This is quite core 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) that 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

Connectionism: Worse

ACT-R: Mixed

□

Just because a system works well on small problems one has no guarantee that it will work well on large problems. There have been numerous analyses of the scaling properties of neural networks. In models like Nettek 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, McNaughton, & O'Reilly, 1995) are addressed to 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, 1999) and past-tense learning (Taatgen, 2001b). However, these 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 progress) 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

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). On the other hand, as exemplified in this model, connectionism's conception of the connection between 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 is not a conception that has actually been carried through in models that operate in dynamic environments.

Many ACT-R models have closed the loop particularly in dealing with a 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. On the other hand, 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

Connectionism: Worse

ACT-R: Mixed

We operationalized Newell's symbolic criterion as achieving the intellectual combination that he thought physical symbols were needed for. While ACT-R does use physical symbols more or less in the Newell sense, this does not guarantee that it has the necessary capacity for intellectual combination. There are demonstrations of it making inference (Anderson, Budiu, & Reder, 2001), performing induction (Haverty, Koedinger, Klahr, & Alibali, 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 would find 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 asynchrony. 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 rather to question whether the phenomenon is real.

5.7. Natural Language

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. On the other hand, despite efforts like Elman, 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, 2000b), learning of inflections (Taatgen, in press), 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. On the other hand, 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

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 in the discussion of consciousness and this is implicit memory. 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 have 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

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 more to do. They display rather 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, Bothell, Lebiere, & Matessa, 1998; Anderson & Reder, 1999). 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 one hand, the basic connectionist learning rules have 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 the human these inferences are better understood as more complex causal reasoning (Schoppek, 2001).

□

5.10. Development

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 architecture of the brain and the timing of brain development. Connectionist treatment of development is in some ways less problematic than its treatment of learning because the 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 is starting to emerge in the guise of cognitive models of a number of developmental tasks and phenomena (Emond & Ferrer, 2001; Jones, Ritter, & Wood, 2000; Simon, 1998, submitted; van Rijn, Someren, & van der Maas, 2000; Taatgen, in press). The emerging theory is one that models child cognition in the same architecture as adult cognition and which sees development as just a matter of regular learning. Related to this is an emerging model of individual differences (Lovett, Daily, & Reder, 2000; Jongman & Taatgen, 1999) that relates them to a parameter in ACT-R that controls the ability of association activation to modulate behavior by context. Anderson, Lebiere, Lovett and Reder (1998) argue that development might be accompanied by an increase in this parameter.

□

5.11. Evolution

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. They 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, etc. However, this would be almost as awkward as getting computational universality from a single-tape Turing machine and 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 human cognitive plasticity is totally discontinuous from other species.

□

5.12. Brain

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 non-classical connectionist approaches take these problems as starting points for their efforts.

□

ACT-R has parts of a theory about how it 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 (Sohn, Ursu, Anderson, Stenger & Carter, 2000; Fincham, VanVeen, Carter, Stenger & Anderson, 2002; Anderson, Qin, Sohn, Stenger, & Carter, in press) have confirmed the mapping of ACT-R processes onto specific brain regions (for instance, goal manipulations onto 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 these will be legitimate complaints and we will want to respond either by defending the grading or perhaps conceding an adjustment in the grading. However, the main point of this paper is that 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 buffers and goal buffers in Figure 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 that they even will acknowledge that these 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 with 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, pp. 288).

