

Vector Subtraction Implemented Neurally: A Neurocomputational Model of Some Sequential Cognitive and Conscious Processes

John Bickle,*¹ Cindy Worley,† and Marica Bernstein‡

**Department of Philosophy and Program in Neuroscience, East Carolina University, Greenville, North Carolina 27858;* †*Department of Neuroscience, Bowman Gray College of Medicine, Wake Forest University, Winston-Salem, North Carolina 27106;* ‡*Department of Biology and Program in Neuroscience, East Carolina University, and Department of Physiology, East Carolina University School of Medicine, Greenville, North Carolina 27858*

Although great progress in neuroanatomy and physiology has occurred lately, we still cannot go directly to those levels to discover the neural mechanisms of higher cognition and consciousness. But we can use neurocomputational methods based on these details to push this project forward. Here we describe *vector subtraction* as an operation that computes sequential paths through high-dimensional vector spaces. Vector-space interpretations of network activity patterns are a fruitful resource in recent computational neuroscience. Vector subtraction also appears to be implemented neurally in primate frontal eye field activity, which computes dimensions of saccadic eye movements. We use this apparent neural implementation as a model and construct from it a general neurocomputational account of an important type of sequential cognitive and conscious process. We defend the biological plausibility of all components of the general model and show that it yields testable anatomical and physiological predictions. We close by suggesting some interesting consequences for consciousness if our model characterizes correctly the neural mechanisms producing a common type of episode in our conscious streams. © 2000 Academic Press

Key Words: Jamesian conscious stream; neural network; vector subtraction; frontal eye fields (FEFs); working memory.

INTRODUCTION

Higher cognition abounds in serial, sequential processes. Here we are concerned with a type possessing the following characteristics: (1) they are extended through time; (2) they proceed in an orderly fashion, from one unified representation or idea to another; and (3) later steps in the sequence depend upon the contents of its earlier states and of prespecified upcoming “target” states. For example, different sequences of ideas might be necessary in distinct instances of a problem-solving task aimed at the same goal state, depending on where earlier ideas land one in the problem space. Finally, (4) temporal limits often require multiple steps in a sequence to be computed in advance and run off “ballistically,” with little opportunity for revision or feedback once the sequence is initiated. In Daniel Dennett’s (1991) picturesque terminology, “Ballistic acts are unguided missiles. Once they are triggered, their trajectories are not adjustable” (p. 145). We are all familiar with initiating a complex linguistic

¹ Address correspondence and reprint requests to John Bickle, Focused Research Program in Computational Neuroscience, Brewster A-327, East Carolina University, Greenville, NC 27858–4353. E-mail: bicklej@mail.ccu.edu.

utterance, realizing mid-sequence that “this isn’t what I meant to say”—yet it is too late to halt or revise that utterance (alas!).

These four features are also apparent in a common type of episode in our phenomenological “conscious streams.” William James described this phenomenon beautifully more than a century ago. Reflecting on one’s conscious awareness of a thunderclap, James describes the continuity of temporal stages in the stream: “Into the awareness of the thunderclap itself the awareness of the previous silence creeps and continues; for what we hear when the thunder crashes is not thunder pure, but thunder-breaking-upon-silence-and-contrasting-with-it” (James 1890, p. 240). He describes the phenomenological effect of earlier events in the stream on the character and content of later ones: “Our feeling of the same objective thunder, coming in this way, is quite different from what it would be were the thunder a continuation of previous thunder” (pp. 240–241). He generalizes these features across presentation modalities: “[I]t would be difficult to find in the actual concrete consciousness of man a feeling so limited to the present as not to have an inkling of anything that went before” (p. 241).

Being staunch neural physicalists about mind, we believe that the mechanisms producing all types of cognition and consciousness will be fully explained by brain science. But philosophical conviction is no substitute for testable theory. *How* does the brain generate these processes? We are not here asking an “in principle” question purely for armchair reflection. Ours is ultimately a question for theoretical neurobiology. At present, the known neuroanatomy and physiology of cognition and consciousness is too underdeveloped to answer our question. But we can push this project forward using the resources of computational neuroscience. We describe a mathematical operation that computes sequences possessing the four features of some common cognitive and conscious processes presented above. We show how this operation can be implemented by activity patterns and dynamics in artificial neural networks and present empirical evidence suggesting that a region of primate frontal cortex implements it. We generalize from this known neural implementation to develop a neurocomputational model that offers testable anatomical and physiological predictions about the mechanisms of one type of conscious and cognitive process.

