An Investigation of Adaptive Behavior
Towards a Theory of Neocortical Function

Jeff Hawkins
2523 Mardell Way
Mountain View, CA. 94043

Introduction

Objective

The objective of this paper is to describe the studies I wish to undertake. I will also describe the research I have done to date. In the end, I hope to provide a clear picture of what problems I want to solve and how I plan to solve them.

Overview

My interests are focused on the human neocortex. Specifically, I want to understand the neocortex from an information processing or theoretical perspective. I believe that major advances in our understanding of thought processes are possible and can be made with the knowledge and technology we possess today.

When I talk to a stranger about a controversial topic, I start by asking some probing questions to reveal how much our views are in common. This helps me interpret and shape the ensuing conversation. The first section of this paper, titled "Approaches to Understanding Behavior", serves the same purpose. It is a quick way of revealing my opinions on several topics related to the study of human behavior. It has little direct relevance to the major ideas presented in the body of the paper and can be skipped altogether if desired. Reading it first, however, may be useful in interpreting the rest of the paper.

The second and major section of the paper is titled "Towards a Theory of Neocortical Function". Here I present a proposal for a theory of the operation of the neocortex. Some of the ideas I present are well established and non-controversial; others are new and speculative. Together they form a cohesive picture of how the neocortex may operate. My goal is not to convince you that all these ideas are correct, but to illustrate my philosophy of how to understand and study the adaptation of behavior.

I conclude the paper with a review of the major points of my thesis and a brief discussion of my plans for the future.

Approaches to Understanding Behavior

The nature of the brain and mind is controversial. There are sharply divided opinions on how best to study the problem and what kind of solutions to look for. There are so many different approaches and philosophies that it is difficult to classify them. However, I have tried to do so in order to reveal my own preferences and opinions. What follows is a short and admittedly biased overview of the different approaches people have taken in an effort to understand and emulate human behavior. This section should prepare you for the topics to be discussed in the latter part of the paper.

Nay Sayers

At one extreme are those who believe that intelligent behavior cannot be understood by ordinary physics, or perhaps cannot be understood at all. They claim that supernatural or metaphysical forces are at the heart of human nature. (Eccles, 1984; Penfield, 1975) It is also sometimes suggested that it might not be possible for brains to understand brains; such knowledge would be paradoxical, a sort of infinite regression. I reject these ideas. I believe that our ignorance of human behavior is a temporary condition that can be overcome through standard scientific methods. If we have trouble imagining how billions of neurons can produce human behavior, it is only a fault of our imagination.

Artificial intelligence

I place the Artificial Intelligence (AI) researchers at another extreme. AI refers to many different disciplines. As commonly used it includes speech recognition, visual pattern recognition, language processing, expert systems and knowledge representation. Most AI researchers aren't concerned with how the brain works in any physical sense. They believe that by programming digital computers they can create intelligent machines. It has been said that "airplanes don't flap their wings", and presumably therefore intelligent machines don't need to be made with neurons. Fifteen or twenty years ago I might have given my allegiance to AI. The power and adaptability of digital computers makes them a very appealing tool for investigating almost any phenomena. However, progress in most AI disciplines has been extremely slow and there is little reason to expect it to change. This seems to be the opinion of both critics and many proponents of AI (Dreyfus, 1979). Only the expert systems area has had any real success but these programs shed little light on the nature of human intelligence. The biggest contribution of AI research is often overlooked. AI has shown how difficult these problems are to solve on a digital computer, and that is my cue to try other approaches.

Psychology

Psychologists have also modeled thought processes such as memory, language and perception. Since I have no formal training in psychology I am not qualified to evaluate their success. Most of what I have read is rich in ideas, but too abstract to be tested or correlated with what we know

from anatomy and physiology. One of the most influential works has been D.O. Hebb's Organization of Behavior (1949). Hebb ventured a little into the world of neurons and their connections. As you might guess, that is where I think we should be looking for the roots of an information processing theory of intelligence.

Neurophysiology and neuroanatomy

There are many successful information processing theories that have arisen from neuroanatomical and neurophysiological studies. At the "low" end we have the theories of the action potential, the synapse and the modifiable synapse. At a somewhat more complicated level, complete circuits for simple invertebrates and reflex arcs in higher animals have been determined. Higher still are the theories describing the operation of the retina and the cerebellum. But as we approach the neocortex, the site of the most interesting action, successful theories become almost nonexistent. This is not to say that we don't know a lot about how the neocortex is organized and what operations it performs. We do. But we have only the vaguest notion of how the whole system works. In the words of Francis Crick (1979):

"What is conspicuously lacking is a broad framework of ideas within which to interpret all these different approaches. ... It is not that most neurobiologists do not have some general concept of what is going on. The trouble is that the concept is not precisely formulated. Touch it and it crumbles."

Some people have tried to develop such a "framework of ideas", but no one has succeeded. Consider the efforts of neurophysiologists. Their main tool for investigating the neocortex has probably been the single unit recording. With this technique the investigator uses a small probe to measure the electrical activity of a single cortical cell. Hubel and Wiesel used single unit recordings to make their now famous discoveries of the cat's visual cortex. But what have we actually learned from the discovery of complex and hypercomplex cortical cells? We have learned that cells with these receptive fields exist. We don't know how their properties are created, and much more importantly, we don't know why they exist. There is no theory here, only observation.

As one moves away from sensory and motor areas it becomes much more difficult to characterize individual cortical cells. But what if we could? What if we could determine every cortical cell's "receptive field"? We still wouldn't know how the brain works. We still wouldn't be able to build machines to perform similar functions. Physiologists realize these limitations. Discussing the elusive behavior of the majority of cortical cells, Vernon Mountcastle (1984) wrote the following:

"Many observations suggest that at this level the method of single-neuron analysis is inadequate, for the relevant central signals appear embedded within the patterns of activity of populations or neurons in ways still scarcely imaginable but which certainly cannot be reconstructed *post hoc* by study of single members of the populations, one by one. These signals must depend critically upon the dynamic aspects of population activity."

Similar limitations can be argued for neuroanatomical investigations. Mostly via labeling techniques we can determine the projections and connections of cortical cells. I believe these studies will eventually prove extremely useful, but today they provide very few clues as to how the neocortex works. If we were given a complete "wiring diagram" of a brain, what would it tell us? Without a theoretical framework it would tell us little.

Of course new anatomical and physiological techniques are always being created. But using the techniques available today there is little reason to expect major breakthroughs in our understanding of the neocortex in the near future. I think this view is close to mainstream opinion.

Neural net modeling

Many people have tried to discover a framework of higher brain function by modeling networks of neurons. This has been a rocky road to follow, characterized by a sort of boom-bust cycle. However, in a limited sense I believe this is the right way to proceed, so I will describe some of the trends as I see them.

One of the first forays into neural nets was started with McCulloch and Pitts' paper A Logical Calculus of the Ideas Immanent In Nervous Activity (1943). This paper introduced the notion of modeling neurons as two-state synchronized logic devices. It was shown that any definable state machine could be implemented with these simplified neurons. As best as I can tell, interest in state machines as being potentially useful for the study of brains continued into the late 1960's.. Gradually these studies became more abstract (Arbib, 1969) and most practical applications centered on the emerging field of digital electronics. There are two basic reasons why McCulloch and Pitts' model failed as a paradigm for nervous activity. The first is that nervous systems are not rigidly synchronized* . The second is that the design of state machines traditionally starts with a complete definition of states and their transitions. Brains are not built this way. The number of states in a brain is very large and they are undefinable. It may be possible to model the neocortex as a state machine if it is unsynchronized and if you concern yourself with the limitations of states, not the actual states.

Another chapter in the history of neural network models focused on the perceptron. Perceptrons are pattern recognition devices. In their simplest form they consist of a two dimensional array of "receptors" which innervate additional layers of "neurons". Perceptrons are characterized by random connections, no feedback, and reinforcement training algorithms. It was hoped that these devices could be trained to recognize images projected onto the array of receptors. There was a surge of interest in perceptrons which

---

* State machines do not have to be synchronized. However, non-synchronized state machines are prone to indeterminate state transition (Torng, 1972).

led to wildly optimistic predictions about their capabilities. In 1969, Marvin Minsky and Seymour Papert (1972 2nd ed.) published a very critical book on perceptrons. This book, combined with the rise of interest in AI (led in large part by Minsky) and the death of Frank Rosenblatt (the main proponent of perceptrons), quickly deflated the interest in perceptrons and perhaps neural nets in general.* Perceptron theory probably deserved this fate. Perceptrons had many flaws, the biggest of which was the complete lack of feedback. It seems though that a general plague descended upon many neural net investigations, and this might not have been warranted.

