The Computational Self

TERRENCE SEJNOWSKI

Howard Hughes Medical Institute, Salk Institute for Biological Studies, La Jolla, California 92037, USA

Division of Biological Sciences, University of California at San Diego, La Jolla, California 92093, USA

ABSTRACT: **Your brain is never at rest. Shifting patterns of activity course through your brain at night as you review the events of the day and plan the next day before falling asleep. Rumination is a reflection of the Self that is not directly driven by sensory stimuli. When we record from single neurons in the brain, we discover that even in the absence of sensory stimulation, neurons are continuously active. This is called maintained, or spontaneous, activity, and although it is well documented, it has not been as well studied. Most experiments are designed to look for signals that are elicited by sensory stimuli above the background, without mentioning whether the background has changed too, as it often does. New methods have been developed recently that allow us to study the brain's spontaneous activity and to explore how it might provide clues to the origin and nature of the Self.**

KEYWORDS: **independent component analysis; electroencephalogram (EEG); event-related potentials (ERPs); spontaneous activity; computer models**

There are many different levels of investigation in neuroscience, and FIGURE 1 illustrates these levels ranging from molecules to the entire brain over 10 orders of magnitude of spatial scale. At meetings of the Society for Neuroscience, attended by more than 25,000 neuroscientists working at all of these levels, one can be overwhelmed by the sheer amount of knowledge we have uncovered about the brain, more in the last 10 years than in all previous history. Integrating between levels can help to unify this knowledge and allow us ultimately to understand how complex brain states that give rise to the Self arise from molecular, synaptic, cellular, network, and systems mechanisms.

Address for correspondence: Terrence J. Sejnowski, Salk Institute, 10010 N. Torrey Pines Road, La Jolla CA 92037. Voice: 858-587-0423; fax: 858-587-0417.
terry@salk.edu