Not all cognitive processes nor temporally extended episodes in our “Jamesian” conscious streams possess these four features. Those that do not fall beyond the scope of the model presented here. The key is the third feature mentioned above. Under “Generalizing from this Implementation: A Neurocomputational Network That Computes Multiple-Step Sequences” we emphasize how much of cognition and consciousness processes this feature.

THEORETICAL BACKGROUND: VECTOR-SPACE INTERPRETATIONS OF NEURAL NETWORK ACTIVITY

Over the past 2 decades, cognitive neuroscientists have exploited a powerful mathematical resource for characterizing neural representations and computations. It provides a natural interpretation of activity patterns and dynamics in *neural networks*, both biological and artificial. Neural networks have moved to center stage in cognitive science over the past decade, so repeating the details of their components and opera-

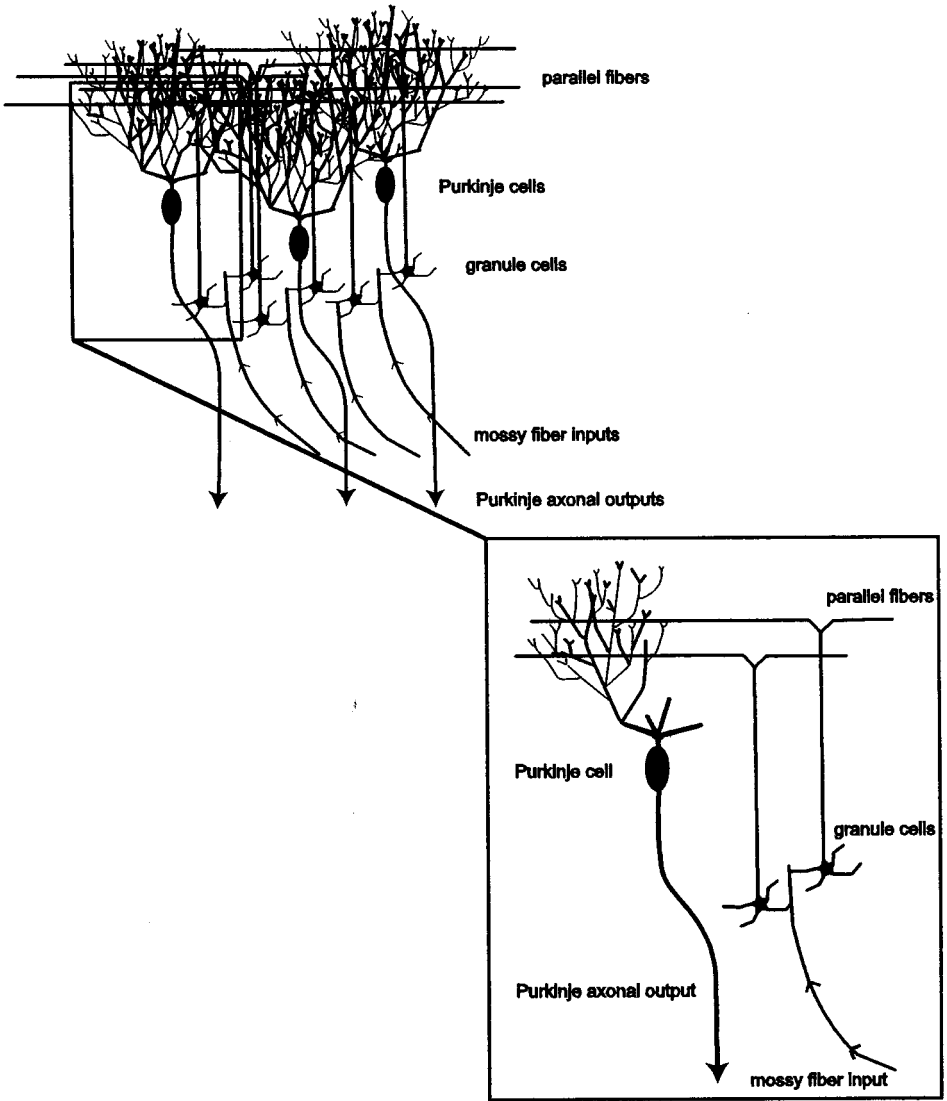


FIG. 1. Schematic illustration of a cerebellar network. Insert reveals the detailed structure of a single Purkinje neuron.