I find it much more difficult to characterize the research into neural modeling that has occurred during the last fifteen years. This may be due to a general lack of research interest in this area or to a shift of interest toward invertebrate nervous systems where less controversial results can be obtained. In any case, I think it is fair to say that we are still without any usable theories of how neurons in the cerebral cortex interact to produce adaptive behavior, and that progress toward this end has been very slow.

Recently two articles on neural networks have appeared in the popular magazines Business Week (June 2.1986) and Science 86 (May 1986). Both of these articles are in large part based on the work of John Hopfield (1982), and particularly on his use of neural nets to solve a well known mathematical problem, "the traveling salesperson problem" (Hopfield, 1985). It is too early to know if Hopfield represents the leading edge of another surge of interest in neural nets. If it is, I hope that it is characterized by conservative optimism and critical self-appraisal. In terms of solving certain mathematical problems, Hopfield's work may be a breakthrough; in terms of understanding thought processes it is a only a small piece of the puzzle. Keeping this foremost in mind will help prevent another "boom-bust" cycle of enthusiasm.

Summary

The fundamental mechanism of human intelligence has completely eluded us. Our past efforts have been sharply divided and marred by periods of unfulfilled optimism. Many people feel that the future for understanding behavior appears equally confused. I don't believe so.

In the next section I describe how I believe the future will unfold. I can't of course predict the future. I can, however, point out the key observations that I feel have been overlooked, and from them speculate on the nature of a "solution" to the mystery of human behavior.

---

* It is interesting to note that AI theorists have not suffered the same embarrassments as the perceptron theorists, although they too have made wild and unfulfilled predictions.

Towards a Theory of NeocorticaI Function

Solving any problem is greatly aided by knowing where to look for the solution and knowing that a solution actually does exist.

I learned this lesson as a child. I was brought up in an old New England house which had three floors, a basement, and a detached garage, all of which I visited often. Being a typical child I would misplace things, let's say a book, and then have trouble finding it. Quickly, I would search each room, race up and down the stairs, into the basement, and out to the garage. I would make two, three, even four complete circuits before my frustration would cause me to ask for help. Then my mother might say, "I saw the book in the living room." With this clue, I would find the book very quickly, usually in less time than I already spent in total in the living room. This was amazing to me. The effect of knowing what room in which to look was very much greater than just the reduction of area to search. It made it easy to solve a problem that seemed unsolvable only minutes before.

Solving the riddle of human behavior is similar. There are too many places to look, too many approaches, and too many theories. So much uncertainty discourages people from really trying to solve the problem.

There are a great number of people who are interested in how the brain works, but relatively few who study the brain on a regular basis. Of these, only a small fraction are actively searching for unifying theories of neural information processing. Yet the day we start to understand and build intelligent machines, there will be an unprecedented flood of new interest in the field. What is keeping this interest at bay is that people don't know where to start. It is unclear which approaches will be successful and which will not.

I think I know where to start. What follows is a collection of observations and proposals. Together they form a set of theoretical constraints on an information processing theory of the neocortex, the "correct room to look in" if you will. Knowing where to look and knowing what kind of solution to look for gives me confidence and optimism that this problem can be solved.

Here then are the major points of this paper.. The order in which I present them roughly corresponds to my confidence in their veracity; the latter ideas are somewhat more speculative.

1) Adaptation is based on consistencies or "patterns" in environmental stimuli.
2) The neocortex is the primary adaptive organ of the human nervous system.
3) The neocortex consists of millions of nearly identical units.
4) The algorithm implemented by the cortical units works independent of afferent modality.
5) A dominant feature of human behavior is prediction of environmental stimuli.
6) Each cortical unit makes predictions about its own environment.

8) Through cortico-fugal projections, the neocortex makes its environment more consistent.
9) Perception is the collective behavior of many cortical units simultaneously making correct or incorrect predictions.
10) The reticular system directs attention guided by incorrect predictions of cortical units.
11) lntracortical connections define limits to patterns which can be recognized and produced.

Following is the development and support for each of these ideas.

1) Adaptation is based on consistencies or "patterns" in environmental stimuli.

The world in which we live is very consistent. Generally things behave the same way from moment to moment, day to day and year to year. The laws of physics don't change. The floorplan of my house doesn't change. Every day humans continue to have two eyes, one nose and one mouth. The definitions of the keys I am typing on are the same as when I first learned to type. Our environment contains many consistencies and patterns. These are essential to the usefulness of any behavior.

In fact, the behavior of all plants and animals is nothing more than the exploitation of the patterns in an organism's environment. An oak tree sends roots into the ground because throughout the evolutionary development of trees, water and minerals were consistently found in the soil. Bees sting predators because their venom consistently causes the predator to withdraw. Behaviors evolve because the processes of evolution and natural selection automatically seek out and exploit consistencies in the environment of the organism.

Throughout most of the history of life on our planet, evolution was the only mechanism for adapting behavior to take advantage of environmental patterns. Evolution is, of course, a slow process and can therefore only successfully adapt to environmental patterns which persist over many generations.

The invention of plastic nervous systems has dramatically changed the rate at which an organism can adapt to its environment. A dog can learn to avoid walking on an electrified grate in only a few seconds. Fundamentally, however, the process of adaptation exhibited by a nervous system is similar to that which occurs via evolution. There is a consistent pattern in the dog's environment: i.e. walk on grate - get a shock. The animal's behavior changes to exploit that pattern. If electric grates were a common terrestrial feature for the last several million years, animals would have evolved the instinctive behavior to avoid them just as they have evolved the instinctive behavior to avoid their natural predators.

The point of all this is that the human brain is continually exposed to a stream of inputs representing various features of its environment. The only way that the brain can adapt to its environment is by finding and

exploiting patterns in this stream of inputs. If there are no consistencies or patterns, then-adaptation is impossible.

You might find it useful to visualize a brain isolated from its sensory organs. The environment of such a brain consists only of action potentials from afferent fibers. We can consider that these action potentials don't represent light, sound, pressure or anything else in particular; they are all there is. Yet there are consistent temporal and spatial associations in the firing of these afferent fibers, and the brain adapts to them.

What makes the human brain so special is the complexity of the patterns it can recognize. Our brains can adapt to very obscure patterns hidden in input received over long periods of time. Understanding language, appreciating music, and practicing science all require the recognition of subtle associations. Indeed, the tests we use to measure intelligence consist of sequences of figures which have a non-obvious association. The better you are at finding patterns, the better you score.

My primary approach to developing a theory of adaptive brain function is to continually think of brains as recognizing and adapting to environmental patterns, because that is all that the brain has to work with. The more you can explain in terms of generic stimuli and adaptation, the better the theory.

2) The neocortex is the primary adaptive organ of the human nervous system.

This statement should not need too much explanation or justification. The most distinguishing feature of the human brain is the size of its neocortex. It is known to be intimately involved with almost all high-level human capabilities, and it is generally believed to be the site of most learned behavior.

One might take issue with my implicit assumption that the neocortex can be understood as a semi-independent organ. It does, after all, connect to many other areas of the brain, and therefore it may be impossible to gain any insight into the operation of the neocortex as an independent entity. I don't believe this is a valid concern. Later I will show how most of the connections of the neocortex can be interpreted within a single paradigm of adaptation of behavior.

Also, plasticity of function has been clearly demonstrated in other areas such as the hippocampus and cerebellum, but it has not been clearly demonstrated in the neocortex. This, however, is almost certainly a shortcoming of laboratory technique. It is generally believed that most of the adaptation constituting language, vision and other mental processes occurs in the neocortex.

So I will proceed with the assumption that the neocortex can be understood

as a semi-independent organ, the function of which has great relevance to the majority of learned human behavior.

3) The neocortex consists of mi1lions of nearly identical units.
4) The algorithm implemented by the cortical units works independent of afferent modality.

In the introduction to his book Vision, David Marr (1982) proposed the need for different "levels of explanation" to understand human behavior, and vision in particular.

"Almost never can a complex system of any kind be understood as a simple extrapolation from the properties of its elementary components. Consider, for example. some gas in a bottle. A description of thermodynamic effects ... is not formulated by using a large set of equations, one for each of the particles involved. Such effects are described at their own level, that of an enormous collection of particles; the effort is to show that in principle the microscopic and macroscopic descriptions are consistent with one another. If one hopes to achieve a full understanding of a system as complicated as a nervous system. .. then one must be prepared to contemplate different kinds of explanation at different levels of description that are linked, at least in principle, into a cohesive whole."

The approach embraced by Marr is to first understand vision on its own terms, and later see how this could be implemented with neurons.

In search of a high level understanding of vision, Marr discusses contours, edges, surface textures and contrasts, He then combines these elements to synthesize images, $2^{1/2}$-D sketches and the representation of form. This kind of approach was not originated by Marr, nor is it limited to the study of vision. People who investigate language talk of parsers, syntax checkers and formants. Those who study motor control speak of limb trajectories and motor programs. Their common assumption is that human behavior can be partitioned into different functions, each of which can be understood independently and with algorithms specific to the area of study.

