Symposium on Cognitive Architecture

Participant: Jeff Elman[1] (elman@cogsci.ucsd.edu)
Department of Cognitive Science, University of California, San Diego, La Jolla, CA 92093-0515

Participant: Robert Hadley[2] (hadley@cs.sfu.ca)
School of Computing Science, Simon Fraser University, Burnaby, B.C., V5A 1S6, Canada

Participant: Gary Marcus[3] (gary@psych.nyu.edu)
Department of Psychology, New York University, 6 Washington Place, New York, NY 10003-6634

Symposium Organizer: Robert Hadley

Introduction

Between 1965 and 1982, classically symbolic AI was often touted as the most promising basis for a general theory of Cognitive Architecture. By 1986, however, prominent researchers began to herald the emergence of a new paradigm, Connectionism, which would (they argued) eventually displace the classically symbolic methods dominant in AI and Cognitive Science. Since then, a variety of competing positions have emerged, some purely connectionist, some hybrid, and some purely classical. This symposium examines the present and future prospects for contending paradigms in this realm. A variety of arguments, arising from computational, psychological, biological and philosophical bases, are presented.

1. Connectionism: Whence, Wither, and Why (Elman)

It has now been over a decade since the PDP books came out (though a half-century since Hebb's proposal for associative learning). During the past 10 years a flurry of modeling activity has focused a great deal of attention on core issues about the mechanisms which underly human cognition. Much has been learned, many issues are still in contention, and there is an enormous amount which remains undone.

In this talk I will present what I believe to be some of the core computational principles which distinguish at least one broad class of connectionist models from what has been called the "classical" theory to cognition. This is the "why" part of the talk. I then will discuss what I see as the major insights and accomplishments of this work ("whence"), as well as the major gaps. I close by proposing a program for the future ("whither"). This includes broadening the scope of inquiry and relevant data to include more attention to biology, to the social and affective aspects of cognition, to the relations between behavior and environment, and to the role of evolution.

2. Novel Combinations of Skills: Implications for Cognitive Architecture (Hadley)

In the late 1980s, there were many who heralded the emergence of connectionism as a new paradigm – one which would eventually displace the classically symbolic methods then dominant in AI and Cognitive Science. At present, there remain influential connectionists who continue to defend connectionism as a more realistic paradigm for modeling cognition, at all levels of abstraction, than the classical methods of AI. Not infrequently, one encounters arguments along these lines: given what we know about neurophysiology, it is just not plausible to suppose that our brains are digital computers. Thus, they could not support a classical architecture.

I argue here for a middle ground between connectionism and classicism. I assume, for argument's sake, that some existing or future form of connectionism can provide reasonably approximate models – at least for lower-level cognitive processes. Given this assumption, I argue on theoretical and empirical grounds that most human mental skills reside in *separate* connectionist modules or "sub-networks". In addition, it is shown that humans certainly employ *novel combinations* of skills in rule following and, plausibly, in problem solving. During the course of argument, it emerges that only an architecture with classical structure could support the novel patterns of *information flow* and interaction that would exist among the relevant set of skill modules. Such a classical architecture might very well reside in the abstract levels of a hybrid system whose lower-level modules are purely connectionist.

3. Rethinking Eliminative Connectionism (Marcus)

Humans routinely generalize universally quantified abstract relationships to unfamiliar instances. If we are told "if glork then frum", and "glork", we can infer "frum"; any name that serves as the subject of a sentence can appear as the object of a sentence. These universals are pervasive in language and reasoning. One account of how these universals are generalized holds that humans possess mechanisms that manipulate rules, symbols and variables; an alternative account holds that rules and symbols can be eliminated from scientific theories in favor of descriptions couched in terms of networks of interconnected nodes. Can these "eliminative" connectionist models offer a genuine alternative?

In this talk, after briefly reviewing the empirical literature on universals, I show that contemporary eliminative connectionist models cannot account for how we extend universals to arbitrary items. The argument runs as follows. First, I show that if these models, as currently conceived, were to extend universals to arbitrary instances, they would have to generalize outside the space of training examples. Next, in simulations, I show that two prominent network architectures, the feedforward network and the simple recurrent network, cannot extend universals outside the training space. I then show how this limitation follows from the mathematics that underlies the models. Finally, I show how this limitation can be avoided through the use of an architecture that implements symbol manipulation.