□

References

- Ackley, D. H., Hinton, G. E., and Sejnowsky, T. J. (1985). A learning algorithm for Boltzmann machines. Cognitive Science, *9*, 147-169.
- Altmann, E. M. & Trafton, J. G. (2002). □Memory for goals: An activation-based model. Cognitive Science, *26*, 39-83.
- Anderson, J. R. (1976). Language, memory, and thought. Hillsdale, NJ: Erlbaum.
- Anderson, J. R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.
- Anderson, J. R. (1990). The adaptive character of thought. Hillsdale, NJ: Erlbaum.
- Anderson, J. R. (1991). Is human cognition adaptive? Behavioral and Brain Sciences, *14*, 471-484.
- Anderson, J. R. (1993). Rules of the Mind. Hillsdale, NJ: Erlbaum.
- Anderson, J. R. (2000). Learning and Memory, Second Edition. New York: Wiley.
- Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. Psychonomic Bulletin and Review, *8*, 629-647.
- Anderson, J. R., Bothell, D., Lebiere, C. & Matessa, M. (1998). An integrated theory of list memory. Journal of Memory and Language, *38*, 341-380.
- Anderson, J. R., Budiu, R., & Reder, L. M. (in press). A theory of sentence memory as part of a general theory of memory. Journal of Memory and Language.
- Anderson, J. R. & Douglass, S. (2001). Tower of Hanoi: Evidence for the Cost of Goal Retrieval. Journal of Experimental Psychology: Learning, Memory, & Cognition, *27*, 1331-1346.
- Anderson, J. R. & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Erlbaum.
- Anderson, J. R., Lebiere, C., Lovett, M. C., & Reder, L. M. (1998). ACT-R: A higher-level account of processing capacity. Behavioral and Brain Sciences, *21*, 831-832.
- Anderson, J. R., Qin, Y., Sohn, M-H., Stenger, V. A. & Carter, C. S. (in press.) An information-processing model of the BOLD response in symbol manipulation tasks. Psychonomic Bulletin & Review.
- Anderson, J. R. & Reder, L. M. (1999). The fan effect: New results and new theories. Journal of Experimental Psychology: General, *128*, 186-197.
- Baddeley, A. D. (1986). Working memory. Oxford: Oxford University Press.
- Bever, T. G., Fodor, J. A., & Garret, M. (1968) A formal limitation of association. In T. R. Dixon & D. L. Horton (Eds.) Verbal behavior and general behavior theory, 582-585. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Bock, K. (1986). Syntactic persistence in language production. Cognitive Psychology, *18*, 355-387.
- Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? Journal of Experimental psychology: General *129*, 177-192.

- Botvinick, M., & Plaut, D. C. (submitted). Doing without schema hierarchies: A recurrent connectionist approach to routine sequential action and its pathologies.
- Brown, R. (1973). A first language. Cambridge, MA: Harvard University Press.
- Browne, A. & Sun, R. (2001). Connectionist inference models. Neural Networks, *14*, 1331-1355.
- Budiu, R. (2001) The role of background knowledge in sentence processing. Doctoral Dissertation, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA.
- Budiu R., & Anderson J. R. (submitted). Interpretation-based processing: a unified theory of semantic processing. Cognitive Science.
- Burzio, L. (1999). Missing players: Phonology and the past-tense debate. Unpublished manuscript, Johns Hopkins University at Baltimore, MD.
- Byrne, M. D. & Anderson, J. R. (1998). Perception and action. In J. R. Anderson, & C. Lebiere (Eds.). The atomic components of thought, 167-200. Mahwah, NJ: Erlbaum.
- Byrne, M. D., & Anderson, J. R. (2001). Serial modules in parallel: The psychological refractory period and perfect time-sharing. Psychological Review, *108*, 847-869.
- Chase, W. G., & Ericsson, K. A. (1982). Skill and working memory. In G. H. Bower (Ed.), The psychology of learning and motivation, (Vol. 16). New York: Academic Press.
- Chomsky, N. A. (1965). Aspects of a theory of syntax. Cambridge, MA: MIT Press.
- Clark, A. (1998). The dynamic challenge. Cognitive Science, *21* (4), 461-481.
- Clark, A. (1999). Where brain, body, and world collide. Journal of Cognitive Systems Research, *1*, 5-17.
- Cleeremans, A. (1993). Mechanisms of Implicit Learning: Connectionist Models of Sequence Processing. Cambridge, MA: MIT Press.
- Cohen, J. D., & Schooler, J. W. (1997). (Eds). Scientific Approaches to Consciousness: 25th Carnegie Symposium on Cognition. Mahwah, NJ: Lawrence Erlbaum Associates.
- Collins, M. (1999). Head-driven statistical models for nature language parsing. Doctoral dissertation, University of Pennsylvania.
- Cosmides, L. & Tooby, J. (2000). The cognitive neuroscience of social reasoning. In M. S. Gazzaniga (Ed.) The new cognitive neurosciences, 2nd Edition, 1259-1272. Cambridge, MA: The MIT Press.
- Dennett, D. C. (1991). Consciousness explained. Boston : Little, Brown and Co.
- 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.
- Dolan, C.P. & Smolensky, P. (1989). Tensor Product Production System: a Modular Architecture and Representation. Connection Science, *1*, 53-68.
- Elman, J. L. (1995). Language as a dynamical system. In R.F. Port & T.v. Gelder (Eds.), Mind as Motion: Explorations in the Dynamics of Cognition (pp. 195-223.): Cambridge, MA: MIT Press.
- Elman, J. L., Bates, E., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). Rethinking innateness: A connectionist perspective on development. Cambridge, MA: MIT Press.
- Emond, B. & Ferres, L. (2001) Modeling the false-belief task : An ACT-R implementation of Wimmer & Perner (1983). Second Bisontine Conference for Conceptual and Linguistic Development in the Child Aged from 1 to 6 Years. Besançon (France), March 21-23 2001.
- Emond, B. (in progress). ACT-R/WN : Towards an implementation of WordNet in the ACT-R cognitive architecture.

- Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102, 211-245.
- Fellbaum, C. (1998) (Ed.). WordNet: An Electronic Lexical Database, MIT Press.
- Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, 25, 348–368.
- 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 National Academy of Science*, 99, 3346-3351.
- Fodor, J. A. (1983). The modularity of mind. Cambridge, MA: MIT/Bradford Books.
- Fodor, J. (2000). The mind doesn't work that way. Cambridge, MA: MIT Press.
- 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
- Gigerenzer, G. (2000). Adaptive thinking: Rationality in the real world. New York: Oxford University Press.
- 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, 581-598.
- Greeno, J. G. (1989). Situations, mental models and generative knowledge. In D. Klahr & K. Kotovsky (Eds.), Complex information processing: The impact of Herbert A. Simon. Hillsdale, NJ: Erlbaum.
- Harnad, S. (1990). The symbol grounding problem. *Physica, D* 42, pp. 335-46.
- Harnad, S. (1994). Computation is just interpretable symbol manipulation; cognition isn't. *Minds and Machines*, 4, no. 379-90.
- Hartley, R. F. (2000). Cognition and the computational power of connectionist networks. *Connection Science*, Vol. 12, No. 2, pp. 95-110.
- Hartley, R., Szu, H. (1987). A comparison of the computational power of neural network models. In Proceedings of the First International Conference on Neural Networks, Vol. 3, pp. 15-22. San Diego, Ca.
- 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), 249-298.
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feed forward networks are universal approximators. *Neural Computation*, 2: 210-215.
- Hinton, G. E. & Sejnowsky, T. J. (1986). Learning and relearning in Boltzmann machines. In Rumelhart, D. E., McClelland, J. L., and the PDP group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations, MIT Press, Cambridge, MA.
- Hopfield, J.J. 1982. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. In Proceedings of the National Academy of Sciences, USA 79, 2554-2558.
- Hummel, J. E., & Holyoak, K. J. (1998). Distributed representations of structure. A theory of analogical access and mapping. *Psychological Review*, 104, 427-466.
-
- Jones, G., Ritter, F. E., & Wood, D. J. (2000). Using a cognitive architecture to examine what develops. *Psychological Science*, 11(2), 1-8.
- 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.

