4

Implications of neural networks for how we think about brain function

David A. Robinson
Departments of Ophthalmology, Biomedical Engineering, and Neuroscience, Wilmer Institute, The Johns Hopkins University School of Medicine, Baltimore, MD 21287

Abstract: Engineers use neural networks to control systems too complex for conventional engineering solutions. To examine the behavior of individual hidden units would defeat the purpose of this approach because it would be largely uninterpretable. Yet neurophysiologists spend their careers doing just that! Hidden units contain bits and scraps of signals that yield only arcane hints about network function and no information about how its individual units process signals. Most literature on single-unit recordings attests to this grim fact. On the other hand, knowing a system's function and describing it with elegant mathematics tell one very little about what to expect of interneuronal behavior. Examples of simple networks based on neurophysiology are taken from the oculomotor literature to suggest how single-unit interpretability might decrease with increasing task complexity. It is argued that trying to explain how any real neural network works on a cell-by-cell, reductionist basis is futile and we may have to be content with trying to understand the brain at higher levels of organization.

Keywords: connectionism; coordinate transformations; hidden units; neural networks; oculomotor system; pursuit eye movements; saccadic eye movements; signal processing; vestibulo-ocular reflex

1. Introduction

The function of the brain is to process signals. Whether a group of cells is analyzing sensory signals, storing maps in memory, shaping motor commands, or doing other things, its essential action is to receive a signal, usually distributed over an ensemble of fibers, alter the spatial and temporal patterns of the signal, and pass this new signal, also distributed over its axon ensemble, to downstream networks. How this happens is the main goal of basic neuroscience. Although the recent discoveries of molecular biology reveal fascinating genetic events in assembling molecules and should lead to methods of treating or preventing diseases of the brain, they are, as far as one can see, of little use in illuminating the basic problem of how the brain processes signals or, if you will, how it thinks. This problem is compounded by the suspicion that various subdivisions of the brain may solve similar problems in different ways. This aspect of brain function, probably an *ad hoc* one, suggests that generalized algorithms for neural signal processing are unlikely to emerge. Nevertheless, the die-hard reductionists among us have taken it as a matter of faith that we will someday understand how the brain works, or at least bits and pieces of it, in the same sense that we now understand how an electronic circuit works.

This idea has long been thought naive, but the developments of the last decade in neural networks, which can do clever brain-like things with neuron-like unit behaviors even though they are just electronic circuits (or simulations thereof), offer a reason for taking another look at the problem. This target article makes no attempt to review the history or the scope of artificial neural networks. I am primarily interested in neural network models that simulate real neural systems where much of the neurophysiology is known. There are very few such systems, but the brainstem oculomotor system is one of them. The following is therefore an account of my impressions in trying to simulate parts of the oculomotor system with neural networks.

Unfortunately, the term "neural network" can be applied to many schemes involving artificial, neuron-like elements developed as far back as just after World War II. Within the last decade, however, there has been a surge of interest in networks that, in their most common form, consist of three layers and use a learning algorithm such as back-propagation. It is such networks that have stirred up so much recent interest in the possibility of shedding light on brain function. The following discussion concerns this kind of neural network – the kind that contains hidden units and learns by adjusting synaptic weights in response to some error-driven learning algorithm.

Such artificial neural networks contain neuron-like units connected by things resembling plastic synapses. They are capable of learning extraordinarily complex tasks, but after they have done so we cannot understand the basis of their success at the single-unit level. The same can be said of the brain or any of its parts. It contains neurons connected by synapses, most if not all of which are modifiable, and it also learns extraordinary tasks. These similarities are what give artificial networks their uncanny resemblance to real neural networks. Beyond this, however, there is a sharp departure in attitude. The engineer who uses a neural network to do a job, such as

running a chemical processing plant (a young but growing field of neural network applications), chooses the network because the operation is so complex that he despairs of ever creating a designed, hardwired controller. The virtue of the neural network is that it can learn to perform tasks beyond the scope of the design engineer – that is why he chooses it. He has no intention of examining the hidden units. Why? If he could have predicted their behavior, he could have designed the system himself. He knows that their behavior will be largely inexplicable; to examine and worry about them would defeat the whole purpose of using a neural network in engineering. Consequently, recording from single units would be regarded by the applied engineer as amusing but not very constructive.

The main message of neural networks for the neurophysiologist is that the study of single neurons or neuron ensembles is unlikely to reveal the task in which they are participating or the contribution they are making to it. Conversely, even if one knows the function of a neural system, recording from single units is not likely to disclose how that function is being fulfilled by the signal processing of the neurons. A corollary is that being able to describe that function mathematically tells little about what to expect when recording from single neurons.

Examples of these problems abound in the neurophysiological literature concerning single-unit recording in behaving animals. A typical report tells us that in area *X*, in the alert monkey, 27% of the cells were phasic, 18% were tonic, 38% were phasic-tonic, and 17% did nothing. The conclusion drawn from the study is that in area *X*, 27% of the cells are phasic, 18% tonic, 38% phasic-tonic, and 17% do not respond. This is not meant as a criticism; I have written similar papers myself. The point is that the problem raised with such painful clarity by neural networks are attested to by a great deal of the relevant literature in neurophysiology.

2. A brief introduction to neural networks

For those who are unfamiliar with neural networks and their capabilities, the following is a simplified description of their essential features. Figure 1 shows a typical scheme (e.g., Anastasio & Robinson 1989; 1990a; Sejnowski & Rosenberg 1987). Each unit is a simple model of a neuron. The output, a_j, of the jth cell, is to be interpreted as a real neuron's discharge rate in spikes/sec. The membrane depolarization of the jth cell is the sum of the activities, a_{ij}, of all cells, $i = 1, 2, \ldots, N$, projecting to it in proportion to their synaptic weights, w_{ij}. This sum is passed through a nonlinearity, NL, which recognizes that cells cut off at zero rate and eventually saturate at some high rate. The exact shape of NL is usually not important.

The cells are customarily arranged with a minimum of three layers, as shown in Figure 1b. Usually, all cells in one layer project to all cells in the next; projections in this type of network are always forward, never sideways or backward, because this would create reverberating feedback loops. Initially, the slate is wiped clean of experience by randomizing all the synaptic weights. A value is applied to each unit of the input layer, forming a spatial input pattern, and the resulting output is compared to a desired output pattern determined by an external

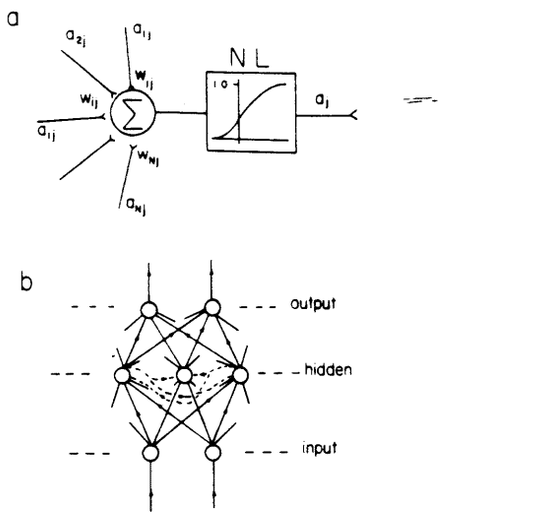

Figure 1. *a:* The behavior of each unit in a neural network. Its model membrane potential is the sum of all of its presynaptic inputs, a_{ij}, weighted by the synaptic strength, w_{ij}. This is passed through a bounding nonlinearity, NL, and the output, a_j, is thought of as a discharge rate to be forwarded to downstream neurons. *b:* Most networks are three layered. Usually, all units in one layer project to all units in the next for spatial transformation networks. When networks are asked to process temporal signals, they are allowed feedsideways (essentially feedback) pathways shown by dashed lines. The number of units in each layer is determined by each application.