Ann. N.Y. Acad. Sci. 1001: 262–271 (2003).
doi: 10.1196/annals.1279.015

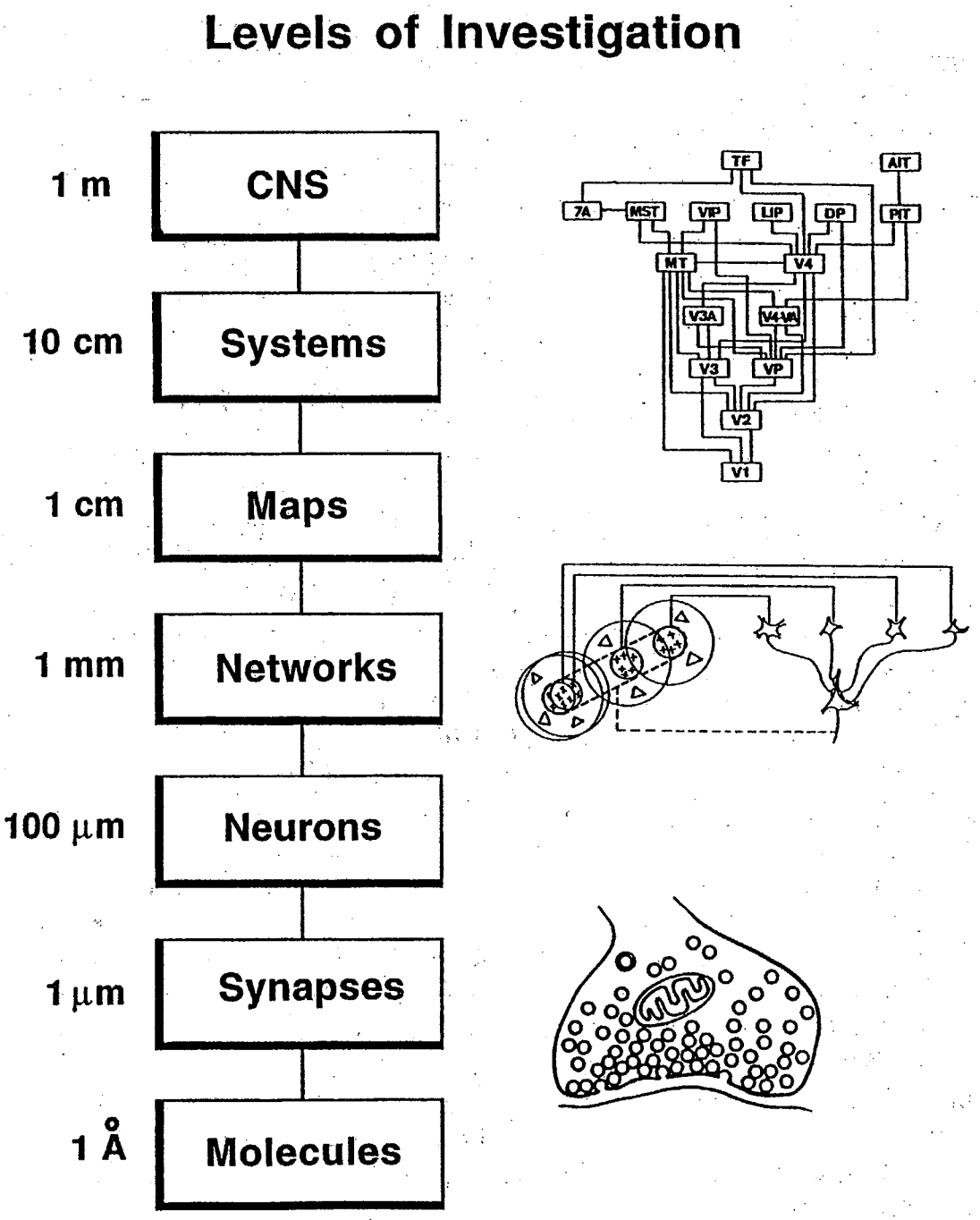

FIGURE 1. Levels of investigation of the brain organized according to spatial scale. Behavior is a property at the highest level involving the entire central nervous system. At the lowest level we can study the individual molecules of the brain such as neurotransmitters and receptors. There are many intermediate levels between these two that could contribute to the origin and nature of Self.

We are now faced with a "Humpty Dumpty" project: We've taken apart the brain and we know almost all of its pieces, but we are like the child who has taken apart his father's watch and is trying to put it back together again. How are we going to do that with as complex and dynamic a structure as the brain?

At the Computational Neurobiology Laboratory at the Salk Institute, we approach the problem of integration by developing new techniques for analyzing brain recordings, new computer models for simulating brain activity, and mathematical theory for understanding the activity. Computer models allow us to study how the many components of the brain interact together. Such models can give us insight into how these interactions give rise to percepts and thoughts. Once a model has been confirmed experimentally it can be mathematically analyzed to extract general principles.

FORESTS AND TREES: UNDERSTANDING NEURONS IN POPULATIONS

Over the last 50 years, there has been an especially strong focus on single neurons, spurred by the seminal development of the microelectrode, a thin piece of wire with a very sharp tip. If you put it into the brain and get lucky, you can record from a single neuron. The advantage of isolating a single neuron is that one can listen to what it is saying and find out in great detail its specific preferences. In the visual system, for example, we can determine which properties of the visual world each neuron responds to best. The trouble is that there are 100 billion neurons. Recording from all of them would not only take a very long time, but in the end, we would have only a huge catalog.

The real problem is that we know too much. We see all the trees, but we don't see the forest. In *The Computational Brain,* Patricia Churchland and I (1992) predicted that 100 years from now, when the history of our period in neuroscience is written, this period will be said to be based on the "theory of the microelectrode"; that we were so focused on the tip of that recording device that we were blinded to the obvious fact that neurons interact in complex patterns. In the brief summary I will present here, I will describe what brain activity looks like at a high level, looking at the forest. At the end we will return to the Self and describe a research program for how to find it in large populations of neurons.