tion is not necessary here. What is less well known is the vector-space interpretation of their activity. On this account, *activity vectors* constitute the brain's principal type of representation and *vector-to-vector transformations* constitute its principal type of computation (Churchland & Sejnowski, 1992). Consider the schematic biological network in Fig. 1. It reflects a common architecture in the cerebellum (Llinás, 1975). The illustrated network contains four input lines (the parallel fibers originating out-

side the network), three computing units (the Purkinje neurons), and three output lines (the axons from the Purkinje neurons projecting outside the network). Each input line carries a signal strength (reflecting, e.g., frequency of action potentials). The input to this network thus constitutes a four-element activity vector, with each value representing the signal strength on one input line. Each neuron computes its new activation value, depending on the input vector, the weight value at each synapse, and its previous activity value. This new value is transmitted as a signal strength (e.g., frequency of action potentials) down the neuron's output line (axon). The resulting set of new activity or output values in the network's neurons constitutes a three-element activity or output vector. Each value represents the new activity or output rate in one neuron. Activity patterns across neural networks constitute vector-to-vector transformations. In our simple example, network activity transforms a four-element input vector into a three-element activity or output vector.¹

We can graph these vector representations and transformations in an *activation vector space*, with a separate dimension for the activation rate of each neuron in the network (or in a relevant subset like those in a specific column or layer). We can represent activity or output vectors in the simple network of Fig. 1 on a three-dimensional activation space. A vector of activity values, one from each neuron in the network, is a point in the space. Geometrical properties of the space and its partitions reflect features of and relations between activity patterns and dynamics in the neural network (Churchland & Sejnowski, 1992.). For example, in a network whose output categorizes inputs into a variety of types, activity or output patterns generated to input vectors belonging to the same type will be similar to one another. Thus the (hyper-) points representing these activity vectors will lie in close proximity to one another in the activity vector space. Measurements of their geometrical proximity can provide quantitative analyses of, e.g., similarity to a prototype (Churchland, 1989).

We need not limit our use of this resource to activity vector spaces. We can interpret the dimensions of a vector space along any number of biological or representational parameters. For example, Paul Churchland (1995), drawing on earlier work by Susan Brennan, interprets patterns of activity in an artificial neural network designed to recognize faces using a 10-dimensional "face-feature" space. Each dimension represents the value of one quantifiable facial feature (e.g., nose width, eye separation). A point in the 10D space represents a face with that quantity of each of the ten features. He shows how the artificial neural network encoding this face space in its trained activity patterns can characterize prototype faces from a set of examples and construct caricatures by manipulating the distance between a face's point in the space and the location of the prototype.

¹ The popularity of this level of analysis among computational *neuroscientists* is declining rapidly. *Compartmental* modeling enables modelers to mimic computations in and interactions between patches of neuronal membrane. Hence the topology of membrane structure in individual neurons, variations in ionic channels distributed across membrane patches, locations of synapses on dendrites, axons, and soma (e.g., distances from axon hillock), and a host of other biophysical properties that determine frequencies of action potentials in real neurons are accessible to modeling control and manipulation. Modelers can "custom-build" the neurons in their simulated networks to any desired level of neurobiological realism without sacrificing the capacity to track circuit properties and activity dynamics across the entire network. But the vector-space interpretation still stands. With compartmental modeling, not only are neural networks interpretable as vector transformers. The neurons comprising them are too.

