Cognitive Architecture as Revolutionary Science

Nicholas L. Cassimatis
Rensselaer Polytechnic Institute
Troy, NY 12183

Paul Bello
Office of Naval Research
Arlington, VA 22203

Many proposals for evaluating cognitive architectures rely partly on an analogy between cognitive science and other sciences. We therefore reflect on the relations between cognitive science, cognitive architectures and other sciences. Some important differences caution against the misguided use of some evaluation approaches. Some unique characteristics of human intelligence motivate the need for new approaches to evaluating cognitive architecture design. We briefly propose one such approach.

In science there is a general expectation that when you do something, you have some way of showing that it has achieved your goal, that it is in some way valid. For example, when a theory generates predictions that other theories do not, then confirming these predictions is evidence it is true. Another way to support a theory is true is to prove theorems about some of its desired characteristics. These approaches successfully used by so many scientific communities that people naturally expect cognitive modelers to adopt them. This is based on a misunderstanding of both science and of the cognitive architecture approach.

First, let us recall the distinction (Kuhn, 1970) between *revolutionary science* and *normal science*. These can be illustrated with an example. Many scientists working within the Newtonian framework accept Newton's conception of space, time, inertia, mass and force and also accepted his three laws regarding the relations between these. Their work involves applying this apparatus to generate explanations or predictions of specific phenomena. They conduct experiments to confirm that the observations their work leads them to expect are indeed observed. If an experiment fails, they generally assume it is a problem with their application of the Newtonian framework instead of a flaw in the framework itself. Khun calls this *normal science*. He calls the project of developing a framework such as Newtonian physics, molecular biology or evolutionary biology *revolutionary science.*

As Kuhn, Quine (1951) and others point out, one does not confirm or disconfirm a scientific framework with one or a small number of experiments. When a particular prediction of a framework fails, one normally first blames their application of the framework. For example, clouds and birds do not obviously obey Newton's laws. Rather than rejecting Newtonian physics, one thinks very carefully about the specific forces involving cloud and bird motion and attempts to conceive of them so that they are consistent with Newton's laws. When scientists accepting Newtoinan physics observed Uranus and Neptune's orbits were not what they expected, they did not reject Newtonian physics but instead posited the existence of a large mass disturbing the orbit. This mass turned out to be Pluto.

These examples illustrate that the observation-hypothesis-experiment caricature of the scientific method does not apply to developing a scientific framework.

It is our thesis that developing cognitive architectures has some of the characteristics of revolutionary science and that developing models within a cognitive architecture has more of the characteristics of normal science. A cognitive architecture provides a basic computational framework for explaining and predicting human cognition. For example, in many architectures, rules, declarative memory and a process for using these to select actions function like the concepts of mass, inertia and Newton's laws do in physics. They provide a basic framework for explaining and predicting a wide range of phenomena. Models of cognition in a particular situation are similar to explanations of specific physical phenomena in that they use the basic apparatus of the conceptual framework without disrupting it.

The fact that cognitive architectures are more like a scientific framework than a specific explanation of a phenomena has several implications. First, just as specific experiments are not sufficient to evaluate scientific frameworks, cognitive architectures cannot be evaluated or compared merely by employing the traditional model fitting or hypothesis testing from experimental psychology. For example, when an ACT-R model makes an incorrect prediction it is more likely a problem with a production rule, parameter or declarative memory element in the model than the architectural framework of ACT-R itself.

How then should cognitive architectures be evaluated? Typical standards used to evaluate other scientific frameworks are: Do they explain and predict a wide range of phenomena that were previously unexplained? Are they

parsimonious? Do they explain phenomena that were previously unexplained? Do they enable predictions of formerly unexpected phenomena? Although it is worth considering whether and how this standard should be manifest in evaluating cognitive architectures, we will focus on some consequences for evaluation the stem from the unique characteristics of cognitive science.

In addition to the overall similarities between cognitive science and other sciences we have reflected on, there is a difference which has important consequences. One of the central goals of cognitive science is to understand, not a specific behavior, but an *ability* to generate a wide range of behaviors that achieve a goal in certain circumstances. Although pendula and the motion of planets were once difficult to explain, it was not because of some general wide-ranging ability these systems had. They behaved in certain ways all the time and one wanted to know why. Unlike most physical objects, cognitive systems generate a wide range of behaviors that are in some sense appropriate for a situation. For example, we need to explain the weight, hardness and color of a rock. We do not need to explain why rocks flatten when they are used to make a road or become sharp when they are used to cut wood. We do not need to explain such behavior because rocks do not exhibit it. They are not intelligent. They do not sense the environment and change their behavior to adapt to it and achieve goals.