FIGURE 2 summarizes a popular model that dominates our current view of how the brain computes. The figure shows how the brain of a monkey responds to a visual stimulus. First, neurons in the retina are activated and then 30 milliseconds later, following some processing in the retina, the signal arrives in the lateral geniculate nucleus, a visual relay nucleus in the thalamus. Shortly thereafter, it arrives in the primary visual cortex, where perception

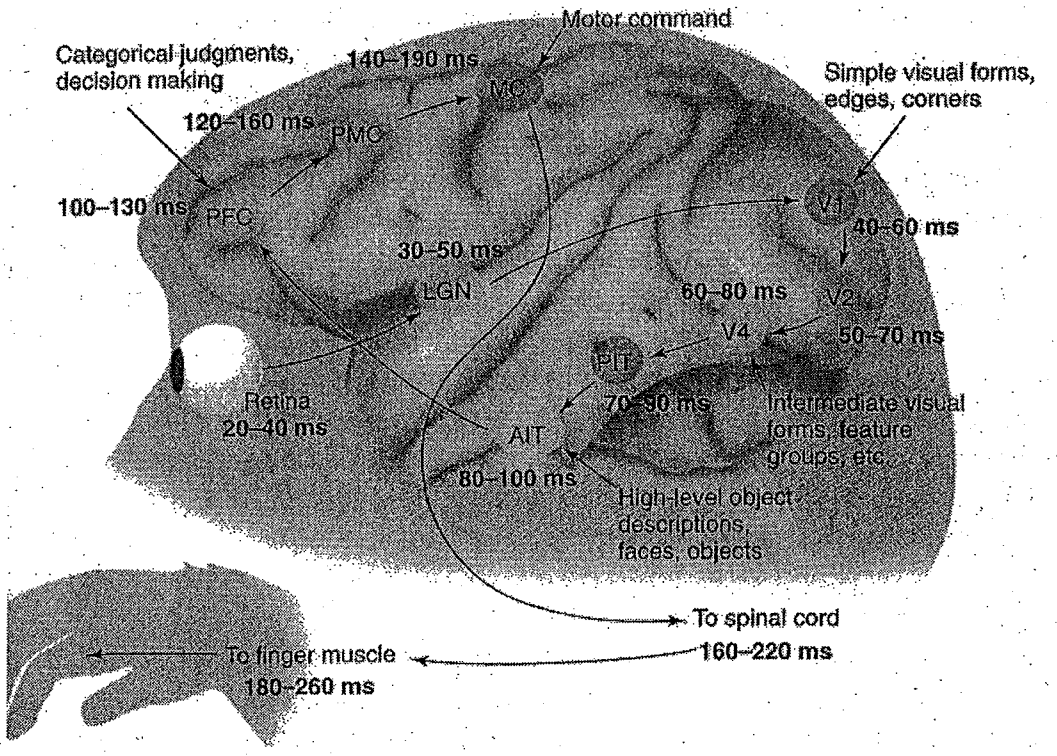

FIGURE 2. Feedforward model of signal processing in the monkey brain. According to this model, there is a feedforward flow of information from the stimulus through a hierarchy of areas in the visual system, where it is recognized and then sent to the frontal cortex, where an action plan is formulated and finally to motor cortex, where motor commands are issued to subcortical structures. This model ignores the spontaneous activity found throughout the cortex and the extensive feedback projections that accompany each feedforward connection. (Adapted from Thorpe and Fabre-Thorpe [2001].)

begins. The signal then goes from the area V1 to V2, and from V2 to V4. From V4 it then travels to the inferotemporal cortex, taking about 90 milliseconds. The visual information then travels to the front of the brain, the prefrontal cortex, where other neurons may go through another sequence, finally arriving at the motor cortex where activation occurs to produce an action.

This is a purely "feed-forward" architecture, a chain of events that occurs in linear fashion leading from sensory stimulus to motor act. It dominates the way most experiments in cognitive neuroscience are designed, especially those on awake and behaving monkeys: A monkey is restrained in a chair, given a complex sensory stimulus, and then trained to respond in particular ways while neurons are recorded from different parts of the brain. On the basis of these experiments we know about how neurons respond during reflexive tasks. The neurons in each area represent something about the task: sensory neurons represent features of the world; motor neurons represent actions, including which muscles are going to be activated; neurons in prefrontal cortex represent what is being planned, and activity may be maintained even in the absence of a sensory stimulus. The goal of this approach is to

understand what the brain represents at each stage of processing. The three questions being asked are: representation, representation, and representation.