This kind of analysis makes sense as long as you restrict yourself to subcortical structures and particularly the sense organs. The retina really does process visual edges, contrasts and motion, and is best understood in those terms. But when you reach the cerebral hemispheres, and in particular the neocortex, this is no longer a successful way to understand vision.

I too believe that a layer of understanding above the neuron is needed. But Marr looked for the wrong layer. He assumed that it had to do with vision. It has nothing to do with vision or hearing or touch or motor control The algorithm of the neocortex is much more generic. It has to do with recognizing patterns and predicting what is going to happen next. The algorithm would work with any kind of sensory organ in any environment, as long as the environment contains consistent patterns and associations

that can be detected by the sense organs.*

I am not suggesting that the neocortex is an amorphous blob in which the details of internal connections are inconsequential. Its patterns of connectivity are specific to the types of environmental associations it is likely to encounter. An area involved in vision will be "wired" differently than an area involved with speech, but both areas operate on the same basic algorithm.

The anatomy of the neocortex supports this view of a single cortical algorithm. Despite its dominant role in the great variety of behaviors we consider uniquely human, the neocortex is surprisingly uniform throughout its entire area. Vernon Mountcastle wrote at length about the evolution and uniformity of the neocortex when he introduced the concept of the cortical column. While agreeing that the neocortex is not completely uniform, he argued for a single cortical algorithm (Mountcastle, 1978).

"I conclude that cytoarchitectural differences between areas of neocortex reflect differences in their patterns of extrinsic connections. These patterns are in no way accidental. They are detailed and precise for each area; indeed, they define it. The traditional or usual 'functions' of different areas also reflect these differences in extrinsic connections; they provide no evidence whatsoever for differences in *intrinsic* structure or function. This suggests that neocortex is everywhere functionally much more uniform than hitherto supposed and that its avalanching enlargement in mammals and particularly in primates has been accomplished by replication of a basic neural module without the appearance of wholly new neuron types or qualitatively different modes of intrinsic organization."

Mountcastle presents a compelling argument for a single cortical algorithm. He also proposed a neural unit which is replicated millions of times over the surface of the neocortex. What makes one unit different from another is the connections it makes with other areas of the neocortex and the rest of the brain; in all other ways they are the same.

Mountcastle proposed that the physical neocortical unit is a column which is perpendicular to the cortex surface, 30-500um in diameter, containing several hundred neurons. Later I will propose the need for another unit in addition to the cortical column. In the meantime I will use the term

---

*There is nothing which precludes the approach advocated by Marr from working. It may be possible to design and build machines (using very specific algorithms) which recognize images, understand speech or solve problems in mathematics. Many people have tried, and some have achieved limited success. But these machines do not work like a brain, nor does it appear likely they will ever work as well as a brain. Evolution has produced a machine which through adaptation can far out perform anything we can currently engineer. Therefore, if you want to build a machine which performs as well as a brain then you have to build a brain and train it. In the future we won't *design* machines which can see or design machines which can understand speech; instead we will design machines which, if attached to a pair of eyes, will *learn* to see or, if attached to a pair of ears, will *learn* to listen.

"cortical unit"; it will still be appropriate if you visualize this as a cortical column.

Interestingly, Mountcastle did not suggest. what the cortical unit does. He only said that "elucidation" of its function will be very useful.

"Put shortly, there is nothing intrinsically motor about the motor cortex, nor sensory about the sensory cortex. Thus the elucidation of the mode of operation of the local modular circuit anywhere in the neocortex will be of great. generalizing significance."

I have already stated that the function of the neocortex. is involved with recognizing patterns and making predictions about its. environment. I believe that this concept also applies to each individual cortical unit. I will elaborate on these ideas later.

Thus the two ideas of this section, that the neocortex is composed of millions of identical units, and that the function of these units is independent of afferent modality, are supported by the same anatomical evidence. If the reader remains skeptical of these conclusions, I recommend reading Mountcastle's excellent essay.

5) A dominant feature of human behavior is prediction of environmental stimuli.

Scientists and engineers who study vision are always asking, "How do we visually recognize an object like a chair?" Similarly, people ask, "How do we recognize a spoken word, someone's voice, or an object placed in our hand?" These are extremely difficult problems, which haven't been solved even for simple objects in controlled environments. I propose that we ask a slightly different question. Instead of asking, "How do we recognize something?", let's ask "How do we recognize something is different?"

When we interact with a familiar object we can notice slight changes in any of a thousand different things. Imagine coming home some day and finding a different door on your house. The new door is identical with the old door except that the top is rounded. You would notice this difference instantly; not because it doesn't look like a door, it does; and not because the motions you use to open the door wouldn't work, they would. Nevertheless, you would notice that the door had changed, and you would stop to consider how this could have happened. The recognition that something is different must correspond to a neural event, either an increase, decrease or some other change in the firing rate of neurons.

As you would notice the top of the door is round, you would also notice if it opened on the wrong side, or if it opened in the wrong direction, had different dimensions, was a different color, had a different style doorknob, was tilted in any direction, was raised above the sill, gave you an electric shock, had a window in the middle, made an unusual sound, or weighed 1000 lbs. In the one or two seconds you normally spend opening the door and entering your house you can notice any of thousands of potential

changes to your door. Noticing any one of these must correspond to some neural event.

The more often you are exposed to a particular relationship of stimuli, the more likely you are to notice a change in that relationship. Because you experience your own door every day you would notice any slight change in its appearance. If however, you were opening a door on a house you never visited before, any "reasonable" door would go unnoticed. Some patterns in our environment are so consistent that if they ever change we find it disturbing. For example, faces with two eyes are an extremely common association. When you encounter a forehead you normally see two eyes below it. If you saw someone with three eyes, loud bells would go off in your head indicating that something is wrong. You would find it difficult not to stare and puzzle at this oddity.

Of course the potential for noticing a change in a familiar object does not occur only while opening your front door or looking at faces. It occurs during every second of your waking life and involves all of your sensory pathways.

Think about this phenomenon carefully. Look around you right now and notice all of the familiar relationships in your immediate environment. Imagine how you would react if any of these relationships changed. Imagine if the ink in your pen changed color, or a friend sitting next to you suddenly had a different voice, or this paper felt sticky, or the edges of a book were no longer parallel. All these things would catch your attention. The possibilities for noticing changes in your environment are nearly endless. How can we explain this phenomenon? Remember that noticing a change brings your attention to that change and therefore must correspond to some neural event. When familiar environmental associations have not changed, this neural event does not occur.

There is only one explanation for this phenomenon. The brain is continually making predictions about what is going to happen next in its environment. It predicts that environmental associations that have consistently occurred in the past will occur again in the future. Therefore, as you approach your house your brain predicts it will see a door. As you near the door your brain makes detailed predictions that it will have the features it has had on previous occasions.

If your door does not have the features your brain predicts, the sensory input conflicts with the internal predictions. A substantially incorrect prediction causes our attention to be focused on the area where the incorrect prediction occurred.

The prediction of environmental stimuli is a dominant function of the neocortex.

These observations may seem so obvious and everyday that I am afraid you

will glance over them. Please don't. They tell us something very real and important about the fundamental operation of our brains. The prediction of environmental associations is not just one attribute of learned behavior, it is a primary attribute.

6) Each cortical unit makes predictions about its own environment.

Now I will address the question, "What is the function implemented by each cortical unit?" (or if you prefer, each "cortical column"). We should be able to figure this out. After all, due to its limited size, the unit can't be too complicated and there are millions of copies in every brain. But most importantly, the function of the cortical unit must somehow be intimately connected to all of human behavior.

I propose that each cortical unit recognizes associations in its own environment and adapts to predict these associations.

Each cortical unit can only make predictions about the inputs it receives, i.e. its "environment". Units of the neocortex whose inputs are dominated by sensory afferents naturally make predictions about "low-level" sensory events. A cortical unit far removed from sensory and motor areas can only make predictions about correlations of the inputs it receives; since these inputs may arise from other areas which are widely dispersed and poorly characterized, it is unlikely that we can ever describe the function of such a "remote" cortical unit. In other words, the environments of some cortical units correspond closely to the environment of our senses. The environments of other cortical units are completely embedded within the cortex, are much more difficult to define and actually change with experience.

The operation of the neocortex as a whole must be understood as a cooperation of many cortical units operating simultaneously. (I am not advocating a variant of the "Grandmother cell" theory.) For example, recognition that an object is a chair requires that many cortical units are simultaneously making correct predictions. Together they form a semi-stable state of cortical activity. On the other hand, confusion occurs when many cortical units simultaneously make incorrect predictions. The gestalt "Aha" is the sudden transition from many incorrect predictions (an unstable condition) to many correct predictions (a more stable condition). Much of human behavior can be understood by this model, but first we should get a clearer picture of what a cortical unit is, and how it "predicts".