"teacher." That is (and this is an important difficulty in proposing these networks as models of brain function), some element outside the network (the teacher) evaluates the output and "knows" whether it is "correct." The difference is the error, which is used to reward good weights and punish bad weights by calculating which weights and units were more or less active when the error was large or small. Usually, one finds some method of estimating the partial derivative of the error with respect to each weight, w_{ij}. The weight is then changed in proportion so that the error is reduced most rapidly in what is called the "method of steepest descent." This process is repeated until the error is driven below some tolerance level and the network has learned to produce the desired output. Usually, there are several input patterns, with each applied in turn until the network has simultaneously learned the correct response to each. It should be appreciated that a steepest descent method is entirely for the benefit of the investigator who would like the network, in computer simulation, to find a solution quickly. There is no indication that the nervous system uses such a method. In some lower animals, escape behavior, which may be required almost immediately after birth, is probably genetically determined; human children, on the other hand, have months to years available to learn many types of motor behavior.

In a similar vein, the issue is often raised whether a particular learning algorithm is physiological. The currently popular, back-propagation method (Rumelhart et al. 1986b) is fairly efficient in converging rapidly but is criticized as being unphysiological. There exists a variety of learning algorithms, however, that are more or less efficient and less or more physiological; all of these end up at the same goal: a network that has learned. In most

cases, the properties of the hidden units – our main concern – appear to be similar, regardless of the learning method. This observation cannot be proven rigorously and has not been tested in many specific situations, so it remains only a desired probability. If this is true in general, one could use a rapidly converging algorithm for efficiency, while noting that a more physiological method could have been used without changing the final outcome.

The end result is a network that has learned to give correct responses for a variety of spatial patterns presented as input. Our interest is in the behavior of the hidden units: If we record from 1 to 10 or 100 hidden units, can we tell what the network is trying to do and how each hidden unit is helping to do it?

A nice biological touch is that these networks do not find a unique solution. Each time one randomizes the weights and retrains, a new network is formed, one that still does the same job. Of course, no two cats (or people) are likely to be wired up identically, synapse for synapse, and neural networks reflect this feature, too. The reason is easy to see. Most networks have a much larger number of hidden units than output units. For example, with 40 hidden units and 2 output units, the convergence would be described by 2 equations with 40 unknowns. There is no unique solution. This situation is called the "overcomplete problem" because our university training has always emphasized questions with only one right answer. For neural networks, however, overcompleteness is not a "problem," it is a way of life. The networks are asked only to reduce the error to zero, not to find some hypothetical unique solution. They will use any solution that works. This sounds very biological.

An example of the power of neural networks is the reading machine called NETtalk (Sejnowski & Rosenberg 1987). Engineers have tried for decades to make a reading machine for the blind. They tried to hardwire into it all the complicated rules of English spelling and grammar as well as the many exceptions to those rules that together constitute, as we all know, a morass of contradictions. Engineers declared that a comfortably understandable reading machine could not be built. The NETtalk device receives as input a string of seven letters (and spaces) of English text and is trained to produce as output the phoneme corresponding to the letter in the center of the string. The rest of the string provides context. In Sejnowski and Rosenberg's work, NETtalk told when it did well and when it did badly and, like a child, it learned to read. After training on one text, it could extrapolate to others and eventually cope with an entire English dictionary.

What did the hidden units do? Some preferred consonants, some vowels, but there were just too many tasks and too many synapses (18,600) to attempt any reasonable explanation of how the network achieved its results on a synapse-by-synapse basis. The network learned the task, but observing its hidden units offered only the most arcane clues as to how.

This example illustrates that neural networks can solve exceedingly complicated problems – more complicated than those that can be solved by conventional design techniques. To put it another way: Tell the network *what* you want it to do, but do not even think about telling it *how* to do it.

Neural networks have a variety of applications. They can be used in Artificial Intelligence studies where there is no concern about whether or not they mimic brain function. Neural networks have also been seized upon vigorously by engineers as a new working tool for analyzing and controlling very complicated operations, such as chemical processing plants, where there are so many complex interactions between so many variables that a reductionist analysis is next to impossible. In addition, there have been many attempts to emulate brain function at abstract perceptual levels with no pretense that the network's units have any resemblance to neurons. There are also network models in which one or two of the layers have units that behave like real neurons but the other layers are simply conceptual and correspond to no known neurophysiology.

There are, nevertheless, a few network models in which microelectrode recordings have provided data on all the neurons in a real neural circuit, and it is these models that can begin to give us some credible indications about how the behavior of hidden units becomes more and more divorced from intuitive behavior as their tasks become more and more complicated.

2.1. Spatial and temporal networks. The networks described so far learn to identify spatial patterns presented at the input layer and respond with a spatial pattern of activity at the output units. These particular networks do not accept or generate time-varying signals and so are not very useful in modeling the neural control of movement. This limitation may be remedied by adding a membrane time constant to each neuron. This amounts to inserting the Laplace operator $1/(s\tau + 1)$ just before NL in Figure 1A, where τ is the membrane time constant (about 5 msec) and, if there are N neurons, solving N simultaneous differential equations. There is then no need for the units to project only forward; they may project sideways and backward to form multiple feedback loops that allow the network to generate interesting output waveforms. For a given input waveform, the output waveform is compared to the desired output and the root-mean-square difference over a performance interval is used as the error. Otherwise, learning occurs as before.

To distinguish these two types of neural networks, they will be called *spatial transformation networks* and *temporal transformation networks*. Obviously, some networks can do both types of transformations.

3. Examples of hidden messages in hidden layers

The difficulty in interpreting hidden units is illustrated at one extreme by artificial networks such as NETtalk, and at the other by almost any real network. Neither illustration offers much insight into how obfuscation progresses in hidden layers. Fortunately, the oculomotor system has a number of simplifying features, for example a single "joint," straight muscles, no stretch reflex, and linear behavior in premotor areas. These simplifications allow neural network models of the oculomotor system to have hidden units that, while offering an interesting amount of disarray, are also reasonably comprehensible. Moreover, numerous microelectrode studies have provided reasonably thorough descriptions of the behavior of the units in

all three (or more) layers of the real network, allowing us to determine whether the models reflect reality. These models only serve to point out ways in which hidden units will become harder and harder to understand as the complexity of the tasks increases.

3.1. The oculomotor neural integrator. The vestibulo-ocular reflex (VOR) starts in the semicircular canals, which provide a signal, coded in discharge rate modulation, proportional to instantaneous head velocity. As this signal leaves the vestibular nucleus, it constitutes an eye-velocity command to create an equal, but opposite, compensatory eye velocity so that the images of the visual world remain relatively stationary on the retina in spite of head movements. The eye muscles are mainly position actuators, however, and need to be given an innervation signal proportional to the desired eye position. It has been demonstrated that there is a neural network in the caudal pons (shared by the vestibular and the prepositus hypoglossi nuclei) that integrates (in the sense of Newtonian calculus) input signals with respect to time (see Robinson 1989, for a review), thereby changing the vestibular velocity signal into an oculomotor position signal.