- Jongman, L. & Taatgen, N. A. (1999). An ACT-R model of individual differences in changes in adaptivity due to mental fatigue (pp. 246-251). Proceedings of the 21st annual conference of the cognitive science society. Mahwah, NJ: Erlbaum.
- 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*, 411-428.
- Lave, J. (1988). Cognition in practice: Mind, mathematics, and culture in everyday life. Cambridge: Cambridge University Press.
- Lebiere, C. (1998). The dynamics of cognition: An ACT-R model of cognitive arithmetic. Ph.D. Dissertation. CMU Computer Science Dept Technical Report CMU-CS-98-186. Pittsburgh, PA.
- 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, pp. 635-640.
- 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, pp. 302-307. Mahwah, NJ: Erlbaum.
- Lewis, R. L. (1999, March). Attachment without competition: A race-based model of ambiguity resolution in a limited working memory. Presented at the CUNY Sentence Processing Conference, New York.
- Logan, G. D. (1988). Toward an instance theory of automatization. Psychological Review, *95*, 492-527.
- Lovett, M. C. (1998). Choice. In J. R. Anderson, & C. Lebiere (Eds.). The atomic components of thought, 255-296. Mahwah, NJ: Erlbaum.
- 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.
- MacDonald, M.C., Pearlmutter, N.J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. Psychological Review, *101*, 676-703.
- Magerman, D. (1995). Statistical decision-tree models for parsing. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 276-283.
- Marcus, G. F. (2001). The algebraic mind: Integrating connectionism and cognitive science.
- Matessa, M. (2001). Simulating adaptive communication. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA.
- Matessa, M., & Anderson, J. R. (2000). Modeling Focused Learning in Role Assignment. Language and Cognitive Processes, *15*, 263-292.
- McClelland, J. L. (1979). On time relations of mental processes: An examination of systems of processes in cascade. Psychological Review, *86*, 287-330.
- McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A Bayesian approach to the effects of experience in recognition memory. Psychological Review, *105*.
- McClelland, J. L. & Plaut, D. C. (1999). Does generalization in infant learning implicate abstract algebra-like rules? Trends in Cognitive Sciences, *3*, 166-168
- McClelland, J. L. & Rumelhart, D. E. (1981). An interactive model of context effects in letter perception: I. An account of basic findings. Psychological Review, *88*, 375-407.
- McClelland, J. L., & Rumelhart, D. E., (1986). Parallel distributed processing. Explorations in the microstructure of cognition (Vol. 2). Cambridge, MA: MIT Press/Bradford Books.
- McCloskey, M., & Cohen, N.J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G.H.

Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol.24, pp. 109-165).

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.

Minsky, M. L., & Papert, S.A. (1969). Perceptrons. Cambridge, MA: MIT Press.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium (p. 283-310). In W. G. Chase (Ed.) Visual information processing. New York: Academic Press Inc.

Newell, A. (1980). Physical symbol systems. Cognitive Science, *4*, 135-183.

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

Newell, A. (1992). Précis of Unified theories of cognition. Behavioral and Brain Sciences, *15*, 425-492.

Newell, A., & Simon, H. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall

Oaksford, M. & Chater, N. (Eds.) (1998). Rational models of cognition. Oxford: Oxford University Press.

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

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

Pinker, S. & Bloom, P. (1990). Natural language and natural selection. Behavioral and Brain Sciences *13* (4): 707-784.

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

Pomerleau, D. A. (1991). Efficient training of artificial neural networks for autonomous navigation. Neural Computation, *3*, 88-97.

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.

Ratcliff, R. (1990). Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions. Psychological Review, *97*, 285-308.

Ratcliff, R., Van Zandt, T, & McKoon, G. (1999). Connectionist and Diffusion Models of Reaction Time. Psychological Review, *106*, 261-300.

Reder, L.M., Nhouyvangsivong, 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, & Cognition, *26*, 294-320

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. Psychological Review, *107*, 358-367.