7) The neurophysiological meaning of "prediction" leads to a new understanding of Hebbian synaptic plasticity.

I have proposed that the neocortex as a whole makes predictions about its sensory inputs, and that this is accomplished by many cortical units simultaneously making predictions about their own inputs. It is now time to precisely define what “prediction” means in neurophysiological terms.

First we should resolve just what is the "cortical unit". As I already mentioned, Mountcastle proposed that it is a narrow column transecting the cortical surface. The cortical column receives inputs from other cortical and subcortical regions, it processes this input and then distributes the results, again to other cortical and subcortical regions. I have difficulty accepting the cortical column as the functional unit of the neocortex because the output of a column consists of many axons originating from different cells in different layers within the column. If the activity of these output axons differ, then there are really several outputs to the column. Therefore a column is not a "unit" but a related collection of units. One could debate whether the outputs of a column are independent or not. For the purposes of my argument I will consider that the functional unit of the neocortex comprises a single projection neuron combined with all of the other neurons in its immediate environment which affect its behavior. I refer to this unit often enough that I have given it a name. I call it a "NAC", for Neocortex Algorithmic Component* . The output of a NAC may send collaterals into its own neighborhood and it may project to many different regions of the brain, but unlike a cortical column a NAC performs a single function and has a single output axon. A cortical column is comprised of many NACs with similar but not identical inputs (see figure 1).

Because the cellular and synaptic organization of the neocortex is very complex, it is difficult to predict from anatomical observation what kind of processing a NAC might perform. So I will only make some very general assumptions about its behavior. First I assume that the output of a NAC, and indeed all neurons, is best understood as a frequency encoded real variable. Second, I assume that a NAC is essentially a continuous function over the entire range of its inputs. (The inputs to a NAC are all the axons which affect the NAC and which originate outside its immediate vicinity.) There may be tens, hundreds or thousands of input fibers to a NAC. If a NAC has n inputs, its function $f(i_1..i_n)$, has a unique value for every location in the n-dimensional space defined by its inputs. It is easiest to visualize the output of a NAC by seeing how it changes when only two of its inputs are varied. In this case we can plot the output as a smooth surface above a cartesian plane (see figure 2).*[1]

For understanding what "prediction" means we don't need to make any assumptions about the shape of a NAC function. Nevertheless we can

---

*I derived the term NAC from CMAC, which is a term that James Albus (1981) used for the functional unit of the cerebellum. A CMAC (Cerebellar Model Arithmetic Computer) is the algorithm implemented by a purkinje cell and associated local neurons.

*[1]We shouldn't get too comfortable with such a diagram because it can be very misleading. If any of the n-2 dimensions not plotted are changed, then the shape of the plot might be completely different.

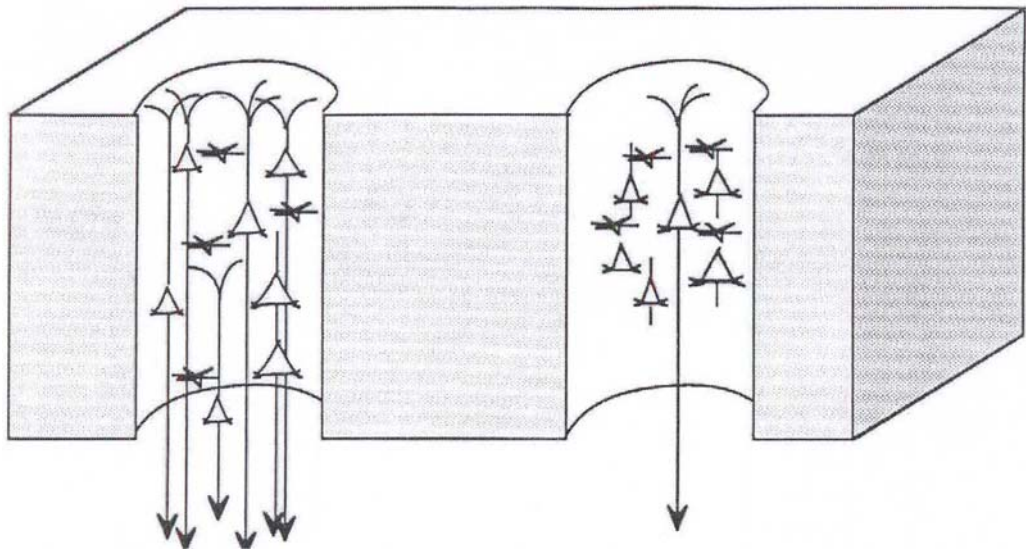

Figure 1. The cortical column on the left consists of several hundred neurons which receive inputs (not shown) and therefore have related physiological properties. The output of a cortical column is of many axons arising from many cells at different layers within the column. Since there are many outputs, the cortical column may best be considered as a collection of related units.

On the right side is shown a proposal for a new additional cortical unit which has only one output. It includes a pyramidal cell (whose axon is the sole output of the unit) and every other cell in its immediate vicinity which affects the output cell through local interactions. This new unit is called a NAC for Neocortex Component. A NAC is therefore similar to a column but it has the advantage of a output which allows us to characterize its behavior more readily.

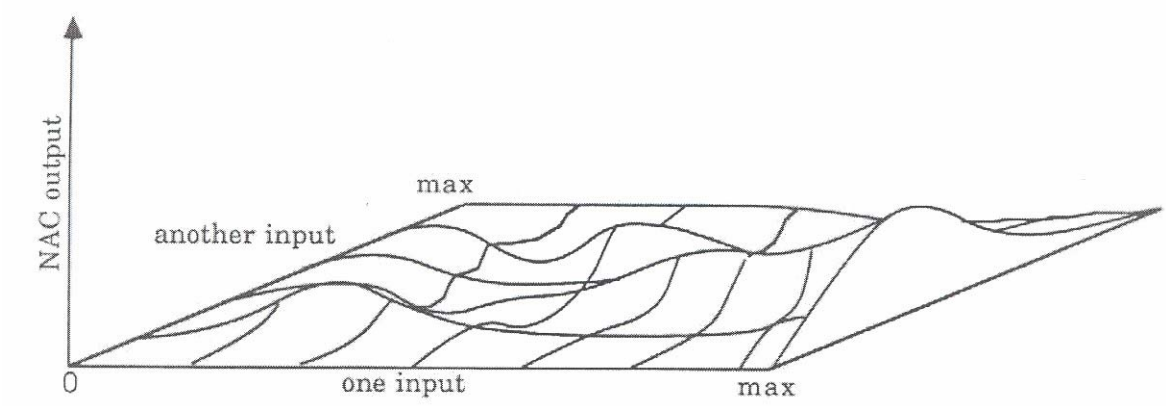

Figure 2. The output of a NAC is a continuous function over the space defined by its many inputs. By varying only two of its inputs, the NAC function can be visualized as a surface above a Cartesian plane.

describe the shape of a NAC function with some degree of certainty and this will be useful later on. First we can divide the inputs to a NAC into three categories; 1) specific excitatory inputs arising from distant areas of the cortex and subcortical structures, 2) inhibitory inputs arising from neighboring columns (Intercolumn collaterals are probably responsible for pericolumnar inhibition (Mountcastle, 1978)), and 3) non-specific subcortical inputs. An increase in the activity of many of the inputs converging from remote regions will lead to an increase in the output of the NAC. An increase in activity of only a few inputs from remote regions will probably have little effect. And an increase in the activity of many inputs converging on a NAC in an adjacent column will increase the output of the adjacent NAC and decrease the output of the NAC of primary concern. Thus the shape of the NAC function is probably similar to that exhibited by neurons in many other brain regions; it has converging excitatory inputs, it has an inhibitory surround, and is somehow modulated by non-specific reticular inputs. (Throughout the rest of the paper when I refer to inputs to a NAC I will generally mean the specific excitatory inputs arising mostly from other cortical areas. If I want to refer to the local inhibitory inputs or the non-specific inputs I will explicitly say so.)

Now let's get back to the question "What does prediction mean in neurological terms?" Consider another example of how we predict environmental associations. Occasionally I will listen to an album of music that I haven't heard for several years. Before playing the album I might not be able to recall any of the melodies. However, as I listen to the music and reach the end of one song, the beginning of the next song often pops into my head. Not only can I remember the lyric or melody of the next song, but I actually remember the exact pitch on which it begins. In a very real sense I "hear" the beginning of the next song before it starts. Obviously no sound is reaching my ears, but the neural activity within my brain is somehow similar to what will occur when the sound of the song does reach my ears. In terms of the NACs, some of them are firing as they would if the song had already begun. Since no sensory input is actually reaching the brain, this activity must be caused by the internal state, or context, of the brain. Thus "prediction" means using internal state to produce neural activity which is similar to what is expected to happen.