This network is interesting in that it is relatively circumscribed anatomically and has a function that can be specified with precision. Lesion studies suggest that integration is a property of the network as a whole, rather than its individual neurons, and positive feedback seems an appropriate way to model it. In contrast, our relative ignorance about most spinal cord networks in awake animals makes it hard to guess the extent to which similar integrators are needed to change descending phasic commands into tonic innervation for muscles in the control of limb movements.

Since the real integrator network calibrates itself in the first few months of life (Weissman et al. 1989) and requires constant monitoring to maintain accurate function, it retains the ability to learn throughout life and has been modeled by a temporal transformation neural network (Arnold & Robinson 1991). The input signal, illustrated in Figure 2, arrives in a push-pull arrangement from a pair of semicircular canals reflecting a brief head movement. The model is freely connected; the inputs project to all interneurons, which in turn project to all other interneurons and to the output motoneurons. The output is a more or less compensatory eye movement that can be used to predict the error signal, that is, the rate at which visual images would move on the retina. This signal is transduced by direction-selective cells in the retina and distributed to the brainstem by the accessory optic system. The learning algorithm for the network in Figure 2 changed the synaptic weights one at a time and observed how sensitive the error was to this change. This partial derivative was then used to adjust each weight with the steepest descent method to drive the error toward zero. To do this, the network had to learn not only to integrate but to frequency-compensate the plant (eye muscles and passive tissues) by driving it with a combination of an eye-position and an eye-velocity signal. The model network did all this successfully.

One interesting emergent property of our simulation is that every cell in the network carries a combination of the position and velocity signals and no other signals, just as is seen experimentally with microelectrodes. It has been proposed that integration might be done step-by-step, each cell partially integrating the signal and passing it on for improvement. This does not happen in the model or in the real network. There are no partially integrated sig-

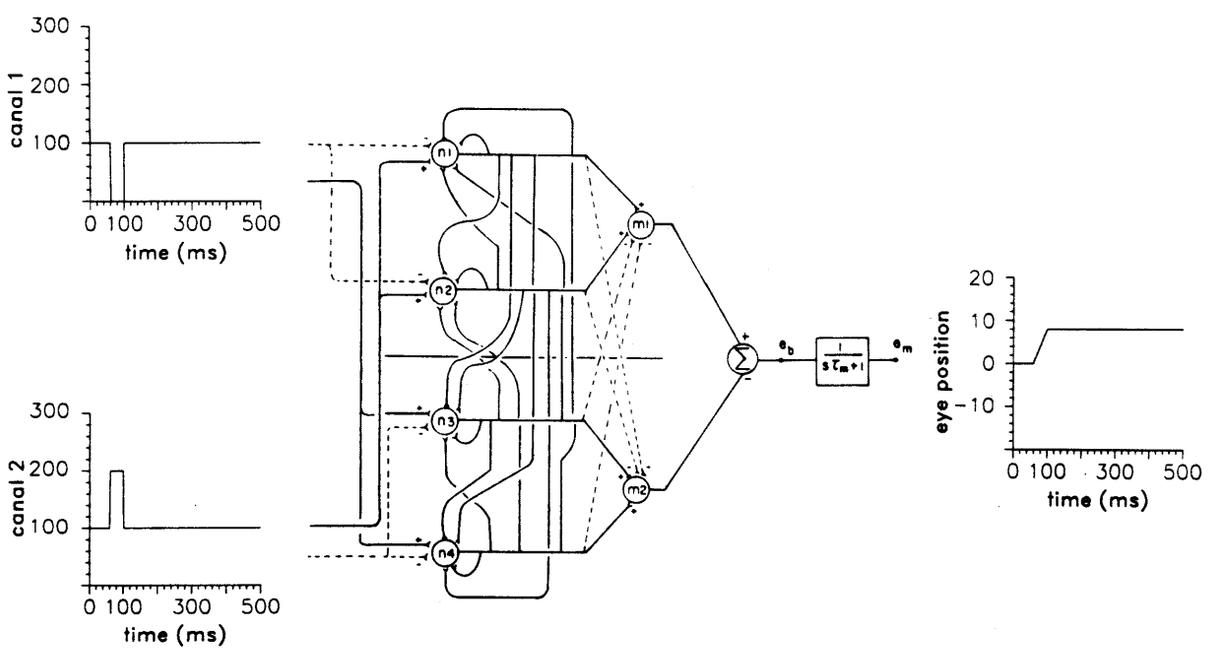

Figure 2. Push-pull, head-velocity signals come in on the left from a pair of semicircular canals. They are rectangular pulses of 100 spikes/sec on a background rate of 100 spikes/sec. These signals project to all interneurons, 4 in this example ($n1$–$n4$). All interneurons project to each other and to the two motoneurons, $m1$ and $m2$. All synapses are modifiable. The difference in their rates is passed through the oculomotor plant with a time constant, τ_m, of about 0.20 sec. The output eye position, right, is trained to become the time integral of the head-velocity input. In some versions of the general model, some projections from the canals and to the motoneurons were restricted to being excitatory (solid lines) or inhibitory (dashed lines). (Reprinted with permission from Arnold & Robinson 1991).

nals. If the eye-position signal appears, it appears fully integrated. This observation is particularly clear in the time domain (e.g., using step inputs), though it would be obscured in the frequency domain, such as when using sinusoidal inputs. For example, a phase lag of 45 deg could give the impression that a particular neuron was half integrating a signal by producing half of the needed 90 deg phase lag. A step input, rich in many frequencies, corrects this misapprehension. The integrated signal arises *de novo* from the circuit properties.

The interesting result, from our standpoint, is the variability in gain with which real and model cells carry these signals. Just as in experiments the sensitivity or gain of each model cell is measured for changes in eye position (k) and velocity (r) in (spikes/sec)/deg and (spikes/sec)/(deg/sec), respectively. Each cell has its own combination of k and r. All the model cells are different, as one would expect in a biological system. There is no correlation between k and r values; the network seems to treat using two signals as separate problems. To put it another way, the position and velocity signals are thoroughly distributed over the cells of the network. Distribution of signals is a natural property of neural networks and derives from the rich connections and the initial randomization of the synaptic weights followed by learning. Since this property imparts a biological flavor to simulations, it is hard not to believe that similar features also shape the real networks.

Occasionally, the model produces a cell that has its velocity and position signals going in opposite directions, or a cell that excites, or inhibits both of a pair of antagonist motoneurons. I call these "rogue" cells: They could be said to hinder rather than help, and one would discount them as model misbehavior if they were not actually observed from time to time with microelectrodes. That rogue cells exist tells us that the system works in spite of such cells; if 90% do right, they can easily overpower 10% doing wrong. The model suggests that such cells do not have a special function. The hidden units offer a menu of behavior patterns and the output layer picks and chooses whatever leads to a solution. When learning stops, because the error has become zero, some hidden units are left in a behavior pattern that seems contrary to the engineering mentality. The network, of course, only seeks a solution and is not concerned with good behavior. The point is that rogue cells are predicted by neural networks and exist in real ones.

