STEPHEN GROSSBERG
Wang Professor of Cognitive and Neural Systems
Professor of Mathematics, Psychology, and Biomedical Engineering
Chairman, Department of Cognitive and Neural Systems
Director, Center for Adaptive Systems
steve@bu.edu

My CV and Selected Articles can be downloaded from:
http://www.cns.bu.edu/Profiles/Grossberg

MY INTERESTS AND THEORETICAL METHOD

My work aims at understanding how a brain gives rise to a mind. Said in a more classical way, how can we solve the Mind/Body Problem? Although there has been enormous experimental and theoretical progress on understanding brain or mind, establishing a mechanistic link between them has been very difficult, if only because these two levels of description often seem to be so different. Yet establishing such a linkage between brain and behavior is, I believe, crucial in any mature theory of how a brain or a mind works. Without such a link, the mechanisms of the brain have no functional significance, and the functions of behavior have no mechanistic explanation.

As of this writing, there is a rapidly growing number of models available that can quantitatively simulate the neurophysiologically recorded dynamics of identified nerve cells in known anatomies *and* the behaviors that they control. In this restricted sense, the Mind/Body Problem is at last being understood.

A particular type of theoretical method has been elaborated over the past thirty years with which to link complex behavioral and brain phenomena. The key is to begin with behavioral data, typically scores or even hundreds of parametrically structured behavioral experiments in a particular problem domain. One begins with behavioral data because the brain has evolved in order to achieve *behavioral* success. Any theory that hopes to link brain to behavior thus needs to discover the computational level on which brain dynamics control behavioral success. Behavioral data provide a theorist with invaluable clues about the functional problems that the brain is trying to solve. One works with large amounts of data because otherwise too many seemingly plausible hypotheses cannot be ruled out.

A crucial metatheoretical constraint is to insist upon understanding the behavioral data – which comes to us as static numbers or curves on a page – as the emergent properties of a dynamical process which is taking place moment-by-moment in an individual mind. One also needs to respect the fact that our minds can adapt on their own to changing environmental conditions without being told that these conditions have changed. One thus needs to frontally attack the problem of how an intelligent being can *autonomously adapt to a changing world*. Knowing how to do this, as with many other theoretical

endeavors in science, is presently an art form. There are no known algorithms with which to point the way.

Whenever we have attempted this task in the past, we have resisted every temptation to use homunculi, or else the crucial constraint on *autonomous* adaptation would be violated. The result has regularly been the discovery of new organizational principles and mechanisms, which we have then realized as a minimal model operating according to only locally defined laws that are capable of operating on their own in real time. The remarkable fact is that, when such a behaviorally-derived model has been written down, it has always been interpretable as a neural network. These neural networks have always included known brain mechanisms. The functional interpretation of these mechanisms has, however, often been novel because of the light thrown upon them by the behavioral analysis. The networks have also typically predicted the existence of unknown neural mechanisms, and many of these predictions have been supported by subsequent neurophysiological, anatomical, and even biochemical experiments over the years.

Once this neural connection has been established by a top-down analysis, one can work both top-down from behavior and bottom-up from brain to exert a tremendous amount of conceptual pressure with which to better characterize and refine the model. A fundamental empirical conclusion can be drawn from many experiences of this type; namely, the brain as we know it can be successfully understood as an organ that is designed to achieve successful autonomous adaptation to a changing world. I like to say that, although I am known as one of the founders of the field of neural networks, I have never tried to derive a neural network. They are there because they provide a natural computational framework with which to control autonomous behavioral adaptation to a changing world.

Because of this fact, our work has also been useful in technological applications. Many of the outstanding problems in technology also involve understanding how a system can autonomously adapt to a changing world, and can do so in a way that is synergetic with human understanding. That is why every contribution to understanding biological intelligence provides an opportunity for developing a more humane form of Artificial Intelligence that many of us like to call Natural Intelligence. Quite a few biological neural networks, suitably specialized, are finding their way into applications as artificial neural networks for this reason.

A successful real-time analysis of autonomous adaptive behavior in a changing world often requires that one have knowledge, and even mastery, of several disciplines. For example, it has always proved to be the case that the level of brain organization that computes behavioral success is the network or system level. Does this mean that individual nerve cells, or even smaller components, are unimportant? Not at all! One needs to properly define the individual nerve cells and their interactions in order to correctly define the networks and systems whose interactive, or emergent, properties map onto behavior as we know it. Thus one must be able to freely move between (at least) the three levels of Neuron, Network, and Behavior in order to complete such a theoretical cycle.

Doing this requires that one has a sufficiently powerful theoretical language. The language of mathematics has proved to be the relevant tool, indeed a particular kind of mathematics. All of the self-adapting behavioral and brain systems that I have ever derived are nonlinear feedback systems with large numbers of components operating over multiple spatial and temporal scales. The nonlinearity just means that our minds are not the sum of their parts. The feedback means that interactions occur in both directions within the brain, and between the brain and its environment. The multiple temporal scales are there because, for example, processes like short-term memory activation are faster than the processes of learning and long-term memory. Multiple spatial scales are there because the brain needs to process parts as well as wholes. All of this is easy to say intuitively. But when one needs to work within the tough honesty of mathematics, things are not so easy. Most of the difficulties that people seem to have in understanding what is already theoretically known about such systems derives from a literacy problem in which at least one, but often more than one, of the ingredients of neuron, network, behavior, and nonlinear feedback mathematics are not familiar to them.

A second important metatheoretical constraint derives from the fact that no single step of theoretical derivation can derive a whole brain. One needs to have a method that can evolve with the complexity of the environmental challenges that the model is forced to face. This is accomplished as follows. After introducing a dynamical model of a prescribed set of data, one analyses its behavioral and brain data implications as well as its formal properties. The cycle between intuitive derivation and computational analysis goes on until one finds the most parsimonious and most predictive realization of the organizational principles that one has already discovered. Through this analysis, one can also identify various of the "species-specific variations" of such a prototypical model, and apply them to different types of data. Such a theoretical analysis also discloses the *shape* of the boundary, within the space of data, beyond which the model no longer has explanatory power. The shape of this boundary between the known and the unknown then often clarifies what design principles have been omitted from the previous analyses. The next step is to show how these additional design principles can be incorporated into a more powerful model, which can explain even more behavioral and neural data. In this way, the model undergoes a type of evolution, as it tries to cope behaviorally with environmental constraints of ever increasing subtlety and complexity.

The metatheoretical constraint that comes into view here is an *embedding* constraint; in other words, one needs to be able to embed the previous model into the new model. Otherwise expressed, the previous model needs to be "unlumpable" as it evolves into an increasingly complex "brain". This is a type of *correspondence principle* that places a surprisingly severe test on the adequacy of the previously discovered theoretical principles. Models that fail the embedding constraint tend to come and go with surprisingly rapidity, and do not get integrated into burgeoning theories of ever greater predictive power.

The crucial importance of being able to derive behavioral mechanisms as emergent properties of real-time brain mechanisms, and being able to embed a previous model into a more mature model that is capable of adapting to more complex environments, led me to the name Embedding Fields for my earliest models of brain and behavior that were derived in the 1960's. The word "fields" is a short-hand for the neural network as a computational unit whose interactions generate behavioral emergent properties; the word "embedding" refers to the unlumpability constraint. Many stages of model evolution have occurred since the mid-1960's and all of them have successfully built a foundation for their progeny.