Rogers, T. T. & McClelland, J. L. (2003). Semantic Cognition: A Parallel distributed processing approach. Manuscript in preparation, to appear Spring 2003. MIT Press: Cambridge, MA.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: pt. 2. the contextual enhancement effect and some tests and extensions of the model. Psychological Review, *89*, 60-94.

Rumelhart, D. E., & McClelland, J.L. (1986). PDP models and general issues in cognitive science. In D. E. Rumelhart, J.L. McClelland, & The PDP Research Group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations (pp. 110-146). Cambridge, MA: MIT Press.

Salvucci, D. D., & Anderson, J. R. (2001). Integrating analogical mapping and general problem solving: The path-mapping theory. Cognitive Science, *25*, 67-110.

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 (pp. 823-830). San Francisco: Morgan Kaufmann.

Schneider, W. & Oliver, W. L. (1991). An intractable connectionist/control architecture: Using rule-based instructions to accomplish connectionist learning in a human time scale. In K. VanLehn (Ed.) Architecture for intelligence: The 22nd Carnegie Mellon Symposium on Cognition, 113-146. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Schoppek, W. (2001). The influence of causal interpretation on memory for system states. In J. D. Moore & K. Stenning (Eds.), Proceedings of the 23rd Annual Conference of the Cognitive Science Society, 904-909, Mahwah: Erlbaum.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. Complex Systems, *1*, 145-68.

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, Pittsburgh, pp. 440-449.

Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. Journal of Experimental Psychology: General, *117*, 258-275.

Simon, T. J. (submitted). De-mystifying magical object appearance with a theory of the foundations of numerical competence. Developmental Science.

Simon, T. J. (1998) Computational evidence for the foundations of numerical competence. Developmental Science, *1*, 71-78.

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

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 National Academy of Science. 13448-13453.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. Psychological Review, *99*, 195-232.

Suchman, L. A. (1987). Plans and situated actions: The problem of human-machine communication. Cambridge, Cambridge University Press.

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

Sun, R. (2002). Duality of the Mind: A Bottom-Up Approach Toward Cognition. Mahwah, NJ: Erlbaum.

Taatgen, N. A. (1999). Learning without limits: from problem solving toward a unified theory of learning. Ph.D. Thesis.

Taatgen, N. A. (2001). Extending the Past-tense debate: A model of the German plural. Submitted to the Cognitive Science Conference.

Taatgen, N.A. (2002). A model of individual differences in skill acquisition in the Kanfer-Ackerman Air Traffic Control Task. Cognitive Systems Research, *3*, 103-112.

Taatgen, N., & Anderson, J. R. (submitted). Why do children learn to say "Broke"? A model of learning the past tense without feedback. Cognition.

Taatgen, N.A. (in press). Extending the past-tense debate: a model of the German plural. In Proceedings of the 23rd Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum.

Tesauro, G. (2002). Programming backgammon using self-teaching neural nets. Artificial Intelligence *134*, 181-99.

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. 226-233. Groningen: Universal Press.

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

Wallach, D. (1998). Komplexe Regelungsprozesse: Eine kognitionswissenschaftliche Analyse [Complex system control: A cognitive science approach]. Wiesbaden: Duv.

Wallach, D. & C. Lebiere, C. (2000). Learning of event sequences: An architectural approach. In N. Taatgen (Ed.). In Proceedings of the 3rd International Conference on Cognitive Modeling, pp. 271-282. Groningen: Universal Press.

Wallach, D., & Lebiere, C. (in press). Conscious and unconscious knowledge: Mapping to the symbolic and subsymbolic levels of a hybrid architecture. In Jimenez, L. (Ed.) Attention and implicit learning. Amsterdam, Netherlands: John Benjamins Publishing Company.

Wise, S. P., Murray, E. A., & Gerfen, C. R. (1996). The frontal cortex—Basal Ganglia system in primates. Critical Reviews in Neurobiology, *10*, 317-356.

Footnotes/Endnotes

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.

□

□

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

Acknowledgements

□

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 any stance on anything in this paper.