3.2. A model with multiple motor commands. We wanted to use a neural network model to show how signals that create saccades and smooth pursuit movements, as well as vestibulo-ocular movements, could become distributed over the cells in the caudal pons (Anastasio & Robinson 1989). For this purpose, it is unnecessary to include a neural integrator in the model. It is sufficient to use a spatial transformation model in which the input signals represent eye-velocity commands in deg/sec, that is, discharge rates in a pair of push-pull encoding neurons; the output is eye velocity, representing the difference in discharge rate of two output, push-pull motoneurons. Thus, there were six input neurons: a push-pull pair for each of vestibular, pursuit, and saccadic velocity commands and two push-pull output neurons. All but saccadic input units had resting discharge rates of 0.5 (50% of maximum rate, NL, Fig. 1a) indicating a resting state with tonic cocontraction.

If the vestibular input pair changed from {0.5, 0.5} to {0.6, 0.4} indicating that the head was moving, say, to the left, the output neurons were asked to produce the opposite response of {0.4, 0.6} to produce a compensatory eye velocity to the right. If the pursuit input pair changed to {0.6, 0.4}, the output should respond with {0.6, 0.4} to produce an equal eye velocity in the same direction as the request. Saccades were exceptions, because their velocity commands are a burst of activity for ipsilateral saccades with no firing at any other time. Thus, the background rates of the input neurons were {0.0, 0.0}, saccadic inputs to one side were {1.0, 0.0} and to the other {0.0, 1.0}. The output neurons were required to change from a background rate {0.5, 0.5} to {1.0, 0.0} or to {0.0, 1.0}, respectively.

The network quickly learned all three of these simple tasks. We then looked at the 40 hidden units and found the 3 signal types to be distributed over them in a seemingly random way. As usual, the gains of the hidden units were calculated as the change in their activity divided by the change in one of the push-pull inputs. Just as is found experimentally (Tomlinson & Robinson 1984), each cell could then be described by:

$$R = (kE) + r_p\dot{E}_p + r_v\dot{E}_v + r_s\dot{E}_s \qquad (1)$$

where R is its firing rate, $\dot{E}_p$, $\dot{E}_v$, and $\dot{E}_s$ are eye velocities during pursuit, vestibulo-ocular, and saccadic movements, respectively, and the r's are the corresponding gains. The term (kE) represents the eye position term from the neural integrator (which is being ignored here for simplicity). Each cell (model and real) had a different sensitivity to each type of eye movement and all combinations of r_p, r_v, and r_s could be found. In summary, initial randomization and error-driven learning produced a group of simulated neurons with behaviors closely resembling the premotor oculomotor neurons in the caudal pons.

The network did not treat all the signal types completely independently. If a hidden unit helped move the eye in one direction for pursuit, it usually did the same for a vestibular signal, although with a different gain. There were also rogue cells, however; these might discharge, for example, when the eye went left in pursuit but also when it went right in a vestibulo-ocular movement. Again, these cells seemed to serve no special purpose; they were the result of the initial randomization, and the network found a solution in spite of them. There were similar rogue hidden units for saccades. Most hidden units behaved sensibly, bursting for saccades in one direction and pausing in the other. But, like real interneurons, some burst (or paused) for saccades in one direction and did nothing for saccades in the other, and a few even burst (or paused) for saccades in both directions. Again, the model suggests that it would be a mistake to assign any special function to these rogue cells.

This study has the expected result that a premotor neuron participates in many or all of the motor acts theoretically permitted by its anatomical connectivity. In the oculomotor system there are only four major conjugate systems (we have neglected the optokinetic system); each of these clearly has a different, identifiable function and can be independently activated. The simplicity of this

arrangement allows one to appreciate just how commands are mixed on premotor neurons. In the spinal cord, the situation might be much more complex, in part because our wide repertoire of limb movements might not be constructed from the sum of a small number of distinct, functional subsystems.

The equation describing the overall input-output behavior of our vestibulo-oculomotor network provides little or no insight into how these movements are organized. The seemingly randomly distributed signals of the hidden units are pieced back together again carefully at the output by synaptic learning so that the gains from input to output are all exactly 1.0. That is, the output should equal the input (a gain of 1.0) for pursuit and saccadic movements, but should equal minus the input (a gain of −1.0) for a vestibular input. The equation describing this behavior would not suggest that the three signals involved were distributed over the entire hidden layer and then reassembled on the motoneurons. This distribution and reassembling mechanism probably developed on an evolutionary scale, creating redundancy and robustness to lessen the impact of lesions, but the input/output equation makes no such predictions.

3.3. Hidden units in coordinate transformations. So far we have looked at the temporal construction and distribution of vestibulo-ocular hidden signals. Next, we wanted to look at signals with spatial orientations (Anastasio & Robinson 1990a). We chose the vertical vestibulo-ocular reflex (VOR), a two-dimensional coordinate transformation, for simplicity. It creates compensatory eye movements using the cyclotorsional eye muscles for vertical head movements in any combination of pitch and roll sensed by the four vertical semicircular canals. Thus, there were four input units, one for each vertical canal, and four output units, a motoneuron for each of the superior and inferior recti and superior and inferior oblique eye muscles of, say, the left eye. As in the previous section, a spatial transformation network is adequate to model these movements in which the activity levels of units in the input layer represent head velocities as reported by the canals, and the activity levels of motoneurons in the output layer represent eye velocity in the pulling directions of the muscles.

The system was trained by rotating the model about many axes in the horizontal plane to create many combinations of roll and pitch. The canal excitations for each rotational axis are related simply to the geometry of the canals, and for the output motoneuron activity, the rotational axis of the eye is related simply to the muscle geometry (Robinson 1982). Thus, the model concentrates only on the neural portion of the process – what happens between canal inputs and motoneuron outputs. The error signal is represented by a failure of the motoneurons to generate signals that would completely compensate for head movements. Once again, the error signal is a measure of the motion of visual images on the retina.

The network required only 500 iterations to learn to generate an accurate vertical VOR; eye velocity compensated for head velocity in all combinations of roll and pitch. We used 40 hidden units. To inspect the behavior of each hidden unit, we found its sensitivity axis, that is, the axis of rotation that creates the maximum modulation of activity. For the canals, this is the axis perpendicular to the canal plane. For a motoneuron, this is close to the axis around which its muscle rotates the eye. The sensitivity axes of the hidden units pointed over a wide distribution of directions (Fig. 3, left), many of which clustered loosely near or between the principal axes of the canals and muscles, but many were scattered in other directions. The experimental results of Fukushima et al. (1990) for real interneurons are shown in Figure 3, right. The scatter in the real neurons is not as pronounced as in the hidden units of our model, but clearly there is a distribution of sensitivity axes over a pool of interneurons.

Evolution has designed the semicircular canals to resolve head velocity vectors with maximum accuracy (Robinson 1982); the canals are oriented almost exactly at right angles to each other. It therefore seems odd that their signals would be intermixed among hidden units, which would decrease the signal-to-noise ratio. On the other hand, because the VOR is an adaptive system, it is constantly being optimized, and errors are driven to zero. This could compensate for the signal degradation. But then, why did evolution go to all the trouble to orthogonalize the canals so well? Whatever the reason, information from the canals appears to be distributed and intermixed over a set of interneurons and then reassembled vectorially at the motoneurons. Again, the behavior of the hidden units does not, by itself, suggest that a coordinate transformation is occurring, or how it might occur.