Viewed from this interpretation, the representational and computational power of the primate brain is inspiring (Churchland, 1989). Biological neural networks possess features lacking in the two-layer “perceptions” criticized by Minsky and Papert (1969) for their computational limitations. Functions implemented in real neurons to compute new activity and output rates are extensively nonlinear. Hence biological networks interpreted as vector transformers can compute nonlinear input–output functions, broadening dramatically their computational range. Biological networks also implement a variety of computational architectures, employing layers, columns, recurrency (feedback), and a variety of excitatory, inhibitory, and modulatory synapses. Paul Churchland expresses eloquently the brain’s representational capacity when its activity is interpreted in this way:

Given high dimensional spaces, which the brain has in abundance, those spaces and the prototypes they embody can encompass categories of great complexity, generality, and abstraction, including those with a temporal dimension, such as harmonic oscillator, projectile, traveling wave, Samba, twelve-bar blues, democratic election, six-course dinner, courtship, elephant hunt, civil disobedience, and stellar collapse. . . . In principle, then, it is no harder for such a system to represent types of processes, procedures, and techniques, then to represent the “simple” sensory qualities. From the point of view of the brain, these are just more high-dimensional vectors. (1989, p. 191)

Given the architectural complexity of the human brain, no concept seems inexpressible as a point or subvolume in a neurally implementable high-dimensional vector space.

Vector spaces also provide a helpful resource for representing sequential processing in neural networks. (Churchland alludes to this at the end of the above quotation.) On the vector-space interpretation, a sequential transformation in a neural network from one representation to another is a *path* through (hyper-) points or subvolumes in the appropriate vector space. Figure 2 illustrates this. A sequential transformation from representation 1 to 2 to 3 is represented as the path from the network activation pattern located at (hyper-) point 1 to that located at (hyper-) point 2 to that at (hyper-) point 3. In the spirit of Churchland’s passage, any sequence of representations that can be expressed as a path through points or subvolumes in an appropriate vector space is a process that can be implemented by an appropriately structured biological neural network.

FEATURES OF SEQUENCES COMPUTED BY VECTOR SUBTRACTION

Our goal is (1) a *neurocomputational* operation that (2) could generate cognitive and conscious processes possessing the four sequential features emphasized under “Introduction.” In this section we argue that *vector subtraction in the appropriate space* meets condition 2. Since we aim to make our model accessible to an interdisciplinary audience, we describe this mathematical operation in explicit detail and use simple examples. We apologize in advance to mathematically sophisticated readers, but see no other way to realize our aim.

Figure 3 illustrates vector subtraction in a two-dimensional Cartesian space. We can compute the dimensions of the second vector in the sequence ($A \rightarrow B$) using

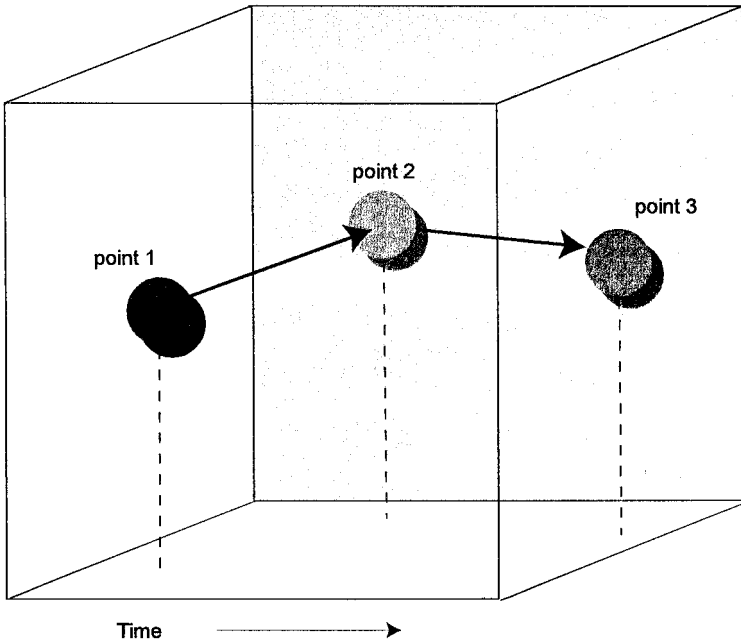


FIG. 2. Graph of a sequential process in a three-dimensional vector space. The neural network implementing this process proceeded from the activity pattern located at 1 to that at 2 to that at 3. In keeping with the account of cognitive representation discussed in the text, each point reflects a cognitive state with a specific representational content.

the dimensions of the first vector (origin $O \rightarrow A$) and of the second target from origin O ($O \rightarrow B$). By the law of parallelograms we know that

$$O \rightarrow B = O \rightarrow A + A \rightarrow B.$$

Rearranging this equation yields

$$A \rightarrow B = O \rightarrow B - O \rightarrow A.$$

Plugging in the values from Fig. 3 yields

$$A \rightarrow B = \langle 4, 7 \rangle - \langle -1, 3 \rangle.$$

Solving yields

$$A \rightarrow B = \langle 5, 4 \rangle.$$

These are exactly the dimensions of the vector necessary to get to B from A.

