Beyond Marr's Levels: On the equivalence of hierarchically structured representation (abstract)

David Baldwin

November 27, 2006

The history of cognitive science holds many examples of broad architectures. The classical approaches embodied by Newell and Simon - physical symbol systems - postulated that human cognition could be explained solely in terms of production system-like rules operating over symbolic representations of the world. More recent approaches have specified alternative conceptions of how to best explain cognitive phenomena. Connectionism, in all of its various flavors, suggests that the brain is best described in terms of the combined computation of model 'neurons'. Dynamical systems perspectives use sets of coupled differential equations that give rise to complex interactions between parameters of the system and the trajectory though a state space.

In all of these cases, the general approach to claims about cognitive architecture often takes the following form - phenomena X (language/perception-action/development) is best explained by process Y (symbols/dynamics/connectionism) because it offers some additional explanatory power that other systems do not account for.

The important question common to these arguments, and one that is often a source of great confusion, is the (in)equivalence of these various frameworks for understanding cognition - when properties of a particular class of models (e.g., symbol systems) give it greater explanatory power than some other system. A wide variety of such debates exist in recent literature.

Fodor and Pylyshyn (1988) attempted to show that neural nets do not have the properties necessary to model language. They attempt to show that the properties of language, namely its combinatorial structure, cannot be captured by a non-symbolic architecture. They argue not that the connectionist system lacks the ability to represent language - but that any system that does so *must* be implementing a symbol system.

van Gelder (1998) has made broad claims that dynamical systems are a more elegant and powerful way of characterizing a number of cognitive processes. He contends that the dynamical nature of cognition is best expressed in terms of the complex interaction between several component parts. Smith and Samuelson (2003) argue extensively that dynamical systems and connectionist approaches are complementary, but "... these approaches are different in important ways. More specifically, the basic components of connectionist and dynamical system models ... take different views on the nature of knowledge".

One of the largest problems with evaluating these arguments is that each of these different class of systems is that there is no way (other than optimistic intuition) of distinguishing whether 'basic components' are inequivalent. This leads Smith and Samuelson to proclaim "There are many dynamical systems which are not connectionist models (*and share none of their properties*) [and vice versa]. Nonetheless, if one takes the mathematics as defining of sameness of theories, these are theories of the same general class". The 'properties' that

are not common between dynamical systems and connectionist models ostensibly refer to ontological commitments each framework purports to make - Smith and Samuelson go on to list four core areas which they differ (e.g., the claims each make about the nature of developmental change). In other words, dynamicists ascribe importance to time-evolution laws and their properties, while connectionist approaches ascribe importance to the functioning on basic units.

However, Smith also notes the commonality between the mathematics of these frameworks. The basic mathematical tools involved in connectionism and dynamical systems clearly embody these ontological commitments. However the particular 'components' of complex systems (e.g., neurons, differential equations, etc.) give rise to many hierarchical levels of behavior. For example, Elman's (1990) analysis of a recurrent neural net is a common example of this type of abstraction - by analyzing the trajectory of activation in a set of artificial neurons, he showed evidence that his model learned context sensitivity in word meaning. A trajectory is a set of activations of more basic units - it is a higher order abstraction over the places of state space these activations might occupy given some input.

The question we should be interested in, though, is not whether the underlying properties of the trajectory (the component patterns of activation) are embodied in some time (or 'neural'/symbol) dependent system, but whether these ontological (and mathematical) commitments are necessary for explanation of the behavior in question. As Fodor (1998) and others have suggested, simply having additional structure (e.g., time) does not automatically imply that that such structure is necessary to account for a particular phenomenon.

The lack of a common and formal language for examining this question leaves many arguments of the strength of representation between systems poorly defined and misleading. When two machines exhibit hierarchical structure, at which level do we compare their representation and how do we know it is a valid comparison? In what fashion can we systematically claim that additional structure is irrelevant to a system?

In this paper, I will consider this notion, which I will call *representational equivalence* - the process of evaluating the relative strengths of the representation in two systems. First, I will further discuss the notion of hierarchical representation and define some of the formalisms used throughout the paper. I will then review Marr's levels and show how they are an insufficient and often misleading way of characterizing this notion. By reviewing the work of McClamrock (1991), Markman (1998), Chalmers (1996), among others, I attempt to show three primary reasons why using Marr's levels to study representational equivalence leads to confusion. Through this discussion, I will suggest an alternate means to characterize these debates.