The mathematical description of this transformation is both interesting and amusing. The head rotation vector, $\dot{\mathbf{H}}$, produces a neural output from the canals that is also a vector $\mathbf{C}$ (a triplet of push-pull activities from three pairs of canals). The conversion of $\dot{\mathbf{H}}$ to $\mathbf{C}$ can be described by a 3×3 matrix $[C]$, which describes the canal geometry. (We can revert to all three canals in all three dimensions for this discussion.) Similarly, motoneuron activity of three pairs of push-pull motor nuclei constitutes a neural vector, $\mathbf{M}$, that produces an eye-rotation vector, $\dot{\mathbf{E}}$. This conversion can be represented by a 3×3 motor matrix $[M]$, which describes the geometry of the extraocular muscles. The brainstem, the neural component of the VOR, takes the canal signals, $\mathbf{C}$, and connects them to the motoneurons, $\mathbf{M}$, to effect an appropriate eye movement $\dot{\mathbf{E}}$. This process can be described by a 3×3 brain stem connectivity matrix $[B]$. Thus, the whole process can be described by

$$\dot{\mathbf{E}} = [M]\,[B]\,[C]\,\dot{\mathbf{H}}. \qquad (2)$$

When the VOR is working correctly, $\dot{\mathbf{E}}$ equals $-\dot{\mathbf{H}}$.

The brainstem matrix $[B]$ can be found from Equation 2 (Robinson 1982). Its coefficients represent quantitatively how much each canal pair should excite or inhibit each motoneuron pair. But this seemingly elegant description of brainstem function contains no hint that the vector components of $\mathbf{C}$ become disassembled and scattered over many interneurons with sensitivity axes apparently unrelated to those of the canals or motoneurons. This is an even clearer example of a mathematical description telling us *what* must be done without giving any idea of *how* it is done. This description is even misleading in its beguiling simplicity.

Such mathematical descriptions can be pursued to even further lengths. The vector $\mathbf{C}$ is a covariant vector, because each canal output represents the *projection* of $\dot{\mathbf{H}}$

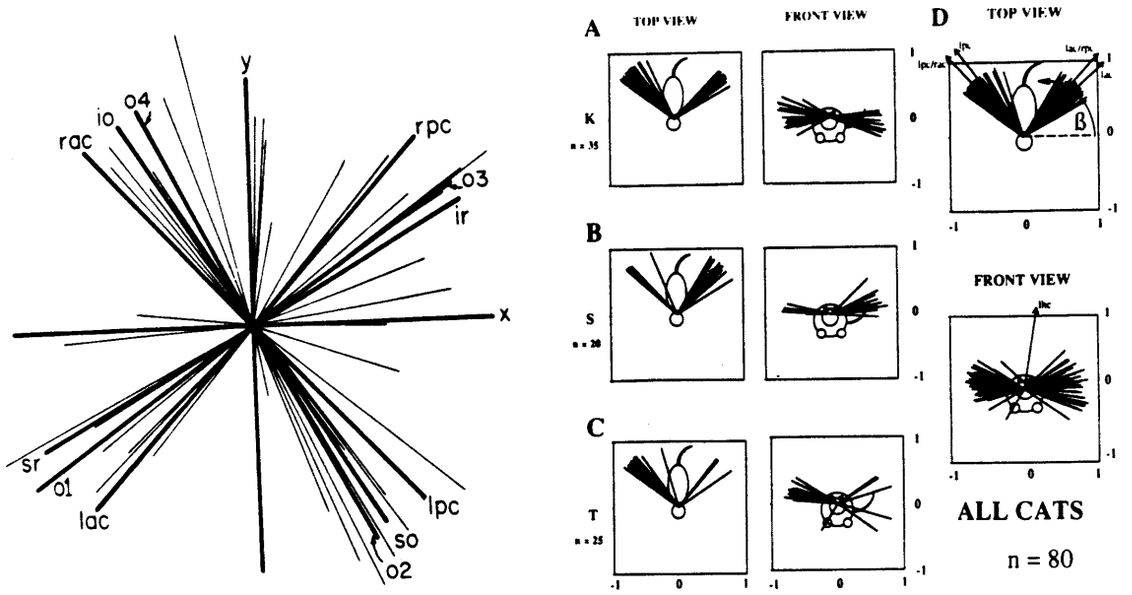

Figure 3. Simulated, left, and actual, right, sensitivity axes of interneurons of the VOR. *Left:* Looking down on the horizontal plane (as in top view at right), y, the roll axis, is anterior; x, the pitch axis, is medial. The axes of the right and left anterior and posterior canals are labeled rac, rpc, lac, and lpc. Those of the left superior and inferior recti and obliques are sr, ir, so, and io. The other vectors show the sensitivity axes, with magnitudes, of the 40 hidden units in the simulation of Anastasio and Robinson (1990). Their axes cluster somewhat around those of the canals and muscles, but there is appreciable scatter and variability. *Right:* Data from Fukushima et al. (1990) of the sensitivity axes of second-order vestibular neurons that project to ocular motoneurons in three cats, A, B, and C. Top view: looking down on the cat (the same view as on the left). Front view: the scatter in the tilt of these vectors out of the horizontal plane. D: the axes of all three cats pooled. The top view should be compared to the model behavior shown on the left. Because recordings were referred to the left side, these axes should be compared to the clusters shown on the left around the lac and lpc quadrants. Note that the cat is upside down compared to the figure on the left. We think the scatter of axes in the model are not dissimilar to those seen experimentally. (Figures reprinted with permission from Fukushima et al. 1990.)

onto the sensitivity axes of the canals. But $\dot{\mathbf{H}}$ is a contravariant vector in which the velocity components along the x-y-z Cartesian coordinates add by the parallelogram rule. To convert the contravariant vector $\dot{\mathbf{H}}$ to the covariant vector $\mathbf{C}$, one must multiply the former by the metric tensor of the canal space. Consequently, $[C]$ does two things: It transforms coordinates from Cartesian to canal coordinates and then multiples the coordinate transformation by the metric tensor to produce a covariant vector (see Robinson 1982, for simple, specific examples).

Now $\mathbf{M}$ is a contravariant vector because muscle forces add via the parallelogram rule, so $[B]$ not only performs a coordinate transformation between canal and muscle coordinates, but multiplies that by the inverse metric tensor. Pellionisz and Llinás (1980) were the first to introduce these concepts in motor control. They further proposed that, since covariant and contravariant vectors appear to be common to all sensorimotor systems, the metric tensor aspects might be handled in a common region of the brain, with the cerebellum being a candidate.

We can now see two extremes in thinking about how the VOR works. In one, basic incontrovertible laws of physics are brought forth, involved in the suggestion that their mathematical expressions are specifically recognized by the brain and dealt with in a compartmentalized manner. In the other, all these considerations are swept aside and a simple network of sensory, motor, and interneurons is asked to wire itself up to eliminate an error signal. The simplicity of the latter scheme, its goal-oriented single-mindedness, its disregard for mathematical elegance, and the naturalness of the sloppy wiring that results draw one to the neural network representation. The wiring is not really sloppy, it is only distributed, but the seeming randomness of hidden unit behavior gives a biological flavor of casualness. Nevertheless, the final solution of the network contains a coordinate transformation and an inverse metric tensor that are very well concealed in the hidden units. A central question is whether anything is gained by recognizing these mathematical transformations. Could these transformations be distinguished neurophysiologically and assigned to different anatomical locations? Given the broad way that signals are distributed over the hidden layer, it would be very difficult, if not impossible, to reconstruct the elements of the $[B]$ matrix by recording from interneurons. Similarly, trying to identify a metric tensor representation in the cerebellum with microelectrodes is likely to be truly impossible. Nevertheless, we can conclude from this example that mathematical descriptions of what a system is trying to do are of little help to the neurophysiologist trying to understand how real neurons do it. It should be added that the laws of physics and mathematical descriptions are perfectly appropriate at the peripheries, in this case the muscles and the canals. This also applies to the recognition of peripheral coordinate systems; canal primary afferents and motoneurons naturally reflect the anatomy of their end organs. The questions raised here, however, pertain to central neurons.