Vector subtraction is an iterative operation. Its general equation is

$$N - 1 \rightarrow N = O \rightarrow N - N-2 \rightarrow N - 1 - \dots - O \rightarrow A. \quad (1)$$

Applying this equation to Fig. 3 yields

$$B \rightarrow C = O \rightarrow C - A \rightarrow B - O \rightarrow A.$$

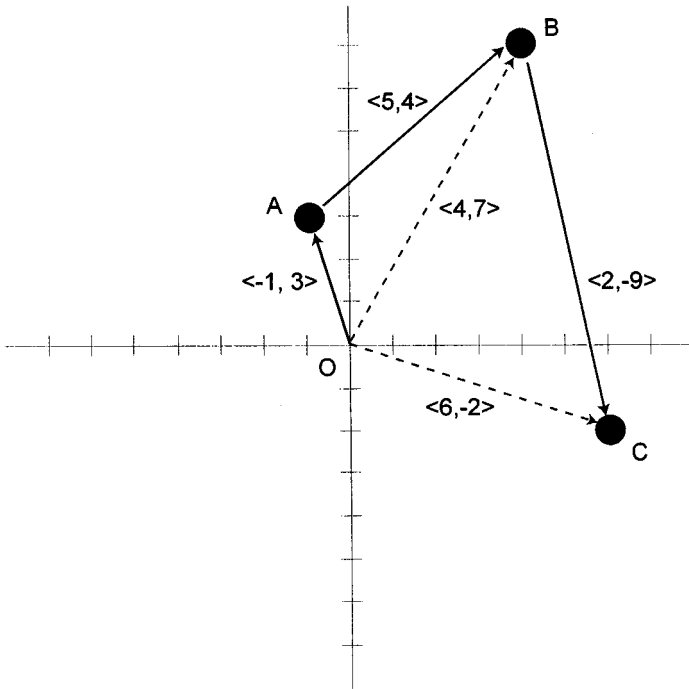


FIG. 3. Graph illustrating the vector subtraction computation described in the text. Solid lines reflect vectors between points whose dimensions were computed by vector subtraction. Dashed lines reflect vector distances between origin and points.

Plugging in the values yields

$$B \rightarrow C = \langle 6, -2 \rangle - \langle 5, 4 \rangle - \langle -1, 3 \rangle.$$

Solving yields

$$B \rightarrow C = \langle 2, -9 \rangle.$$

These are exactly the dimensions of the vector needed to get from B to C.

Thus vector subtraction is an operation by which sequential paths through vector spaces can be computed. We saw at the end of the previous section that sequential processing in neural networks can be characterized as paths through vector spaces. Since vector subtraction uses only information about previous steps in the sequence and the location of the next target from the origin, it can compute multistep sequences through the space. This makes it an appropriate operation for “ballistic” phenomena where time constraints don’t permit “reflection” and recomputation based on explicit feedback from earlier steps. Vector subtraction also uses the contents of earlier steps in the sequence, characterized as vectors between representations (hyper-points or subvolumes) in the appropriate vector space, and the contents of upcoming “target” states (their locations in the vector space), to determine the dimensions of later steps (vectors). Referring to Fig. 3 above, the vector dimensions necessary to reach point

C after the sequence from O to A to B are quite different from those necessary to reach C from O. These are the four features of some higher cognitive processes and episodes in the “Jamesian” conscious streams we seek to explain neurocomputationally.

Combining the results of this and the previous section, we can now see why vector subtraction appears to be a promising operation for a *computational* theory of some cognitive and conscious processes. But could it be a way that *the brain* computes such sequences? Can it assist us in searching for the *actual neural mechanisms* underlying such processes? Before we can develop affirmative answers to these questions, we first need to present evidence that the primate brain implements vector subtraction. We then use the details of this neural implementation to generate a neurocomputational model of sequential cognitive and conscious processing that offers testable anatomical and physiological predictions.

