Panels

The Challenge of Neural Darwinism

Stephen W. Smoliar
USC Information Sciences Institute
4676 Admiralty Way Suite 1001
Marina del Rey, California 90292-6695

Summary

The research program of Gerald Edelman's book *Neural Darwinism* addresses the following significant question: How does an agent form categories in a world which is not explicitly labeled in advance? Very early in his text, Edelman argues that what he calls "information processing models" of cognition, such as those of NewelPs Physical Symbol System Hypothesis, tend to fall back on *a priori* assumptions regarding the existence of such labels. He also claims that the design of connectionist networks ultimately rests on similar *a priori* assumptions. As an alternative, Edelman asserts that Darwinian selection among a vast repertoire of connections between nerve cells is an approach through which perceptual categories may be apprehended without assuming any labels in advance.

The first objective of this panel is to discuss whether Edelman's question about category formation without *a priori* labels is relevant to the study of artificial intelligence. Assuming that it *is* relevant, the second objective is to address how the practice of artificial intelligence should respond to it. Linda Smith will consider the Physical Symbol System Hypothseis from the point of view of human development, David Zipser will consider connectionism, and John Holland will speak from his experience with genetic algorithms. Finally, George Reeke will respond to issues raised by the panelists.

Linda B. Smith
Indiana University
Bloomington, IN 47405

The traditional framework for the study of human categorization is the Physical Symbol System Hypothesis. Within this approach, cognition is the manipulation of structured symbolic representations. Symbols gain their meaning via their correspondence to the external world. Within this Physical Symbol System, thought is abstract and objective. The meaning is in the external world that is represented. Thought is disembodied; it is independent of the nervous system.

In recent years, there has been a considerable effort in the empirical study of human categorization in psychology, linguistics, and anthropology. The accumulated evidence is in conflict with the traditional information processing approach. Human categories are embodied, not disembodied. They grow out of a myriad of kinds of bodily experience. They are put together from the operations of the perceptual systems, movement and interactions of various kinds in the physical and social worlds of people. Human categories are fluid and dynamic; they assemble in context to fit the context. Moreover, human categories have developmental histories that directly link categories to the body and the nervous system.

The developmental history of human categories poses serious problems for a Physical Symbol System. The emergence of new forms is the most profound question of ontogenesis. How can we start with a state that is somehow less and get more? The answer offered by the Physical Symbol System Hypothesis is that categories are taught; the categories we possess are the categories that are labelled as categories in the external world. By definition, in a Physical Symbol System, categories are sets of symbols structured to match the structure of the world.

This assumption does not fit the fluid and emergent nature of human categories. Labelled sets of symbols are static and brittle; human categories are dynamic and flexible. Moreover, the Physical Symbol System's solution to development is tcleological. Categories are prescribed by the environment. The end-state, the category to be formed, does all the work. Changes in representations come about when the current representation does not match external reality. The Physical Symbol System Hypothesis thus presupposes what it seeks to explain: the structure of human categories.

Edelman's theory explicitly recognizes the polymorphic, multimodal and dynamic nature of human categorization. Particularly promising aspects of the theory from a developmental point of view are the selection mechanism, re-entry, primary and secondary repertoires, and degenerate representation. Selection means that we do not have to build new representations but only select from the potential structure that is already present in the variant neural structures. Re-entry provides a basis for uninstructed learning. There need be no teacher or explicit match between internal "representations" and external reality. The distinction between primary and secondary repertoires provides a way to incorporate evolutionary history in the developmental process. The degeneracy of the systems means that within the individual there are multiple solutions to single problems and thereby creativity and the emergence of new forms with development and learning.

Is category learning without *a priori* labels a relevant problem? Clearly, the answer is yes. How shall AI meet the challenge? The key may be in looking at emergent structures in the interactions of diversely structured populations of cells.

David Zipser
University of California San Diego
La Jolla, CA 92093

Our current ignorance of cognition is so great that championing one paradigm to the exclusion of others seems premature, at least, and probably folly in the long run. The test of the usefulness of any particular approach is its ability to solve hard problems. Connectionism was, on these grounds, a weak paradigm until the recent development of learning procedures with great power and generality. Now that we can program networks we see that the strengths and weaknesses of neural networks and traditional AI seem to be largely complementary, so the most productive approach is to develop the strengths of each paradigm while trying to identify its weaknesses.

The question "how does an agent form categories in a world which is not explicitly labeled in advance" has been studied since the beginnings of connectionism. It was first addressed by Frank Rosenblatt in the 1950s. Rosenblatt initiated the concept of "competitive learning," in which individual units compete for the right to respond to inputs. The units start life with different, usually randomly chosen, weights; so they will give slightly different responses to input patterns. Any training rule that strengthens the response of a unit to the current pattern can be used. To implement competition, the amount of this strengthening must increase as some steep function of the degree to which a unit is stimulated. The input patterns will become divided among the available neurons as long as the total response strength of any one unit is limited to prevent a single unit from taking over all the patterns. The group of input patterns to which each unit responds generally have features in common and can be considered a natural classification of the input set.

Stephen Grossberg has pioneered another approach to unsupervised categorization using adaptive resonance theory (ART). Each pattern presented to an ART system can be either an example of an existing category or an exemplar of a new category. The decision is made by comparing the pattern to all existing exemplars. If there is a near match the input is put in the matched category. If no match occurs, the pattern is incorporated into the ART system as a new exemplar. ART systems are more complete categorizer systems than simple competitive learning networks.

Neural Darwinism involves elements taken from these two unsupervised learning schemes, often described in somewhat different terms. It seems to be similar to competitive learning in many ways, although groups of neurons replace single cells as the unit of selection. The required random element that is used to bootstrap competitive learning is motivated by analogy to the immune system but implemented in terms of random initial weights. The Darwin demonstration simulation is a more complete system similar in some ways to ART.