How can we understand this in terms of the shape of a NAC function? Let's say a NAC has 100 excitatory inputs, 10 of which are closely related to direct afferent input from my ears. The other 90 come from other areas of the neocortex which we don't know how to characterize. The first note of the second song on my album just happens to create a lot of activity on the 10 auditory input fibers causing a peak of activity in the NAC. (We might say that this NAC has a receptive field of a particular pitch or maybe a certain musical interval.) Visualize the output of this NAC as a hill or peak in a landscape with 100 dimensions. Figure 3 shows what this hill might look like projected onto two dimensions.

The other 90 inputs say something about the internal state of the neocortex,

and they most likely convey information about other auditory-related areas. (It all depends- on the specific wiring of the neocortex.) If these 90 inputs are always in the same state of activity when the second song begins, then the NAC can predict the forthcoming activity of the 10 direct afferent inputs. If you were a NAC you might say, "Every time these 90 internal inputs are like this, the other 10 afferent inputs become active shortly afterwards. Therefore I'll predict that the 10 affferent fibers will become active very soon and increase my rate of firing now." In terms of the shape of the NACs function, the peak becomes more level or rounded (figure 4).

Another way of saying this is that if x number of inputs are often active at the same time causing the NAC to have some output value y, then the NAC adapts by tending to produce an output close to y even if only some of the x inputs have reached the value associated with output y. In this way the NAC predicts that all x inputs will eventually become active.

I should point out that it is not necessary for the shape of the NAC function to be peaked. Prediction can occur for any shape function as long as the function becomes more "level" around the output value at which association of inputs occurs. From here on out however, I'll assume that the NAC output reaches a maximum at the event of interest. In this example a peak of activity occurs at the beginning of the second song.

It is highly unlikely that all the inputs to a NAC will be correlated with the same event. In our example perhaps only 30 of the 90 internal inputs would have a significant correlation with the peak occurring at the beginning of the song. The other 60 would be inactive or have random values at the time. During some other event, a different set of inputs might be correlated. (Thus there could be many different peaks distributed throughout the function space of the NAC.) Mathematically we can look at the partial derivitives of the NAC function with respect to the different inputs. If a particular input fiber i_k has a significant correlation with a particular peak output value of the NAC, then the partial derivitive of the NAC function with respect to i_k is reduced near the peak of activity.

Thinking again of the shape of the NAC function as a hill, we could say there are 100 different paths that lead to the top of the hill, one along each input. The top of the hill becomes more level only along those paths whose corresponding inputs have a high correlation with the event marked by the peak.

There is no one NAC which foretells the beginning of the second song. Many NACs,. perhaps millions, are simultaneously reinforcing each other with individual predictions. Collectively they predict all the attributes we remember about the beginning of the song. An individual NAC far removed from sensory or motor areas may be involved in millions of different collective predictions, so its "receptive field" may be very difficult to describe.

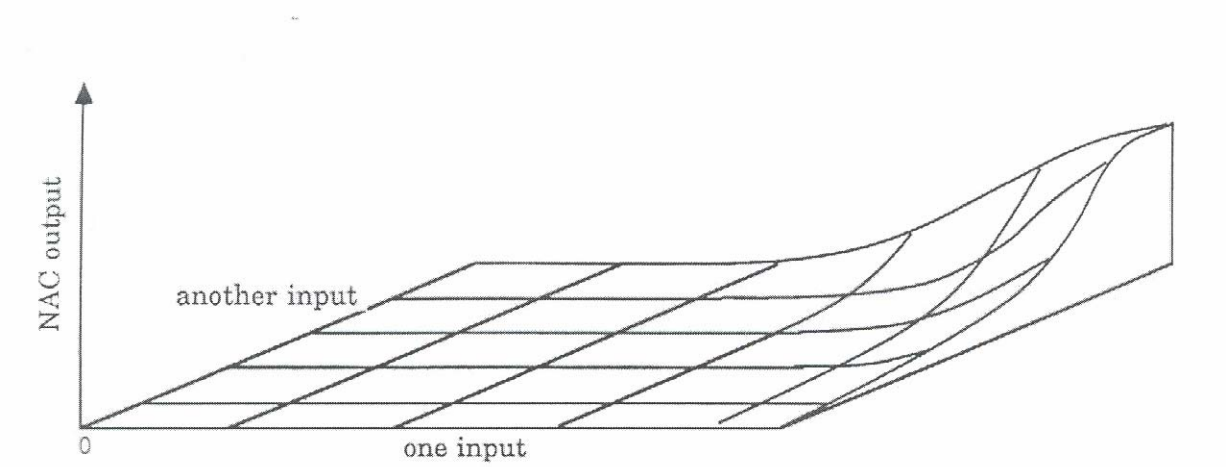

Figure 3. this show a peak of activity in a NAC associated with the concurrent activity of many excitatory inputs (only two of which can be shown).

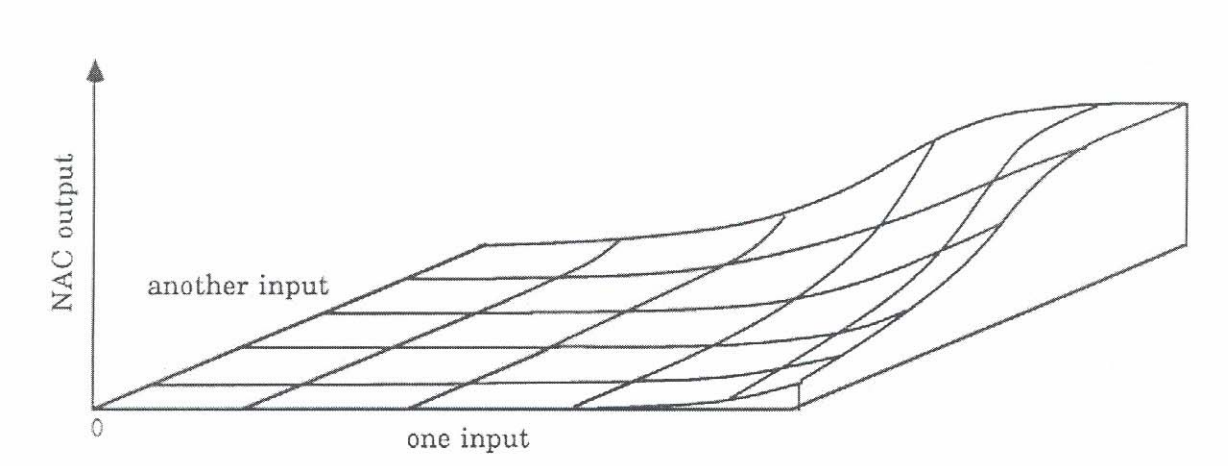

Figure 4. This is the exact same graph figure 3 after repeated exposure to concurrent activity on several of its inputs including the two shown. The NAC reaches its peak value (or close to it) over a broader area, even when all the inputs which were originally required to reach its peak have not become fully active. In way, the NAC has learned to "predict" the peak activity from the partial activity of its inputs.

Figure 4 greatly exaggerates the amount of change that actually occurs in the shape of a NAC function. Most likely you would not be able to see a change in the shape of the function as it adapts to associations of its inputs. Very small changes may be all that is required to achieve significant adaptation of behavior because the columns of the neocortex are in a continual state of competition due to their surrounding inhibition. In addition, the non-specific reticular inputs to the neocortex may keep a particular area of neocortex on the verge of transition. Thus small changes in the "competitiveness" of NACs may bring about large changes in the population of NACs which are active at anyone time. This is all very speculative, but one should be open to the possibility that very small changes are sufficient.

As evidence of this, consider how quickly we can form lasting memories. Think of something you did yesterday which you had never done before. Notice how many details you can remember. The first time I did this exercise I was completely surprised by how much I could remember of a hike I had taken two days earlier. I remembered details to which I was only briefly exposed and had not thought of since. I don't believe, as is sometimes proposed, that memories (i.e. the ability to recall associations) are stored temporarily in one location and later filed away. It is most likely that memories are stored only during exposure or recall. During the few seconds which are sufficient for remembering something, not much morphological change can occur.

Speaking of morphological change, what kind of mechanism could produce the changes that I have proposed for the adaptation of a NAC? The synapse modification rule attributed to Hebb (1949) is a likely candidate. In Hebb's words,

"When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."

This kind of synapse modification is no longer speculation. It has been observed and studied in several nervous systems. (Although I don't believe that it has been demonstrated in the neocortex.). For this discussion we need not be concerned with how the change comes about, only its effect.