VECTOR SUBTRACTION IMPLEMENTED NEURALLY: SACCADE COMMAND EXCITATION IN PRIMATE FRONTAL EYE FIELDS

Saccades are a type of eye movement that bring visual targets onto the fovea, the area of the retina where visual acuity is greatest. When humans foveate a visual target that suddenly moves away, their eyes maintain initial fixation for about 200 ms and then move (“saccade”) quickly to refoveate the target. Even when a visual target remains stationary, humans saccade constantly to gather information about its different features. Unlike other kinds of rapid eye movements, saccades do not require a visual stimulus. They can be made accurately to stimuli of other sensory modalities, to remembered locations, and to verbal commands. Although they are often executed involuntarily, saccades are under voluntary control. Humans saccade on average three times per second, even during activities that don’t use visual stimuli (e.g., doing mental arithmetic) (Goldberg et al., 1992).

The primate saccade-generating system includes a number of cortical and subcortical regions. The circuitry is arranged hierarchically, but there are multiple processing streams and extensive recurrency. (See Goldberg et al., 1992 for an overview of the complete primate saccade-generating system.) Here we focus on the generation of saccade command messages in the *frontal eye fields* (FEFs) of premotor frontal cortex (subdivisions of Brodmann’s area 8). Clinical neuropsychological, neurological, and anatomical evidence has suggested for some time that these frontal cortical regions participate in saccade planning and control (Kolb & Whishaw, 1990; Barbas & Mesulam, 1981; Goldberg et al., 1992). Bruce and Goldberg (1985), Bruce et al. (1985), and Goldberg and Bruce (1990) provided a thorough single-cell physiological and functional analysis of saccade-related activity in FEFs. They trained rhesus monkeys on a variety of saccade-related tasks and recorded from single FEF neurons. They discovered both presaccadic and postsaccadic activity. Presaccadic activity began 50–200 ms prior to saccade onset. Although presaccadically active FEF neurons differed in their individual responses to visual stimuli, together they formed a network receiving retinotopic input from visual association cortex and delivering motor (eye movement) command output to superior colliculi. Individual FEF neurons displayed *movement fields* analogous to the receptive fields of sensory neurons. A neuron’s

presaccadic movement field is the set of eye movements (directions and amplitudes) prior to which the neuron is active. Typical FEF neurons have large presaccadic movement fields, firing prior to eye movements with a variety of amplitudes and directions. But a given neuron was maximally active (highest frequency of action potentials) prior to saccades *of a single particular amplitude and direction*. (This is another way that movement fields are analogous to sensory receptive fields.) Hence motor command messages from FEFs to superior colliculi are coarsely coded, presumably by vector averaging of active neurons' optimal saccade dimensions times activity rates.

Using the *double-step saccade paradigm* developed by Sparks and his colleagues (Mays & Sparks, 1980), Goldberg and Bruce (1990) provided convincing evidence that FEF presaccadic output is coded in oculomotor (eye movement) rather than retinotopic (visual) coordinates. They trained rhesus monkeys to fixate a light (at fixation point FP). The light was extinguished and was followed by a sequence of two lights somewhere in the periphery. Stimulus A appeared and was extinguished, followed by stimulus B at a different location. Both A and B appeared and were extinguished before the monkey initiated the saccade to A. Monkeys were only rewarded for first saccading to the point occupied by A and then to the point occupied by B. This paradigm permits a *dissonant saccade test*, where the *visual vector* to the retinotopic location of B from FP is inconsonant with the *movement vector* necessary to foveate B after the saccade to A. One can measure presaccadic activity of single FEF neurons on a *Right-Eye Movement/Wrong-Receptive Field dissonant double-step task*, where the eye movement vector $A \rightarrow B$ is optimal for the presaccadic cell being recorded from, but neither stimuli (A nor B) appear within the cell's visual receptive field (see Fig. 4). Optimal response by FEF neurons prior to their preferred eye movement (amplitude and direction) in the absence of visual stimuli in their receptive field would show that presaccadic activity is coded in oculomotor rather than retinotopic coordinates.