1 The Hierarchical Nature of Representation

To further explicate how the components of a system might give rise to higher order structures - as well as give formal definitions of these constructs, it is useful to view an example. Van Gelder (1998) appeals to the Watt governor - a device for regulating the speed of a steam engine. In an example oft repeated by dynamicists, the Watt Governor can be described in several ways. One is via a symbolic computational system that samples the state of the machine at discrete time steps and reacts accordingly. Van Gelder argues a more elegant way of describing the system is via a set of differential equations that govern the states of the system - leading to a coupled set of dynamics that determine the rate of change of system at any point in time.

In terms of state transition - the differential equations of system are not unlike those of

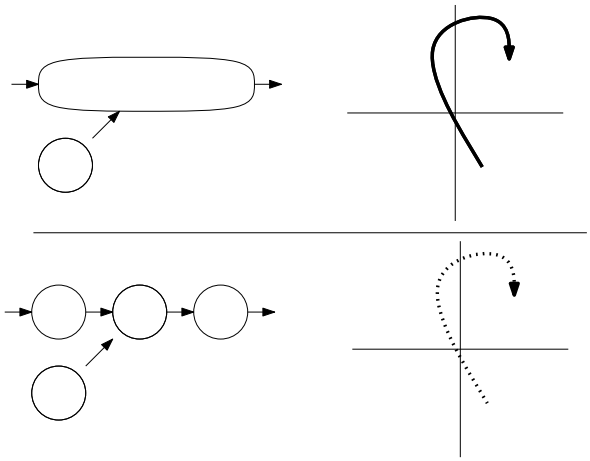

Figure 1: The trajectory as a set of individual state transitions (bottom left) and as a single unified state (top left). The mapping from states to trajectory is a homomorphic, structure preserving one (see text)

a Turing machine or connectionist network . In all these cases, there is a state - a point on the plane, the position of the head and configuration of the tape or activations of artificial neural 'units', respectively. In each state, we deterministically define a successor state (or perhaps a distribution over them, in a probabilistic world) that brings us to the next step in processing. In a continuous system with infinite precision, this set of states may also be infinite.

These sets of states give rise to higher order structures. For example, given an initial state, we can describe the trajectory through Euclidean space for some time interval t - the trajectory consists of the states (points) visited from time 0 to time t as defined by the differential equations. By defining this trajectory we have mapped a set of states in the system to a single, higher order structure that represents the conjunction of these states. Figure 1 shows this graphically.

In doing so, we have created a homomorphic mapping from a subset of states of the plane and their causal transitions (specified by the differential equations) to a single state that encapsulates the processing for that entire set of substates, In making this mapping, however, we ignore the processing (transitions) that the trajectory indicates. In other words, any point starting on or merging with that trajectory will result in a system that resides in the same state as the final point on that trajectory.

We could make these sorts of mapping arbitrarily, and perhaps even ascribe meaning (as a 'word' or something else) to particular trajectories through the space(van Gelder suggests something similar). However, there do exist well defined higher order structures in *every* dynamical system. The mathematics of dynamical systems theory is primarily concerned with describing these structures in a systematic way. For example, a set of points (or equivalently, trajectories) that all converge to the same final state define a basin of attraction. In a system with two attractors, this defines a homomorphic mapping between the entire space of the plane and two *discrete* units.

This distinction is important, so it necessitates some elaboration. The basins of attraction(more formally, the limit sets of the system) are a well defined equivalence class over *any* dynamical system that contains two basins of attraction of the same type - regardless of the differential equations that give rise to them. In moving from a lower level of hierarchical description to a higher one, we have, in effect, broadened the class of systems that might

have the same set of states. More specifically, we have moved from a particular set of differential equations that specify the rate of change at each point(a *unique* set of equations) to a description that is realized by *any* set of differential equations with the same set of basins. In other words, the rate of change at each point in the trajectory does not matter - only that it eventually converges to the fixed point specified by the attractor.