If we were to combine these three models – the neural integrator, multiple command types, and three dimensions – we would produce a Babel of oculomotor spatial and temporal signals in the caudal pons. All interneurons would carry all three types of eye-velocity signals (saccadic, pursuit, and vestibular) as well as the eye-position signal, each with its own sensitivity axis. Based on these models, it is doubtful that the sensitivity axis for pursuit in a given neuron would be exactly aligned with that for vestibular movements or saccades, although some sort of correlation might emerge. The sensitivity axis for the eye-position signal might even differ from that for any of the velocity signals in the same cell. In the extreme, a rouge cell might be activated quite differently for different movements: looking up, pursuit to the left, vestibulo-ocular to the right, and pause for saccades in all directions. One would expect such cells to be rare, but our simulations suggest that they could occur. This picture would be bewildering indeed if we were not lucky enough to know the functions of these three subsystems in all three dimensions and, guided by insights from neural networks, able to allow for the distributed nature of the hidden units.

3.4. Using efference copy to reconstruct the outside world. As a final example, we can turn to a recent study by Zipser and Andersen (1988). A long-standing problem in oculomotor physiology is whether we track targets using only retinal error (the difference between target image and fovea) or re-create the position of the target in space (with respect to the head, or body image, or inertial space) and use this re-creation to formulate motor commands. The latter idea seems realistic because limb pointing must be done in spatial coordinates. We can see a target, close our eyes, turn our bodies away from or toward it, and then point to it with reasonable accuracy. Thus, the position of the target in space is somehow calculated in the brain and can be used to direct our limbs (see also Fetz's, Bloedel's, and Stein's articles, this issue). We also know that if a visual target jumps from point *A* to *B* to *C*, before the eye can move, we can make saccades in total darkness, from *A* to *B* to *C*, if asked to do so (Mays & Sparks 1980). Note that the retinal error of the target at point *C* seen from the eye position at *B* was never available, although the eye went correctly from *B* to *C* without it. We can conclude that eye tracking is not directed by retinal error alone.

In another model for generating saccades (see Fig. 4), we used a local (internal) feedback pathway to compare current eye position, E, obtained from an efference copy, E', to desired eye position, E_D (defined with respect to the head, if the head is stationary, or with respect to space itself; Robinson 1975). This model reduced eye position error to zero quickly, thereby producing a goal-directed saccade. There is much evidence both for and against this local feedback model. Stimulation of many parts of the brain (e.g., the superior colliculus, cerebellum, and frontal eye fields) evokes retinotopic saccades that displace the eye in a fixed direction by a fixed amount from any initial position. In contrast, other regions have been discovered (e.g., the supplementary eye fields, Schlag & Schlag-Rey 1987) where stimulation can bring the eye to a fixed orbital position from any initial position (i.e., goal-directed eye movements).

There are two basic problems with his local feedback

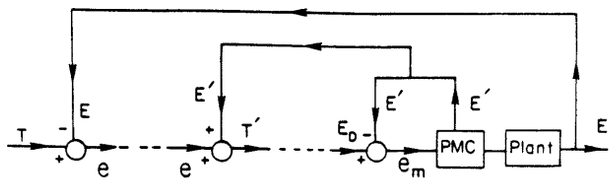

Figure 4. A model of saccade generation that reconstructs the location of a target in space. Retinal error, e, is the difference between physical target position, T, and eye position, E. It is proposed that the CNS adds an internal copy of eye position, E', to e to reconstruct an internal copy of target position, T'. If the CNS decides to make a saccade to T, it sets T' equal to desired eye position, E_D. The difference between E_D and E' is motor error, e_m that enters premotor circuits (PMC) that generate the saccade through the plant (orbital tissues and muscles) and provide the efference copy, E'.

model: It offers no role for the many long-lead burst neurons found in the reticular formation, and it postulates the signal E_D that no one has ever seen. This would require a cell that encoded a visual target position in space, independent of eye position. These problems have caused others to modify the model while essentially retaining the benefits of local feedback (e.g., Scudder 1988). The desired eye-position signal, E_D, in the model of Figure 4, would have to come from another hypothetical calculation: If e is the retinal error (the angle between the target image and the fovea) and E' is an efference copy of eye position, then their sum, T', is an internal re-creation of target position in space. If the brain selects T' as the goal of a saccade, it sets E_D equal to T' and initiates the saccade. As already mentioned, T' probably exists in some form in the brain because we can point to previously seen targets with eyes closed. There remain the questions of how T' is coded and whether it really is used to direct saccades.

Recording from neurons in the monkey parietal lobe, Andersen et al. (1987) found cells that responded to visual stimuli in the usual retinotopic fashion but were also influenced by eye position. These cells responded better to stimulation of a particular point on the retina if the eye was in certain specific positions rather than others. Thus, the cells carried the retinal error signal, e, in the location of its receptive field on the retina. The cell also carried the signal, E, in the form of a multiplicative modulation of the visual response that depended on eye position.

With both signals present, Zipser and Andersen (1988), reasoned that given the power of neural networks, a network could use these signals independently of their format to combine them to yield T', the target position in space. Of course, given the map-like nature of the main input, e, the output would probably be represented in the form of a map with its output signal coded spatially (by location on the map) rather than temporally. Thus, the input layer consisted of a grid representing the retina, with the visual input represented by the activity of units centered around a specific locus on the grid (Fig. 5). The eye position signals were a linear function of E for the four directions: left, right, up, and down. The output was a grid representing external space. Stimulation of the input grid by a target with a constant location in space for many different eye positions had to produce activity at the same point on the output grid. The learning was done by back-