These were exactly the results Goldberg and Bruce (1990) observed in 65 of the 83 presaccadic FEF cells they studied. Even more than half of the cells that responded presaccadically only in the presence of visual stimuli (and never before spontaneous saccades or saccades to a learned location without a visual stimulus) displayed significant activity prior to the $A \rightarrow B$ saccade in the Right-Eye Movement/Wrong-Receptive Field dissonant double-step task. In their own words, presaccadic processing in FEFs

begins with visual input, enables this visual input to be a major but not the only input from which a presaccadic signal is generated, and ends with a saccade movement signal devoid of antecedent sensory information. This movement signal then projects to the saccade effector system in the superior colliculi. (1990, p. 503)

Some of their data even suggested that this coordinate transformation occurs in individual FEF neurons.

Bruce and Goldberg (1985) and Goldberg and Bruce (1990) also discovered *postsaccadic* activity in individual FEF neurons. Postsaccadic discharge (over baseline rate) begins after saccade initiation. Roughly one-half of the postsaccadic neurons studied had tonic discharges, firing until the monkey initiated another saccade. In

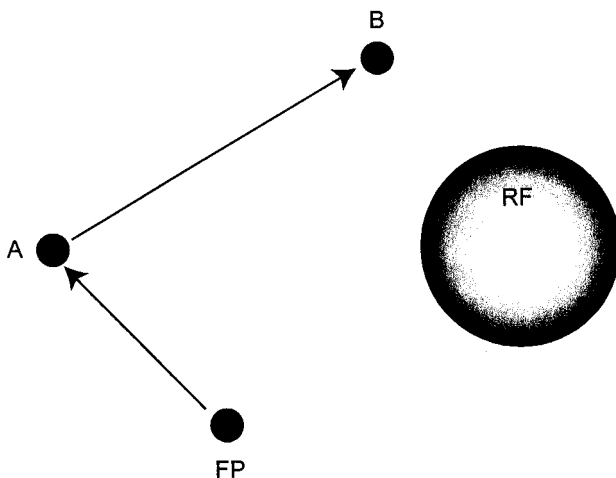


FIG. 4. Illustration of the Right-Eye Movement/Wrong-Receptive Field dissonant double-step saccade task described in the text. RF indicates the receptive field (in retinotopic dimensions) of the FEF neuron being recorded from. $A \rightarrow B$ reflects the neuron's presaccadic optimal eye movement (in oculomotor dimensions). Even though neither visual stimulus (A or B) occurs within the neuron's receptive field, typical FEF neurons exhibit maximum activity 50–200 ms prior to the second (preferred) saccade. This result indicates that FEF presaccadic activity is coded in oculomotor rather than retinotopic coordinates.

some cells, this tonic firing continued for longer than 1 s if experimenters delayed artificially the next saccade. A neuron's postsaccadic movement field is the set of all saccades (amplitude and direction) after which the neuron's activity increases over its baseline rate. Typical postsaccadic movement fields were large, but also displayed optimal activity after saccades of a single amplitude and direction. Postsaccadic activity occurred after every type of saccade, including spontaneous saccades made in total darkness (which typically are not preceded by FEF activity). Goldberg and Bruce (1990) suggest that postsaccadic FEF activity indicates, in coarsely coded fashion, the dimensions (amplitude and direction) of the saccade just executed. They suggest that an "efference copy" of the presaccadic message projected back to FEFs from superior colliculi drives this postsaccadic activity.

Based on this pre- and postsaccadic activity, Goldberg and Bruce (1990) suggest that the FEFs compute the oculomotor coordinates of saccade sequences using *vector subtraction*. In the double-step task, the retinotopic dimensions of stimulus B from fixation point FP ($FP \rightarrow B$) are coded neurally in the activity of FEF neurons that respond primarily to visual input. The oculomotor dimensions of the first saccade from FP to A ($FP \rightarrow A$) are coded in postsaccadic FEF activity. Recall from the discussion of vector subtraction under "Features of Sequences Computed by Vector Subtraction" that this is the information necessary to compute the dimensions of the next vector:

$$A \rightarrow B = (FP \rightarrow B) - (FP \rightarrow A).$$

Applied to saccade command generation, the value $A \rightarrow B$ will code the dimensions of the appropriate second saccade in "eye movement" space. If the intra-FEF net-