Initially I did not see the relationship between adaptation of a NAC function and Hebb's synapse rule. This was because I did not want to make any assumptions about the shape of a NACs function. I only assumed that the function became more "level" in the vicinity of a trained association. Hebb's rule, on the other hand, implies an increase or peak in the output of a NAC at the point of association. In fact Hebb's rule implies an ever-increasing peak amplitude. The more a pre-synaptic and a post-synaptic neuron fire together, the bigger the effect the pre-synaptic neuron has. Of course there are limits to how often a neuron can fire and there are limits to how influential any particular synapse can become. Therefore, the output of an NAC would not increase indefinitely with repetitive training. Instead it

would reach some maximum. value and thereafter spread into a small plateau. Thus Hebb's rule fits nearly perfectly with the adaptation of a NACs function that I proposed earlier.

In retrospect you might say I have "derived" the necessity of Hebb's synapse from an observation about the perception of changes in familiar environmental patterns. After reaching this conclusion, I read Hebb's book, Organization of Behavior. There are other parallels in his theory and what I have developed so far. For example, his concept of a "cell assembly" roughly matches my concept of many NACs collectively making corrections. It doesn't matter whether you view what I have done as an elaboration of Hebb's theories or as being substantially different. Either way I believe there is good reason for optimism in this approach.

The intent of this section was to define the neurophysiological meaning of prediction. We can now see that "prediction" is not a conscious act, nor is an all or nothing affair. It is a graded neurological expectation that some association of neural activity will occur. Hebb's synapse modification rule provides a simple way for each cortical unit to automatically learn to make predictions about its environment,

8) Through cortico-fugal projections, the neocortex make its environment more consistent.

A central theme to my thesis has been that the neocortex operates on a single algorithm. Up to now I have concentrated almost exclusively on perception, i.e. the processing of afferents and internal activity towards the goal of successfully predicting patterns embedded within the afferent inputs. What about the outputs of the neocortex? Is it possible that we can understand its external projections by the same algorithm?

I think we can. This may seem surprising since the neocortex projects to many other nuclei within the brain and spinal cord. How could one algorithm produce the outputs for each of these different destinations? There is a simple and beautiful answer to quandary. The neocortex automatically learns to produce those outputs which make its inputs more consistent and predictable. Therefore the best way understand the output of the neocortex is as an. integral part of perception.

First, I'd like to dispel the idea that the neocortex has many different types of outputs. The neocortex does project to many different subcortical areas, but from its perspective the neocortex doesn't know what its outputs control, or for that matter what its inputs represent. The only thing that the neocortex really knows is the spatial relationship between certain afferents and efferents and how they are connected internally. For example, one common relationship is that cells which receive input from a particular area, say the medial geniculate body, are very close to the cells which project back to that same area. The neocortex doesn’t need to know that this input is related to hearing. All it needs to know is that the outputs from

this region are likely to affect certain inputs in a predictable way.

Next I propose that the output of the neocortex interacts with its environment to make its environment more consistent. I'll illustrate this statement with some examples.

Virtually all sensory afferents ascending to the neocortex project first to a nucleus in the thalamus. Spatial relationships and receptive field properties appear to be maintained on both sides of these nuclei, hence they are often called relay nuclei. It is also well known that each of these nuclei receive massive highly specific innervation from the area of the neocortex to which they project. Usually our description of the purpose of these corticothalamic fibers is something vague, such as "they modify the incoming signal" or "they create feedback loops". There is a simple and precise explanation which fits the model I am proposing. The neocortex is biasing the thalamic nucleus to pass the signals that the neocortex is expecting to receive. A NAC in the primary sensory area, like all NACs, learns to predict the concurrent activity on a number of its inputs. If some of its inputs are from other cortical areas and some are nearly direct afferents, then the NAC will adapt to predict the arrival of these afferents from the state of the cortical inputs. The output of the NAC may be considered as a prediction that a specific afferent input will occur. (So far this is identical to the song example I discussed previously.) Let's say this NAC projects to the thalamus exactly where the predicted afferent is relayed. Let's also say that this NAC output fiber reduces the threshold of the principle thalamo-cortical neuron onto which it projects.* This makes the thalamus more receptive to this specific input. It is now more likely that the input the neocortex is expecting to receive will actually occur. In other words, it is more likely that the pattern that has occurred in the past will occur again. The end result is that the neocortex has made its environment more consistent by biasing the thalamus to relay the information that is expected.

It is hard to judge how large an effect the cortico-thalamic projection will have. It might be very small, casting the deciding vote between two competing afferent inputs. (Like most areas of the brain, it is believed that lateral inhibition in thalamic nuclei causes competition between similar inputs.) Or it might have a very large effect, sending a signal back to the neocortex even without real sensory input. Either way the principle is the same. By manipulating its environment, the neocortex has made it more likely that its inputs will contain familiar and expected patterns. It does so automatically and with the same simple algorithm I postulated as the basis of perception.

The same argument can be applied to the other nuclei to which the

*Anatomical studies have shown this is indeed the most likely effect of cortical projections to the thalamus. (Shepherd, 1979)

neocortex projects, such as the inferior colliculus or dorsal column nuclei. Even efferent projections to the auditory hair cells can be understood in this way. In each case the neocortex encourages the subcortical area to pass whatever inputs the neocortex is expecting to receive. There may be a lot of truth to the adage, "you see what you want to see and you hear what you want to hear".

Next let's turn our attention to the role the neocortex plays in muscle contraction. This topic poses a problem because there is a large amount of complicated neural machinery outside of the neocortex involved in movement. There isn't a straightforward connection between the output of the neocortex and what eventually happens at the periphery of the body. The principles are the same however, so as long as we don't look for too much detail we will stay out of trouble.

We normally think of our environment as a relatively stable collection of objects and their relationships, such as the buildings, tools and people we interact with everyday. This, however, is not the environment that we sense. What we sense is quite different. The information reaching our sense organs is fragmentary and constantly changing. Let's say I blindfolded you and put a fountain pen in your hand. Without moving your fingers you wouldn't be able to say what it was. You could guess at its weight and maybe sense that it is long and straight. But without moving your fingers the amount of information you can gain from touch is very small. At no one instant can you feel "fountain pen". The way you recognize the pen is by moving your fingers over the object creating a serial stream of constantly changing simple sensations.

These constantly changing inputs are not "inflicted" upon the neocortex. The neocortex doesn't sift through fragmentary details trying to figure out what the object is despite the changing perspective. The reality is quite the opposite. The stream of inputs is almost completely under the control of the neocortex and the inputs that the neocortex receives are mostly what it expects to receive. In other words, you move your fingers over the pen in ways such that the subsequent sensations can be anticipated.

The more the inputs match what is expected, the more the neocortex understands what the object is. When the inputs don't match what is expected then the neocortex is confused and tries a different hypothesis.

The algorithm which the neocortex uses to accomplish this is the same one I have described over and over again. The neocortex has adapted to predict a recurring association of events. In this case, some internal state plus some afferent input plus some movement has consistently resulted in some subsequent afferent input. Only those movements which produce consistent results are learned.

It doesn’t really matter if the neocortex is completely responsible for the movement of your fingers, partially responsible, or had nothing to do with

it. The premise of the theory holds for varying amounts of subcortical involvement in the production of movement and can therefore accommodate the inclusion of other structures such as the cerebellum, basal ganglia and spinal cord. All that is necessary is that the patterns embedded in the afferents reaching the neocortex are simple enough that the neocortex can successfully predict their arrival and thus "understand" its environment. All the outputs of the neocortex, whatever their effect on movement, work towards this goal. If other parts of the nervous system independently achieve the same results, all the better.*

The point I want to make is that the everyday notion of our environment being a largely passive landscape of objects is an illusion. Instead, it is a fast changing stream of simple stimuli largely under our own control. Movement, of any kind, drastically changes the stimuli which reaches our senses. We automatically learn to perform those movements which produce consistent patterns in this stimuli.

Nowhere has this concept been ignored more often than in the study of vision. Many people still view vision as a completely parallel process. A two dimensional image is projected onto the retina. Some parallel processing occurs in the retina and the resultant image eventually gets projected onto the neocortex. It is assumed that further parallel processes are at work in the neocortex which eventually allow us to recognize objects or do whatever. This is of course all true, but also incomplete. The fovea and the saccade are two mechanisms which serialize the process of vision. They are analogous to the extra sensitivity of our fingertips and the movement of our fingers when we touch an object to see what it is. As with palpation (sensing an object with your hands), the saccade completely changes the patterns of innervation reaching the neocortex. Many people consider this as an inconvenience which must be endured to get a different view, that somehow "vision occurs between saccades". This is wrong. The saccade is an integral part of vision. It is the brain's method for expressing a prediction about its environment. The saccade performs the same function as moving your fingers over the pen.

