Conceptions and misconceptions of connectionism

Ron Sun

Department of Computer Engineering and Computer Science (CESC), University of Missouri-Columbia, Columbia, MO 65211.
rsun@cecs.missouri.edu **http://www.cecs.missouri.edu/~rsun**

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

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

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

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

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

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

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

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

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

ACKNOWLEDGMENT

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

Poppering the Newell Test

Niels A. Taatgen

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

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

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

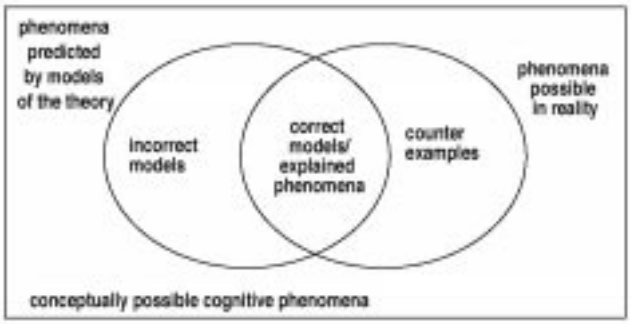

Figure 1 (Taatgen). Diagram to illustrate successes and problems of a theory of cognition.

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

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

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

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

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

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

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

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

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

Cognitive architectures have limited explanatory power

Prasad Tadepalli

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

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

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

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

https://doi.org/10.1017/S0140525X03390132 Published online by Cambridge University Press