There are, however, other types of questions that can be asked. An entirely different class of questions concerns interneuronal communication, and can be illustrated with an analogy. Imagine that we could scale up the brain so that a person would be about the size of a neuron. The brain would be about 20 miles across, about the size of New York City. Now imagine that one such neuron sitting, say, at a conference in Manhattan, has a very important message. It is representing some important fact about the world and wants to communicate it to a particular motor neuron over in the Bronx. With no direct connection, how is it going to get the message there? A single neuron is only connected to about ten thousand others, but there are many others that might need to receive its message. This is a communication problem. The brain's communication problem is even more daunting than this analogy allows since there are only about 10 million people in the New York metropolitan area, but there are 100 billion neurons in a brain.

To continue with the analogy, we'd have to pack the conference room cheek to jowl and stack people 20 miles high to mirror the 3-D structure of the brain. Imagine what it's like to be a neuron in the brain. One would be sitting in sea of people, trying to make sense of signals coming in, making decisions about what to signal out. How can we understand neurons and the brain from the perspective of the forest rather than the trees?

A window into the large-scale electrical activity in the brain has been available for nearly 100 years. Scalp recordings, called the electroencephalogram (EEG), report the summed activity from thousands of millions of synapses. Such averages can indicate whether someone is awake or asleep and can detect epileptic seizures, which typically generate very large spike and wave discharges. Although EEG recordings are helpful to clinical physicians trying to diagnose and assess brain damage, they have taught us little about how neurons represent the world. It's like a Martian trying to understand something about human beings by placing a microphone over a football stadium and recording crowd noises. The Martian might learn about touchdowns and crowd waves but not too much about human beings and the mechanisms through which they interact. It has not been able to disentangle the sources of the gross signals provided by the EEG. We believe our laboratory has solved this problem. Surprisingly, although EEG has not taught us anything about representation in the brain it may have much to teach us about the global communication network (Laughlin and Sejnowski, 2003).

PATTERNS IN THE NOISE

To describe this new methodology let's start with the "cocktail party problem." Imagine oneself at a cocktail party with other people around talking

and perhaps a band playing in the background. If the problem is to pick out one strong signal from this cacophony we might be able to do it. But if we imagine trying to pick up one weak signal from hundreds of other signals in the surrounding environment, the problem becomes significantly more difficult. It might seem that this is not possible without first knowing something about the nature of the signals. For example, suppose the signals are all white noise sources. Mixtures of noise also sound like noise. This problem in blind source separation is now solvable because of recent advances in signal processing called independent component analysis (ICA) (Makeig, Westerfield, Jung, et al., 2002). We have used a computational algorithm for ICA that was developed in my laboratory to dissect out the independent sources of signals from the brain just as we are able to isolate each of the sources of sound in the cocktail party. To solve the brain cocktail party problem, we need to record the EEG from hundreds of locations, and sort out the hundreds of sources that contribute to the EEG.

If we take all raw EEG data from individual trials and apply ICA, we obtain several dozen independent sources, as shown in FIGURE 3. Each source contains two parts, a scalp map, which is a static picture of the "weighting" of each electrode, and the time course by which the scalp map is modulated. Some sources of electrical signals in the EEG come not from the brain itself, but from eye movements, which produce EEG artifacts that are much larger than brain signals, as shown in FIGURE 3 (top left). Other artifacts involve muscle noise such as that generated by temporal muscles when gritting teeth. All of these artifacts are separated by ICA into different output channels, giving us some confidence that the technique can at least help us eliminate artifacts from EEG.