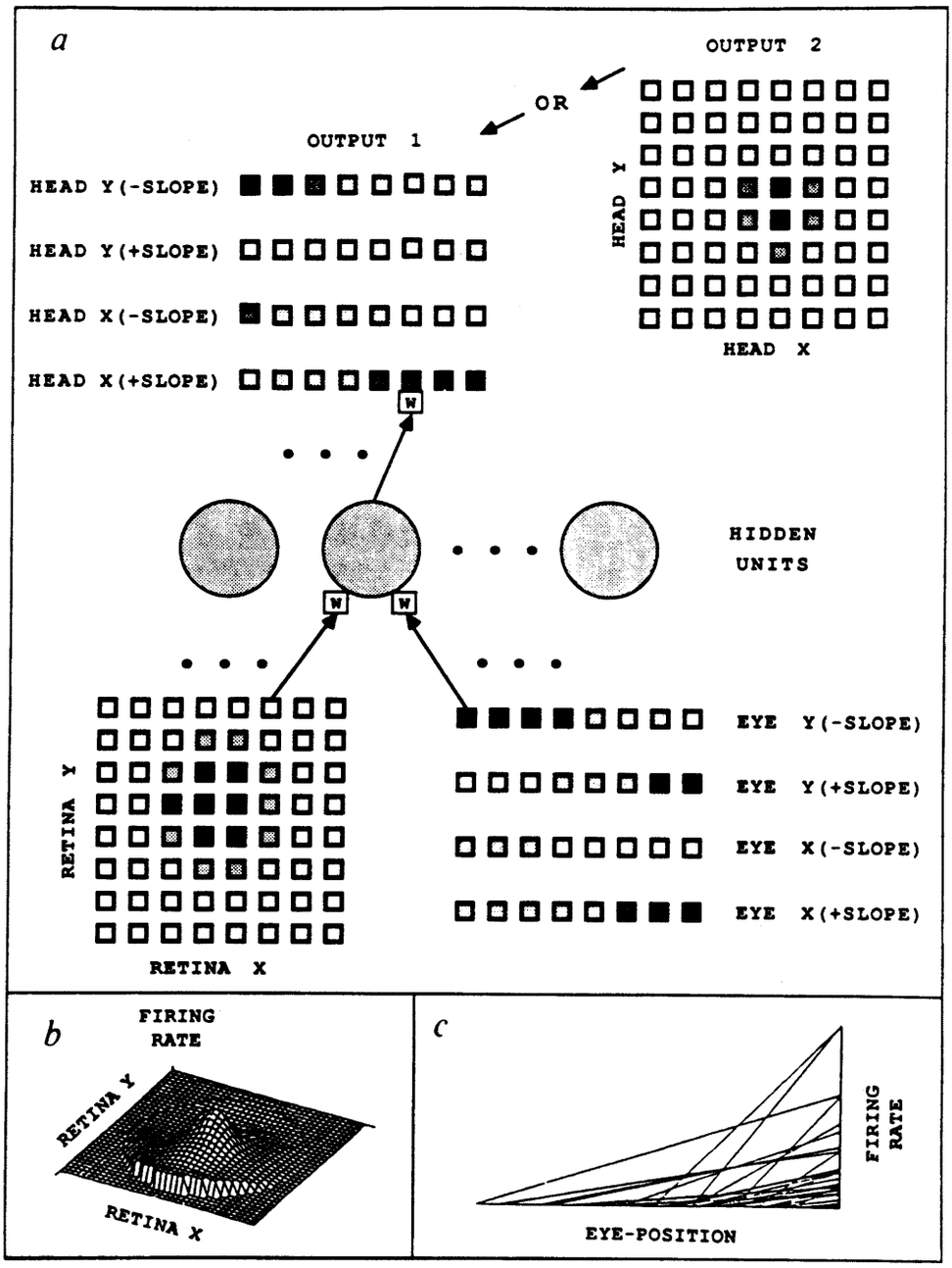

Figure 5. A neural-network scheme (proposed by Zipser & Andersen 1988) uses the behavior of cells in the parietal cortex to calculate the position of a target in space. *a:* At lower left is an 8 × 8 grid of retinal neurons constituting one input layer. The presence of a visual target is indicated by cell activity coded in the intensity of the shading. The second input layer is a set of four rows of eight cells (bottom right) coding eye position for up, down, left, and right. All units in each input set project to all hidden units, which project to all output units. Output units are displayed in two formats. The first, upper left, is an array, similar to the eye position array, lower right, but in head coordinates. A more intuitive display is an 8 × 8 grid representing external space (upper right) showing physical target location. When the network has learned its task, it will combine the target's image location on the retina with the eye position signal and obtain the location of the target in space. *b:* The receptive field properties of a parietal neuron where eye position is held constant and retinal locus is varied. *c:* The dependence of parietal neurons' firing rate on eye position as retinal locus is held constant. (Reprinted with permission from Zipser & Andersen 1988.)

propagation, but as mentioned earlier, the exact learning rule may not be important. The teacher was external – the network was told by the experimenter when it made errors and by how much.

The network learned to make accurate localizations and the hidden units, not surprisingly, resembled neurons in the parietal cortex (Fig. 5). Of course, this network is only a demonstration of how things might be. Just like the cells purported to carry the signal E_D, the cells in this model's output layer have never been observed. One could, of

course, press on with the model and design a network to take T' (or E_D) from a spatial code to a temporal code (a burst of spikes) for generating saccades, setting aside for the moment the problem of the apparent nonoccurrence of the signal T'. The immediate question is whether one could build a neural network capable of doing both – finding T' and generating a saccadic pulse – in which T' or E_D never explicitly appeared. This seems likely.

The main message is that hidden units contain signals that are not just hidden from an external observer, but are even hidden from an invasive observer with a microelectrode. One sees only scraps and pieces of information jumbled together, which can be interpreted only if one knows ahead of time what one is seeking. Again, the mathematical description of what one thinks is being done ($T' = e + E'$) gives no hints of how interneurons might actually do this. If the combined model proves feasible, it will create a problem for those who argue about whether saccades are organized in a retinotopic or craniotopic "coordinate system." Zipser and Andersen's model shows only that craniotopic signals are potentially available. The extended model could make saccades without explicitly realizing such signals. Then what does it mean to ask if the system is craniotopic? The model's hidden units would not be simply represented in retinotopic coordinates either, making the question of coordinate systems an ill-posed one (a frequent result of neural network analysis). The network gets the job done, and for it, coordinate systems are irrelevant; they are a problem of our own making. Obviously, the sensory and motor periphery have their own geometries, which we often call, incorrectly, coordinate systems. (Coordinate systems are human inventions used to measure spatial relationships.) However, we are not concerned with these geometries so much as with whether, in a central neuron, it is even useful to ask in what coordinate system it works. I suggest that it is not.

4. Conclusions

The examples considered here have all been relatively simple because, so far, only a few model networks exist with units that at all resemble real neurons. These "realistic" networks have had to be simple because few people try to model real neural networks with uninterpretable functions. The input signals have consisted of a limited number of types, with only one or two dimensions, and the outputs have been equally simple. The number of simultaneous functions have been few and, more important, they can even be described quantitatively. When one realizes that the network disassembles the input signals, scatters them over the hidden units, and then reassembles them, the behavior of the hidden units becomes understandable. It is not hard to imagine that increasing the number of types of signals, their dimensionality, or the number of functions will lead, long before one gets to memory, language, and consciousness, to a Babel of signal scraps on neurons.

The examples considered above are anecdotal and offer no systematic approach to how this complexity of unit behavior will grow with the complexity of the overall system. The main characteristic to emerge from these simple examples has primarily been the distributed nature of signals. That signals are distributed in systems of neurons is so well known that it is usually brushed aside as an inconvenience. One function of these simple networks is therefore to provide us with concrete examples of signal distribution that will make it more difficult to treat this essential feature of brain organization so lightly. We have also pointed out the phenomenon of rogue units and suggested that to search for coordinate systems in the brain may be a waste of time. One suspects that our attempts to impose our preconceived notions from mathematics, physics, and communication theory (tensors, coordinate systems, sine waves, quaternions, and so on) onto the behavior of central single units is not useful, and indeed, may be holding us back. Modeling increasingly complex systems will probably not reveal anything systematic. Each real network may solve its own problems with its own particular tricks. A related issue is the continuing use of the simplistic lumped, linear membrane summation scheme of Figure 1. Membrane nonlinearities are known to exist that could enrich network behavior, sometimes to our confusion, with such things as oscillations and multiple stable states.