If I saw your face right now, it is likely that my eyes would glance first at one eye, then the other, then the mouth, nose and perhaps back to an eye. Each movement of my eyes completely changes the patterns of innervation reaching my brain. Recognizing that I am looking at a face means successfully making predictions about what input is to occur next. If my neocortex could describe what it was doing it might say, "This looks like an eye on the left side of a face. Every time I have seen a left eye in a situation like this, there has been another eye to the right. Let's look there and I will expect to see a right eye. If I do see a right eye then I am probably looking at

*For example, contraction of the iris is a reflexive behavior which makes the inputs to the neocortex more consistent. It tries to keep the amount of light reaching the retina at a constant level.

a face,"

The brain learns to make those eye movements which produce consistent results, and not to make eye movements which produce varying results. If when you saw a left eye you made a saccade further to the left, the next thing that would come into view might be a hat, some hair or perhaps a mountain in the distance. A saccade to the right however, almost always results in seeing a right eye. The saccade to the right will therefore be chosen as the preferred behavior. The NAC algorithm automatically makes this choice. If for some reason I didn't find a right eye where expected, many NACs would all of a sudden be making wrong predictions and your attention would be brought to the matter. (see figure 5)

So far I have talked about cortico-thalamic projections, palpation, and role of the saccade. I could make similar arguments about how we turn our head to better hear a sound, or place an object under our nose and inhale to get a better smell, or even what we do to taste something (think of the procedures an expert uses to taste wine). No matter how you look at it, behavior is an integral part of perception.

I conclude that although the outputs of the neocortex may project to different parts of the nervous system, they all serve the same purpose of making the inputs to the neocortex follow expected patterns. The same algorithm that I have postulated to be responsible for perception works equally well for determining appropriate behavior. This leads to the realization that the outward behavior of an organism can be understood as an integral part of perception, and not as an independent entity.

9) Perception is the collective behavior of many cortical units simultaneously making correct or incorrect predictions.

10) The reticular system. directs attention guided by incorrect predictions of cortical units.

Up to now I have been rather vague about how to understand the ensemble action of the many NACs and cortical columns in the neocortex. I have stressed that the neocortex is constantly trying to predict what is going to happen next in its environment. To make these predictions it uses both the internal state of the neocortex and the current sensory input. The outputs of the neocortex automatically adapt to make the inputs more predictable. I showed that a single local algorithm can be responsible for all these attributes.

In this section I will elaborate on these ideas. Two concepts I will specifically address are the importance of columnar inhibition and reticular control of activity. Unfortunately many questions will have to remain unanswered since what I describe here is very speculative.

Most evidence suggests that excitatory activity spreads vertically within a cortical column. Therefore, we can consider that all the NACs within a

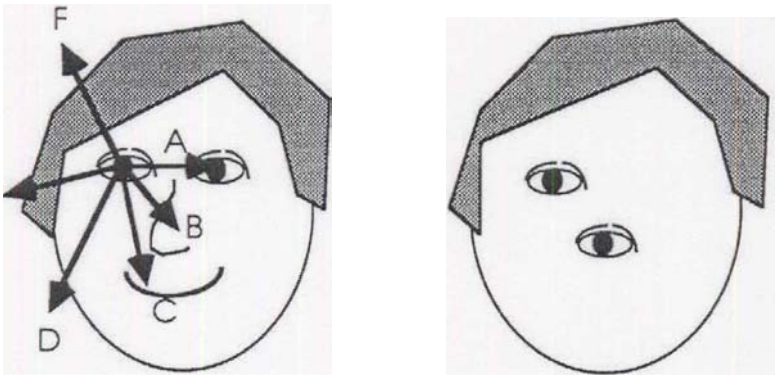

Figure 5. We learn to make those movements which lead to predictable patterns of stimulation. If your eyes are focused on the left eye of a face (shown above left), they may next glance in any direction. Of the possible movements shown, A, B and C lead to predictable subsequent input; D, E and F lead to less predictable inputs, Therefore, movements A, Band C will more likely be performed.

By making a saccade such as B, the brain predicts it will next see a nose. If instead it sees something else (shown above right), the prediction will conflict with the sensory input and your attention win be brought. to the matter.

The brain understands its environment based on the degree to which it can successfully make predictions about sensory inputs, Movements, including the saccade, are a means the brain uses to make its sensory inputs more predictable.

column are either active or inactive at the same time. In addition, it is believed that inhibitory forces exist between adjacent columns. Columns are thus competing with their neighbors for the right to be active.

Now I imagine that at any point in time some subset of all columns are active. Normally this subset is relatively stable. That is, the active columns are reinforcing each other in a stable arrangement of activity. The population of active columns steadily changes over time as our thoughts move from topic to topic. Hebb called such a self-exciting collection of neurons a "cell-assembly". I prefer to look at it more as a set of NACs which are making mutually compatible predictions.*

There are probably several things which make the active population of columns change over time. One is the reticular formation. Studies with associative memories have shown that one can control the playback of a stored sequence of events stored in an associative memory by raising and lowering firing thresholds (Wilwacher reviewed in Palm, 1982). Something similar is probably occurring in the neocortex. Think back to my discussion of noticing a change to your front door. Your attention is drawn to the area of misprediction. I imagine that the reticular system monitors activity throughout the neocortex. If a substantial number of NACs in a particular area, say low-level visual, start making incorrect predictions, the reticular system senses the extra activity and lowers the threshold in that area. That area is now able to change more readily and seek a new stable state which understands (i.e. correctly predicts) the stimuli which caused the confusion. In this way our attention can shift to any area where things are not occurring as expected.

This procedure is not limited to sensory areas. Imagine that a friend tells you something which just doesn't seem correct, perhaps it conflicts with what they had said earlier. Again you notice something is different than expected, except this time the misprediction is occurring far from a sensory area.. To understand this conflict the reticular system inhibits activity in sensory areas and increases activity in the internal area where the conflict occurred. The effect is to "tune out" sensory input and concentrate on the problem of why your friend told you conflicting things.

I suspect pericolumnar inhibition also plays an important role in directing the flow of activity within the neocortex. It can be considered as a means for amplifying small differences between similar patterns. If two adjacent columns are receiving nearly identical excitation, inhibition will force one to become active and the other to become inactive. It seems that the

*There is a general rule of intracortical projections that says if area A projects to area B, it is likely that area B projects back to area A. This is similar to the relationship between the neocortex and the thalamus and can be understood in a similar way. An active NAC biases its environment to produce the inputs which it expects, whether the environment is a sensory input or an input from some other area of the neocortex.

neocortex wants to reach definite conclusions and avoid indeterminate states of activity.

Inhibition also allows small changes in the competitiveness of columns to make major changes in activity. This could perhaps explain how brief exposure to something can make a lasting memory.

These discussions are all very speculative. The only thing I can say for certain is that a successful theory of neocortical function will have to take into account inhibitory effects and reticular contro1. It is not sufficient to merely connect together many simple elements.

11) Intracortical connections define limits to patterns which can be recognized and produced.

By now I have covered the main ideas of my thesis. I believe these ideas have merits on their own, but they fall short of a complete theory. If today you were to ask me to build a neocortex or mathematically describe the operation of one, I wouldn't be able to do it. There is nothing precise enough in what I have proposed so far to allow us to, let's say, build a machine which would learn to understand speech.

I feel capable of creating this more precise understanding. I can't say a priori just how it will turn out, but I have some ideas on what it will look like and how to get started.

As with everything else I have described, the mathematical understanding of neocortical function will deal with generic patterns. It will be phrased in terms that are independent of vision, touch and hearing. It will not specify what specific patterns can be recognized or produced by a neocortex, but will instead define the limits to these patterns. Remember that a neocortex placed in one environment learns to recognize different patterns than the same neocortex placed in a different environment. (The brain of a Tibetan sherpa is pretty similar to mine, but many of the patterns we have learned to recognize and respond to are different.) There is a commonality, however, and that is the limits of the patterns we can adapt to. By "limits", I mean the maximum spatial and temporal complexity of the associations that can be successfully tracked.

For example, consider the complexity of language. Reading comprehension can be looked at as the ability to track and *predict* the patterns inherent in language.* If a typical ten year old child were to try to read this paper he or she would understand very little. The sentences are too long and the concepts too subtle for a child to understand. Similarly there are authors whose writing I have trouble reading for the very same

*Because you understand what I am writing you can predict what word is likely to finish this ...

reasons. A child's reading ability will improve with practice as will mine, but there is a limit to how much any individual can improve. This limit is defined by the number of cortical units in the neocortex and by their specific connections. In the same way, a monkey's brain limits the complexity of language it can understand. No amount of training can improve either of our abilities beyond their limits.

A mathematical model of neocortical function will most likely involve an expression relating two quantities. On one side will be all those factors which constitute the anatomy of a neocortex and on the other side will be the limits to the patterns that can be recognized and produced. Devising such an model may seem like an impossible task, but I think it is possible.

On the bright side we can probably begin with small examples. I would start by seeing how a very simple neocortical-like structure could track and predict simple patterns. Knowledge gained by studying small systems can probably be extended to larger structures.