Even in the absence of a sensory stimulus, there are ongoing sources of brain rhythms, including prominent sources that oscillate at 10 Hz, called alpha rhythms. There are several sources of occipital alpha. There are others, centered over the motor cortex in the hand area, called mu rhythms, which precipitously quench after a motor act, like pressing a button, as shown in FIGURE 3 (bottom right). With ICA it is now possible to discriminate among the various alpha rhythms and pick out their unique properties and roles. For example, we have discovered that these sources and several others not in the figure do not change their amplitude in response to sensory stimuli, but rather their phase (Jung, Makieg, McKeowan, et al., 2001), as shown in FIGURE 3 (bottom middle). When a sensory stimulus is presented to a human, the ongoing rhythms become phase-shifted so that within 100 ms they are reset. When 100 single trials are averaged, the resulting average "event-related potential" (ERP) has a sequence of peaks and troughs that arise from this phase resetting of multiple sources. The traditional way of interpreting the ERP is as a sequence of overlapping activations in different brain regions. With new analysis techniques we can look at individual trials and see a different picture. When the stimulus first appears, it interacts with the ongoing background

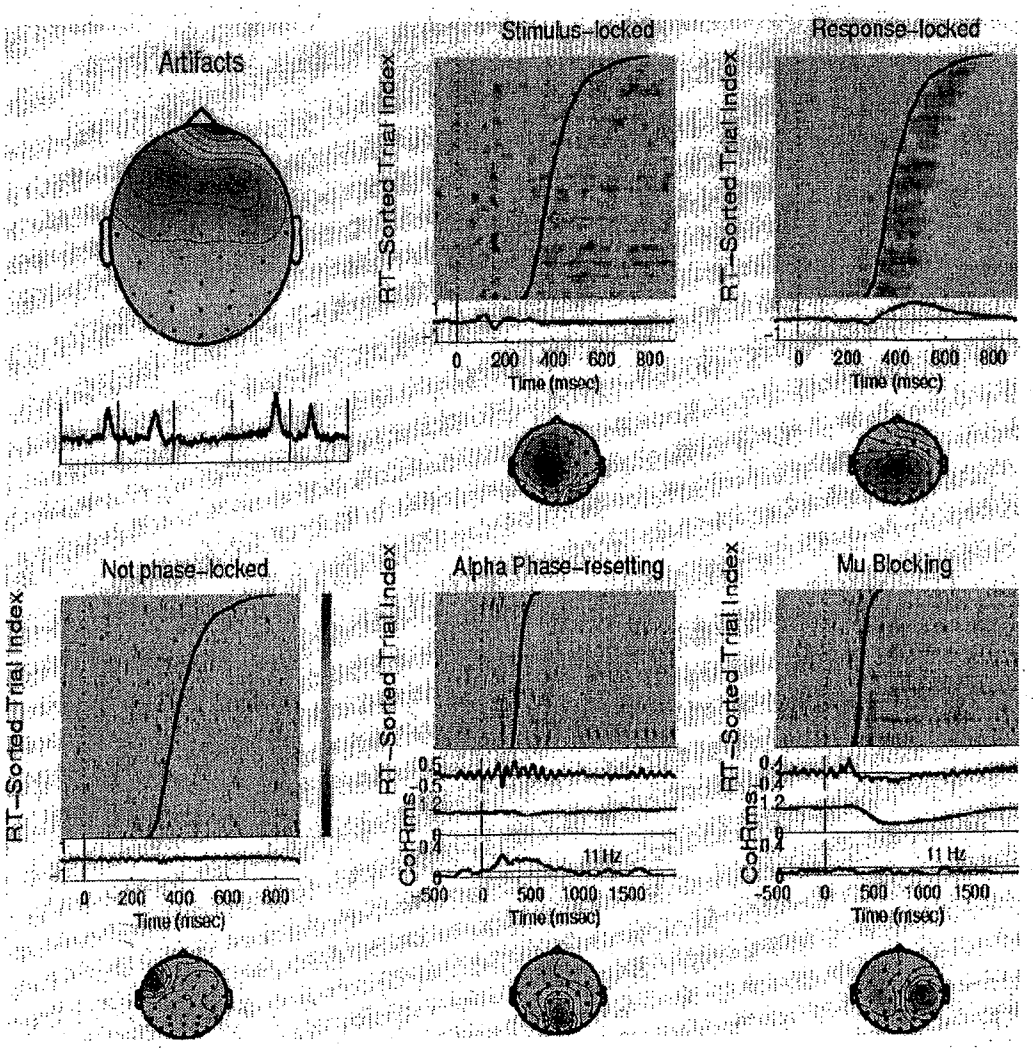

FIGURE 3. Classes of independent components derived from an ICA analysis of single event-related potential (ERP) trials from a visual reaction task. The component in the *upper left* corner is an artifact caused by an eye blink (strong localization to the front of the scalp shown above and with a large-amplitude, slow time course shown below). Each of the other panels shows the ERP image (Jung et al., 2001), formed by sorting each trial by response time (black line) and illustrating positive values of the ERP as black, zero as gray, and negative values as white. The line beneath the ERP image is the average of the ERPs. Some components are time-aligned with the sensory stimulus (*top middle*), some are aligned with the motor response (*top right*), while others are oscillatory (*bottom*). The component shown on the *bottom middle* has an ongoing 10-Hz frequency that is phase-shifted by the stimulus without changing in amplitude (trace below the average ERP). There is a systematic phase shift that increases the coherence between trials (*bottom trace*). The component on the *bottom right,* which is centered over the motor cortex, also has a 10-Hz oscillation, but this decreases in amplitude after the motor response. Some oscillatory components (*bottom left*) are not affected by the stimulus. (Adapted from Jung et al. [2001].)

EEG generators, shifting the phase, and it is the phase shifting itself that gives rise to the peaks. This tells us that we really should think of ERPs as being not separate from, but rather a property of, the ongoing background EEG, the continuous, spontaneous activity.

ATTENTION