So far, we have concentrated on modeling single-unit behavior, yet much brain modeling has taken place at the black-box level. Has this been useful? There is no simple answer. My complaints about input/output mathematical descriptions that put an entire system into one black box are a case in point, but only an innocuous one. They merely describe the behavior and package it in equations. This is a useful exercise, so long as one realizes that it may have limited predictive ability with respect to single-unit behavior. Modelers rarely stop there, however; they want to guess what goes on inside the black box. So they assemble inner black boxes marked, for example, lead, lag, delay, gain, sample and hold, Fourier transform, and so on, putting them in feedforward and feedback arrangements until the simulation behaves, usually over a limited set of inputs, the way the real system does.

What is gained by this approach? Too often – nothing. How many other arrangements of boxes do the same thing? This question cannot be answered. One can almost never say that there are no equivalent circuits or even no equivalent simpler ones if one likes Occam's razor (a device, it has been proposed, used by biologists to cut their own throats). Usually, the interbox variables have never been observed, and if the forecast of neural networks is correct, they never will be. An example from oculomotor physiology is our ready use of efference copy in our models, signals that reflect desired eye position or velocity. The concept of efference copy is not in doubt, it is just that very few cells in the brainstem, or anywhere else, carry such signals. Usually, the position and velocity signals are intermixed. Neural networks suggest that this might be okay; the signals will probably not be separated until the output layer (motoneurons), so their failure to exist in pure form may be irrelevant. Our jaunty hand waving may be justified after all, but with no thanks to most black-box modelers.

Black box models seldom make testable predictions. The most desirable prediction would concern how single neurons should behave. Even if a black box model were conceptually correct, it seems unlikely that single neuron behavior could ever be predicted, given the considerations of this target article. This poses a real challenge to

black-box modeling. If they cannot predict unitary behavior, their usefulness is severely limited.

Nevertheless, some boxes are useful in expressing a concept. There are a few simple examples in the oculomotor system. Direction-selective units in the retina use a small network of cells to transduce the direction and speed of images across the retina. Several plausible models of this network have been proposed, although none have been unambiguously verified. A convenient feature of this network is its anatomical isolation. Its only output must be on retinal ganglion cells. When we set forth a black box with the Laplace operator, s, in it (differentiation) and put image position into it and show neurally encoded image velocity coming out, we feel a useful representation has been made. This is a box not to be thrown out. The same is true of the neural integrator described here: It can be anatomically isolated in the sense that for horizontal head movements its input is a one-dimensional velocity signal from the canals, its output, a one-dimensional signal, one component of which is a position signal funnelled through the motoneurons. All findings so far indicate that the integration is done by a network located in specific regions of the caudal pons. Again, when we set out a box with $1/s$ in it (Laplacian for integration) we feel it is a useful representation.

To throw out these two boxes as part of a grand purge of black boxes would be to throw out the baby with the bath water. However, when one moves centrally and throws in a few more boxes to model the saccadic pulse generator, the optokinetic system, or velocity storage in the VOR (see Robinson 1981, for further descriptions) one's model moves further from reality and becomes less and less testable. Even though this type of model is fairly simple (say 2 or 3 more boxes and 1 positive or negative feedback loop) and based on a fairly well understood system (the oculomotor system, by now subjected to thousands of microelectrode penetrations by dozens of research teams over 15 years), *not a single one has been confirmed or refuted by single-unit recordings*. Over the years, black-box modeling has largely proven to be a blind alley.

Let us return to the applied engineer who uses neural networks but has no desire to look at their hidden units. The engineer must be amused to discover that there is a small army of people in neurophysiology doing just that. What are we learning and what are we not learning by recording from single units? One unrealistic feature of current artificial networks is that they allow all units in one layer to project to all units in the next, thus thoroughly intermixing all inputs. Localization of function in the brain has been one of our major allies, allowing us to concentrate on one modality, such as vision, at a time. Much gross localization had already been done by lesions before microelectrodes came on the scene, but the latter have helped to augment such findings, expand them to their borders, and fill in details in ways that would have been impossible or, at least, tedious to do with lesions. This much localization could have been done by recording field potentials or multiunit potentials, however, since isolating single units addresses a different question – how do neurons process signals – to which there are very few answers. Microelectrode recordings have also pushed localization into other, multimodal areas where lesions would have been uninterpretable. For example, there are 20-odd visual areas fed by primary and secondary visual cortices, some specializing in color vision, others in motion detection. The supplementary eye field in the supplementary motor cortex and the presence of eye-movement-related activity in the parietal lobe are other examples of multimodal localization.

That is the bright side. The dark side is what single-unit recordings have not told us. In the most general sense, they have not told us how neurons, or groups of neurons, process signals. At a very basic level, single-unit studies have shown that direction-selective neurons exist in visual systems and that they extract image motion. This is a very valuable piece of information, but such recordings apparently cannot tell us how the processing is done. Eye-velocity signals converge in the caudal pons, and suddenly one sees eye-position signals in neural activity there, but unit recording does not tell us how this was done. Perhaps these are neural circuit problems that can be solved in another 20 years.

A slightly more complicated situation is the stretch reflex where, to our embarrassment, we do not know after 50 years the purpose of an animal's proprioceptive signals (see also Gandevia & Burke's article, this issue). Ironically, this is because of a lack of unit recording: Most spinal cord physiologists refuse to bite the bullet and record from proprioceptive neurons in behaving animals. The result is a growing body of electro-anatomy, but no signals. This is a problem that could probably be solved by unit recording, but only in behaving animals.

As we move centrally, the situation deteriorates rapidly. The oculomotor system is no exception. Saccades are created by bursts of activity in premotor cells, but rostral to this, in the superior colliculi and frontal eye fields, attempts to relate unit behavior to saccade metrics (i.e., position, velocity, and direction) have opened a controversy, one that has been fueled rather than quelled by single-unit data. This controversy has led to a plethora of black-box models that have gone nowhere. As mentioned earlier, the great majority of single-unit studies reported in the literature describe signal scraps without interpreting their meaning. Unfortunately, the doubts of the engineer (who uses neural networks as a tool) about whether anything useful will come from examining single-unit behavior, have been painfully corroborated by decades of experience in sensory and motor neurophysiology.

Two future developments might help. One is that applications of neural networks are relatively new in engineering, and although they are used to solve "insoluble" problems in the field, it is unlikely that engineering theorists will be content to let it go at that. This area will no doubt become a new field of study and perhaps in 10 years relationships will emerge that can build a bridge between system function and hidden-unit behavior and tell us how to relate one to the other. The second development is a recent one in neural development research. How does synaptic modification occur? This is a rapidly evolving field that is at the moment noted more for its diversity than for providing any simple, universal answer. We all know that real neural networks learn, but we do not yet know how. In artificial networks, a learning rule is stipulated, the network learns, and one does not worry about details. Perhaps we will be forced to adopt the same attitude: Unable to understand real neural networks at a synaptic or cellular level, we may be forced to wave our

hands and say, "Well, we do not know how it works in detail, but we do know that such and such a learning mechanism is at work here to allow synaptic weight changes that are compatible with the observed learning, verified with models, and that's it." This might be as close as we are ever going to get to explaining how a network does its thing. As a result, if we know the learning rules, we may have to accept the inexplicable nature of mature networks.

ACKNOWLEDGMENTS

The author's laboratory is supported by Grant EY00598 from the National Eye Institute, the National Institutes of Health, Bethesda, MD. I thank A. McCracken for preparation of the manuscript and C. Bridges for the illustrations.