Although most biological systems to some extent also display a kind of adaptability, it is normally much more restricted. A cell, for example, can deal with moderate changes in temperature, but it cannot, for example, learn to control fire, wear a coat, move south, etc.

The human ability to observe the world and take actions to achieve goals – we will call this *human intelligence* – is therefore a different kind of phenomenon from those tackled successfully by most sciences. What does this means for architecture evaluation? Consider the following tradeoff. Cognitive Architecture A facilitates models that provide very precise predictions of behavior in many psychological experiments, but does not have the representational or expressive power to support models that have conversations and solve problems in a wide variety of situations. Cognitive Architecture B facilities has been used to create such models, but does not easily support models that make very precise predictions about, say, the exact frequency of a word choice, past tense error rates, etc. A is very likely to have something about it that reflects human cognition, but only B predicts and explains a key datum: humans are intelligence. Although A does a better job of reflecting some specific mechanisms of cognition, which is important, B does embody much more progress towards an important goal of cognitive science.

How, then do we measure a cognitive architectures success at explaining human intelligence? One approach (Bringsjord & Schimanski, 2004) is to use human intelligence tests. We believe that if a single cognitive model could perform well in a wide range of human intelligence tests it would obviously be a huge advance towards the human cognitive ability cognitive architectures are able to model. We do not adopt this approach in near-term research because we fear that monumental, but still intermediate advances towards full human-level intelligence will still perform abysmally on these tests.

Artificial Intelligence, of course, includes several metrics for testing the ability of algorithms to make inferences or solve problems. Success on many of these metrics has been achieved however (e.g., in parsing the Wall-Street Journal corpus) without bringing the field recognizably closer to achieving its ultimate goals (e.g., having computers understand language).

One approach we are exploring is to adopt metrics that are similar in character to those used in AI, but which measure for the cognitive ability we are trying to capture and which require solutions to problems we think occur throughout most domains of cognition.

We will illustrate this with an example from object tracking. The problem is easy to formulate. An agent (human or robot) tracks other agents in an environment where they become occluded. Metrics for this problem are easy to develop. They basically involve the success rate of simple queries of an agent about the location of objects it was tasked to track. This turns out be a very difficult problem (Cassimatis, Bugjaska, Dugas, Murugesan, & Bello, 2007) (because of the size of the state space, incomplete state information and the impossibility of knowing which objects exist in the domain before reasoning begins) that challenges that abilities of the state of the art of inference algorithms. Further, according to the Cognitive Substrate Hypothesis (Cassimatis, 2006) the computational problems that must be solved for cognition in most domains are those that must be solved for the agent tracking problem.

Thus, if we can create a cognitive model that performs very well on this problem, we will not only have advanced the level of ability cognitive architectures are able to model, but also will (according to the Cognitive Substrate Hypothesis) have done this for abilities that are likely key to support a very broad range of cognition.

We believe that benchmarks such as these are complex enough to require important advances in the understanding of human cognitive ability, but are tractable enough in the near term to drive tangible progress.

To conclude, we believe that reflecting on the similarities and differences between cognitive and other sciences is a useful way to think about evaluating cognitive architectures. It suggest that a cognitive architecture is like a scientific framework developed by revolutionary science and is not properly evaluated with straightforward model fitting and hypothesis testing methods from normal science. We believe that cognitive science is unique in the extent to which it aims at an explanation of ability and illustrate an approach to measuring whether cognitive architectures can measure such ability. We believe this approach can, like successful metrics in other parts of AI, drive progress towards cognitive architectures that explain human intelligence.

References

Bringsjord, S., & Schimanski, B. (2004). *"Pulling it all together" vis psychometric AI.* Paper presented at the 2004 Fall Symposium: Achieving Human-Level Intelligence through Integrated Systems and Research.

Cassimatis, N. L. (2006). A Cognitive Substrate for Human-Level Intelligence. *Artificial Intelligence Magazine, 27*(2).

Cassimatis, N. L., Bugjaska, M., Dugas, S., Murugesan, A., & Bello, P. (2007). *An Architecture for Adaptive Algorithmic Hybrids.* Paper presented at the AAAI-07, Vancouver, BC.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions* (2 ed.). Chicago: University of Chicago Press.

Quine, W. V. (1951). Two Dogmas of Empiricism. *Philosophical Review 60, 60*, 20-43.