The modulation of the EEG with state of arousal and attention suggests that it might reflect more dynamic aspects of cortical processing. Further insights into these elusive signals arise from directly recording the local field potentials (LFPs) from the cortex. The same microelectrodes that are used to record from single neurons also carry low-frequency information about population synaptic activity within a restricted region of a cortical column. Normally the LFP is filtered out with a high-pass filter, but there are further clues about how the brain regulates the flow of activity from an analysis of the LFP during an attention task in a monkey.

Fries and colleagues (2001) investigated the synchrony of neurons in area V4 that respond to visual stimuli. Monkeys were trained to fixate on a central spot and to attend to either of two stimuli presented simultaneously and at the same eccentricity. One of the stimuli fell inside the receptive field of a neuron whose activity was recorded. Thus the responses to the same stimulus could be compared in two conditions, with visual attention inside or outside the neuron's receptive field. At the same time, the local LFP was recorded from a nearby electrode. The correlations between single neurons and the neighboring population became more synchronized at high frequencies (30–70 Hz) and less so at low frequencies (0–17 Hz) when attention was directed into the receptive field of the neuron.

How can changes in the degree of correlation be linked to attention? We have shown that the observed changes in synchrony in V4 could have a significant impact on the responses of downstream neurons (Salinas and Sejnowski (2001). Compared to the firing rate of a neuron in response to independent synaptic inputs impinging randomly, even a small amount of correlation in the impinging spike trains produce more output spikes. This occurs in neurons in which the total excitatory and inhibitory inputs are roughly balanced, and as a consequence they are sensitive to the fluctuations in the membrane potential. Correlations in the inhibitory inputs, which are concentrated near the soma of cortical pyramidal neurons, are particularly effective in enhancing the firing rate of a neuron and could also serve as a mechanism for synchronizing thousands of cells in a cortical column. These experimental and modeling studies suggest that top-down spatial attention can regulate the flow of information between populations of neurons in cortical areas through correlations in their spike trains: Signals carried by neurons are boosted by increasing their degree of synchrony.

The EEG is a global measure of correlations among distant regions of the brain. Neurons that are firing spikes that are uncorrelated will result in incoherent electrical signals that will cancel at the level of the scalp. Only those populations of neurons that have a significant degree of synchrony will contribute to the EEG. This suggests that the EEG may provide valuable insights into the global regulation of information flow between the parts of the brain when the brain is engaged in a task. This may explain why the EEG has not been helpful in uncovering how information is represented in the brain—the coherent signals reflect a complex communication network that can be dynamically reconfigured by top-down planning, expectation, and attention rather than the content itself.