One approach that might work is to consider the neocortex as a state machine. NACs and their interconnections can be interpreted as a programmable state machine. The limits of the states that can be associated with one another can be rigidly defined. I have studied this concept and it shows promise. I would like to describe this in a separate paper.

I don't want to speculate further on how this mathematical model will come about or what it will look like. The important thing to recognize is that we must develop such a mathematical model if we ever want to promote brain theory from the speculation stage to the laboratory bench. My personal opinion is that we can develop such an understanding fairly easily and soon.

Discussion and Conclusion

Let me return to the quote from Francis Crick that I gave in the first section of this paper. Summarizing his impressions as a newcomer to neuroscience he observed:

"What is conspicuously lacking is a broad framework of ideas in which to interpret all these approaches."

Crick was right on target. The study of human behavior is drifting without a foundation, without a "framework of ideas". It is amazing how little we understand. We are not lacking in details of the anatomy, physiology or chemistry of the brain. So much data has been published that it is well beyond anyone's capability to absorb it all. There is no one research project which we can point to and say "if only we could do this, then everything would become clear". Even if every anatomical and chemical detail of a human brain were known we would be far from understanding how humans produce the behavior they do. We do not need more data, instead we need a new way of thinking about the data that we already have,

This paper is a proposal for just such a new way of thinking about adaptive behavior. I tried to get to the heart of what it means for an organism to adapt its behavior to its environment. Along the way I made many speculative proposals. If in a year from now I still believe in 50% of the details that I believe today I would be pleasantly surprised. Therefore what I want you to consider most carefully is the underlying philosophy. Following is a review of what I consider to be the most important philosophy.

Review of the major proposals

First and foremost is the recognition that adaptation of behavior is dependent upon patterns in the environment of the organism. By pattern I mean any consistently occurring association of sensory input. You must ignore the everyday notions of what behavior is and consider what the brain is really adapting to. Once you get past the sense organs, everything is the same. Everything (including language, culture and the ideas in this paper) converted to action potentials on afferent axons. It really doesn't matter what caused those action potentials. What matters is the patterns and consistencies in their findings. That is all the brain has to work with.

The second major point takes a small leap of faith. It is that the majority of adaptation in a human is accomplished by one basic algorithm.* This is an optimistic assumption because it makes our job of understanding human behavior a lot easier, but it is supported by a great amount of evidence. It frees us from having to postulate different algorithms for all the different types of adaptation we exhibit (just as it freed nature from the

* I only make this assumption for effected by the neocortex.

burden of evolving different algorithms). It tells us to look for commonality between vision, hearing and touch, and to search for an algorithm that works with any kind of environmental pattern.

The third point is that the basic algorithm of adaptation is the prediction of stimuli. If some association of stimuli repeatedly occur, we learn to anticipate that association. (Adaptation is dependent upon being able to anticipate what is going to happen next.) This algorithm is operating everywhere in the neocortex. The neocortex is composed of millions of units each learning to anticipate patterns in its "environment". When many units of the neocortex are successfully predicting a succession of events we "understand" whatever is causing those events. When many units of the neocortex make incorrect predictions we are startled or confused.

The fourth and final major idea is that the behavior of an organism is an integral part of perception. Every movement we make changes the pattern of stimulation reaching our senses. These changes are not an inconvenience to be ignored but an essential part of perception. We learn to make those motions which lead to predictable or consistent patterns of stimulation. Our behavior changes the environment of the brain (i.e. the afferent signals) from one that is relatively independent to one that is mostly under the control of the brain itself. This concept unifies behavior and perception into one process which can explain such diverse things from how we move our eyes over a strange object to why McDonalds is a popular restaurant chain.*

In addition to these fundamental concepts I proposed a new fundamental unit for the neocortex. I named this unit a NAC and defined what prediction means in terms of the algorithm it implements. This same algorithm worked equally well for prediction of environmental stimuli and for the outputs of the neocortex. Looking for a neurophysiological basis of the NAC adaptation led to a new meaning for Hebbian synaptic modification.

Optimism

I admit that the concepts I am proposing are insufficient to allow us to build or model a neocortex. This has been the fault with every theory of behavior ever devised. However, I am optimistic that these shortcomings can be readily overcome.

A major reason for my optimism is that I expect the final resolution of this problem will be far simpler to understand than most people expect.

---

*It is not a coincidence that every McDonalds restaurant serves the exact same menu in the same environment. The corporation knows that consistency is an important part of their formula for success. People find security in knowing what to expect. They pick the behavior which leads to predictable results.

Nothing in nature is really difficult to understand. What is difficult is figuring things out for the first time. This is a truism of science which we often forget. The nervous system is complicated and immense. We are intimidated by it and imagine that an understanding of adaptive behavior must be equally complex. But complexity is a symptom of not understanding something. Once you gain an understanding it never seems as overwhelming.

Consider the Copernican model of the solar system. The geometry of the solar system is easy to understand. Grade school children learn the general relationships of the planets, moon, sun and stars. A more exact understanding of the planetary orbits can be had with a few years of undergraduate study. But try to imagine what it was like before we knew the solar system was heliocentric. Understanding the motions of the objects in the sky was an extremely difficult problem. These motions are complex and many equally complex theories were proposed to explain them. Lengthy treatises were published extolling all sorts of opinions as to what was going on up there. The eventual resolution of the problem was not trivial, but it was far simpler than what most people were thinking at the time. Everyone had been looking at the problem from the wrong perspective. Nature had fooled them with the obvious and incorrect notion that most celestial objects revolved around the earth.

Today we are in the "pre-Copernican" era of the understanding of adaptive behavior. Nature has fooled us again into taking obvious but incorrect perspectives on what it is that makes humans so special. As in Copernicus' time, we have many unsuccessful theories and volumes of uninterrupted data.

In this paper I have tried to demonstrate that the behaviors which characterize our species can be reduced to a single simple (but not trivial) algorithm. To fully understand this algorithm only requires that we keep an open mind, be persistent, and look for simple truths. If I am right then there is indeed cause for optimism.

References

Albus, J.S. (1981); *Brains, Behavior & Robotics,* Byte. Peterborough, N.H.

Arbib, M.A (1969); Automata theory as an abstract boundary condition for the study of information processing in the nervous system. In: *Information Processing in the Nervous System.* Springer-Verlag

Crick, F.H.C. (September, 1979); Thinking about the brain. *Scientific American,* 241:219

Dreyfus, H.L. (1979); *What Computers Can't Do: The limits of artificial intelligence,* Harper & Row, N.Y.

Eccles, J.C. (1984); *The Human Mystery,* Routledge & Kegan Paul, London

Hebb, D.O. (1949); *Organization of Behavior,* John Wiley & Sons

Hopfield, J.J. (1982); Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. had. Sci.* USA, 79:2554:2558

Hopfield, J.J. & Tank, D.W. (l98S); Neural computation of decisions in optimization problems. *Biol. Cybernetics,* 52:141-152

Marr, D. (1982); *Vision,* W.H. Freeman and Co.

McCulloch, W.S. & Pitts W. (1943); A Logical Calculus of the Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics,* 5:115-137

Minsky, M. & Papert. S. (1972 2nd ed.); *Perceptrons,* MIT Press

Mountcastle, V.B. (1978); An organizing principle for cerebral function: the unit module and the distributed system. In: *The Mindful Brain,* MIT Press

Mountcastle, V.B. (1984); Introduction. In: *Dynamic Aspects of Neocortical Function,* Neurosciences Research Foundation, John Wiley & Sons

Palm, G. (1982); *Neural Assemblies: An alternative approach to artificial intelligence.* Springer-Verlag

Penfield, W. (1975); *The Mystery of the Mind,* Princeton University Press, Princeton, N.J.

Shepherd. G.M. (1979); *The Synaptic Organization of the Brain,* Oxford University Press

Tomg, H.C. (1972); *Switching circuits, theory & logic design.* Addison-Wesley

Things that go …. in the night

In the first half of this century, elevated trains were common in New York City. The tracks of the elevated trains ran up and down the avenues on steel platforms above the street traffic. For some apartments a few floors above street level, the trains passed very close to living room and bedroom windows. The overhead trains were noisy, dirty and blocked out the sunlight, so eventually most of the elevated tracks were torn down and replaced with underground tracks. Late one evening a local police station received several phone calls from people complaining that they were woken by loud noises in the neighborhood. A patrol car was sent out to investigate, but nothing unusual was found. An hour or so later the same thing happened again; there were phone calls complaining about noise, but nothing apparently wrong. This happened several times during the night with no explanation. The next day the precinct chief was pondering the unusual events of the preceding night. He noticed that the times of all the complaints corresponded to the times when the trains passed through the neighborhood. He quickly called the Metropolitan Transportation Authority (MTA) to find out what they were up to. The old trainman who answered the phone denied that the MTA was responsible for the noise. If fact, he calmly stated, "We didn't run any trains last night. The line was permanently closed just yesterday."