I do not see Neural Darwinism as either a totally separate paradigm or as a uniquely new concept. It is not sufficiently powerful all by itself to support an understanding of cognitive computation or neurobiological theory. Neural Darwinism may, however, have some role to play as one of the components in the description of cognitive function and development. The proponents of Neural Darwinism would help us get a better understanding of what this role is if they made more effort to relate Neural Darwinism to other areas of cognitive and neural theory.

John H, Holland
University of Michigan
Ann Arbor, Michigan

Most current AI systems can be assigned to one of two broad classes: The "language-oriented" systems, such as those implementing the Physical Symbol System Hypothesis, and the "stimulus-oriented" systems, such as those investigated by the connectionists. It is important that, for either approach, the input-interface sets the same ultimate limits on the system's powers of categorization. Environmental states that cause the input-interface to generate the same input "message" are indistinguishable; and further processing, however implemented, can only categorize the distinguishable. If a system of either type is computationally complete, with respect to sorting input "messages" into categories, then it has reached the limits of what, categorization can do for it.

In an important, sense, information about the environment, as supplied by the input-interface, always comes with some kind of "labels." These labels may be quite primitive (such as the retinal coordinates of an input neuron) or they may be quite sophisticated (such as labeling a given input image a "chair"). The question, then, is not so much one of *a priori* labeling as it is a question of how primitive the labels are. Stated another way, it is a question of how much "intelligence" the input-interface uses in translating the environment into the input messages processed by the system.

Taking this into account, there are reasons that both the stimulus-oriented and language-oriented approaches should pay close attention to Edelrnan's points about "re-entrant connections.'' For the stimulus-oriented connectionists: (1) A pioneering result of Warren McCulloch and Walter Pitts in automata theory is that most computational routines are impossible for nets without loops. (2) More importantly, internal feedbacks are necessary if the networks are to be able to produce emergent, semi-autonomous internal models that provide predictions and anticipations. This point is closely allied to one made by Donald Hebb in *The Organization of Behavior.*

At the other end of the scale, language-oriented systems are almost always computationally complete because they directly employ some '' universal'' language such as LISP. However, with few exceptions, they are very weak at constructing models based on categories suggested by experience. This is partly the result of using symbols that are pre-defined and close to monolithic and partly the result of designing systems that require inputs ("symbols[1]*) that activate appropriate sections of a high-level interpreter. It is difficult to design such systems so that they can learn using the kind of low-level data supplied by natural environments. The learning mechanisms used for language-oriented systems (such as the "chunking" mechanism of Soar) look much more like compilation than like the origination of new categories.

If a system is to learn to construct plausible internal models, it is essential, I think, that it use experience to extract simple sub-structures (building blocks) that can be combined in a variety of ways to yield competing models. In principle, such a system could yield an "upper" layer that behaves much as described by the Physical Symbol System Hypothesis. However, when it comes to the origination of new hypotheses and models, the upper layer is the servant of the lower layers. Whether one prefers the stimulus-oriented or the language-oriented approach, it seems to me a great risk to ignore processes that construct models by extracting and combining building blocks.

George N. Reeke, Jr.
The Neurosciences Institute
The Rockefeller University
1230 York Avenue
New York, NY 10021

Neural Darwinism is an attempt to account for higher brain function, particularly perception, in a manner that is consistent with the facts of neurobiology, with the unique developmental history of each individual, and with the origins of the nervous system in biological evolution. The application of population thinking to the nervous system has led to the theory of neuronal group selection (TNGS), which proposes that the brain is a selective system functioning in somatic time. According to this theory, the units of selection in the nervous system are groups of interconnected neurons. Selection acts upon preexisting variance during development to generate repertoires of neuronal groups and during experience to strengthen the responses of groups that contribute to behaviors having adaptive value for the organism. Responses of neuronal groups in multiple heterogeneous repertoires are integrated by reentry—a process of ongoing, parallel, recursive signalling among separate maps along ordered anatomical connections.

The TNGS has important implications for connectionism and artificial intelligence. Both of these approaches are in essence based on functionalism, which holds that psychological phenomena are nothing but physical processes that can be adequately described in functional terms independently of the detailed structure and mode of development of the brain. The TNGS, on the contrary, holds that the brain, the phenotype, and the environment are inextricably linked as a result of the experiential history of each organism; accordingly, the brain cannot, be viewed as a computational device operating upon formal representations of information. This view avoids several problems which are introduced by the information processing approach, as detailed in "Real Brains and Artificial Intelligence" *(Daedalus,* Vol. 117, No. 1 (1988)). In summary, the TNGS provides a mechanism by which the nervous system can acquire a working functional organization without predefined categories in the environment, without agreed upon codes or prespecified algorithms, and without an omniscient teacher. A homunculus is never invoked to interpret neural responses as symbols; instead, such responses have meaning only in terms of the behavior they engender.

To test these ideas, we are constructing behaving automata that recognize and associate patterns of sensory input by selective mechanisms. In an approach called synthetic neural modelling, the environment, the phenotype, and the nervous system of such an automaton are integrated into a single computer model. The most developed of these automata, Darwin III, is a sessile "creature" with an eye and a multi-jointed arm having a sense of touch; its environment consists of simple shapes moving on a background; its nervous system consists of some 50,000 cells of 50 different kinds. Darwin 111 can be trained to track moving objects with its eye, to reach out and touch objects with its arm, to categorize objects according to combinations of visual and tactile cues, and to respond in a positive or negative way to such objects depending on previous experience with similar objects. Synthetic neural models give insight into how biological pattern recognizing systems might operate and may provide a basis for the construction of improved pattern recognizing and classifying automata.