COMPUTATIONAL SELF

Going back to the diagram of the pathways shown in FIGURE 2, we now have a different way of understanding what might be happening in the brain during a typical stimulus–response task. First, even before the stimulus appears, there is spontaneous background activity, which is a reflection of the expectation of a stimulus. When the visual stimulus appears, it resets ongoing activity and sets off a chain of events that, under some circumstances, sets up coherent patterns and oscillations that open communication channels, allowing different parts of the brain to talk to each other. This may be how the brain solves the "Manhattan to the Bronx" problem. The key is to examine coherent activity in large populations of neurons, which can be monitored locally through the LFP and globally through the EEG.

Where is the Self in these correlated patterns of activity? It should be possible to explore this question with the techniques that have introduced here. The key will be to devise tasks that are less time-locked to external stimuli, but instead are self-generated. For example, in the block-copy task (Ballard, Hayhoe, Li, and Whitehead, (1992), the subject is shown a pattern of multicolored blocks and instructed to construct a copy of the pattern from a set of spare blocks. The subject is free to determine the order in which the blocks are picked up. How are the communications patterns between brain areas modulated during the conscious choices made during this task? Coordinated eye and hand movements are involved that go beyond simple lever presses. How is the flow of information between sensory and motor regions regulated? These issues and even more complex tasks can be explored, including ones that involve human communication. A trace of the Self should emerge from these studies.

Although we might be able to devise computational theories for the Self based on the coherent responses of neurons in different parts of the brain, will this lead us to a theory of consciousness? There may be some aspects of con-

sciousness that can be explained with these theories, such as visual awareness (Crick and Koch, 2003), but there may be others, such as the subjective aspects of consciousness, that may elude computational accounts. Ultimately the Self may be found by looking more closely at the brain's spontaneous activity, a part of the background that we have ignored for too long. During sleep the background in the cortex becomes more globally coherent than during states of alertness (Destexhe and Sejnowski, 2001). What changes during sleep states is the pattern of activity, and it is in these patterns that traces of the Self may be found.

REFERENCES

CHURCHLAND, P.S. & SEJNOWSKI, T.J. (1992). *The computational brain.* Cambridge, MA: MIT Press.

BALLARD. D.H., HAYHOE, M.M., LI, F. & WHITEHEAD, S.D. (1992). Hand-eye coordination during sequential tasks. *Philosophical Transactions of the Royal Society of London, Series B Biological Sciences, 337,* 331–338.

CRICK, F. & KOCH, C. (2003). A framework for consciousness. *Nature Neuroscience, 6,* 119–126.

DESTEXHE, A. & SEJNOWSKI, T.J. (2001). *Thalamocortical assemblies: How ion channels, single neurons and large-scale networks organize sleep oscillations.* Oxford: Oxford University Press.

FRIES, P., REYNOLDS, J.H., RORIE, A.E. & DESIMONE, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science, 291,* 1506–1507.

JUNG, T.-P., MAKEIG, S., MCKEOWN, M.J., BELL, A.J., LEE, T.-W. & SEJNOWSKI, T.J. (2001). Imaging brain dynamics using independent component analysis, *Proceedings of the IEEE, 89,* 1107–1122.

LAUGHLIN, S.B. & SEJNOWSKI, T.J. (September 26, 2003). Communication in neuronal networks. *Science.*

MAKEIG, S., WESTERFIELD, M., JUNG, T.-P., ENGHOFF, S., TOWNSEND, J., COURCHESNE, E. & SEJNOWSKI, T.J. (2002). Dynamic brain sources of visual evoked responses. *Science, 295,* 690–694.

SALINAS, E. & SEJNOWSKI, T.J. (2001). Correlated neuronal activity and the flow of neural information. *Nature Reviews Neuroscience, 2,* 539–550.

THORPE, S.J. & FABRE-THORPE, M. (2001). Seeking categories in the brain. *Science, 291,* 260–263.