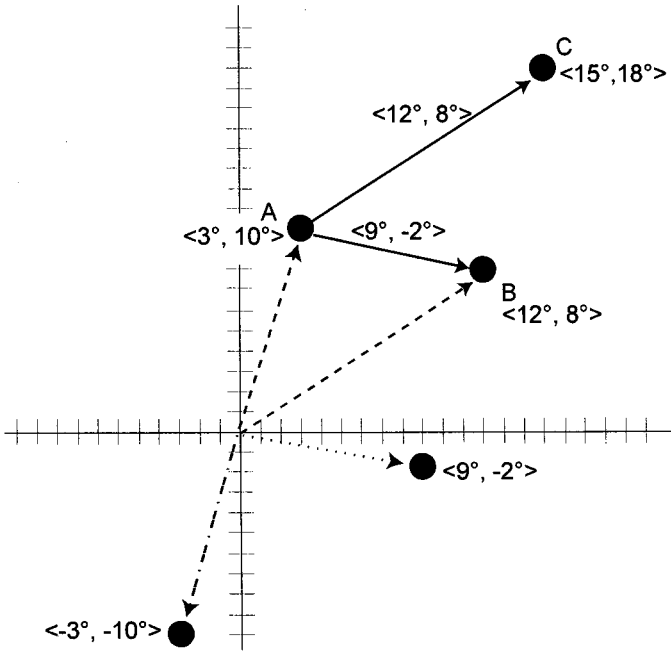


FIG. 5. Computation of the second saccade ($A \rightarrow B$) in eye-movement space. Solid lines reflect saccades from A computed by vector addition ($A \rightarrow C$) and vector subtraction ($A \rightarrow B$). Dashed lines reflect dimensions of postsaccadic activity following $FP \rightarrow A$. Dotted line reflects the dimensions of $A \rightarrow B$ from origin computed by vector subtraction, implemented by the mechanisms of pre- and post-saccadic activity in primate FEF neurons.

work connectivities are present to combine properly this physiologically encoded information, then the FEFs themselves compute the dimensions of the second saccade via vector subtraction.

Figure 5 diagrams an eye location coordinate space (borrowing coordinates from Lee, Rohrer, & Sparks, 1988). Axes represent horizontal and vertical endpoints of visually guided saccades from FP ($\langle 0^\circ, 0^\circ \rangle$). Consider two stimuli in a double-step saccade task, with A at $\langle 3^\circ, 10^\circ \rangle$ and B at $\langle 12^\circ, 8^\circ \rangle$. The retinotopic dimensions to B from FP, that is, a saccade with dimensions $\langle 12^\circ, 8^\circ \rangle$, will not generate the appropriate second saccade. From A, that will land the subject's fovea on C, at $\langle 15^\circ, 18^\circ \rangle$ (the value of $FP \rightarrow A + FP \rightarrow B$). To get from A to B, the subject must execute a saccade with dimensions $\langle 9^\circ, -2^\circ \rangle$. Notice that vector subtraction

$$A \rightarrow B = FP \rightarrow B - FP \rightarrow A$$

yields a value of exactly those dimensions

$$\begin{aligned} &= \langle 12^\circ, 8^\circ \rangle - \langle 3^\circ, 10^\circ \rangle \\ &= \langle 9^\circ, -2^\circ \rangle \end{aligned}$$

from information encoded neurally in FEF visual ($FP \rightarrow B$) and postsaccadic ($FP \rightarrow A$) activity. Goldberg and Bruce (1990) thus insert a "black box" FEF vector subtraction mechanism into their computational diagram of saccade generation.

One would still like to know *how* the FEFs compute vector subtraction, and a “black box” mechanism doesn’t answer that question. An additional result from Goldberg and Bruce’s (1990) study of postsaccadic FEF activity yields an enticing clue (though they don’t note this possibility). Nearly one-third of presaccadic FEF neurons studied in the double-step saccade task (27/83) also displayed postsaccadic activity. The relationship between these “dual-response” cells’ pre- and postsaccadic activity was especially interesting. A cell’s optimal postsaccadic activity typically followed saccades *of exactly the opposite amplitude and direction* from its optimal presaccadic activity. For example, if one of these neurons fired optimally before short saccades up and to the right, then it fired optimally after short saccades down and to the left. This suggests an elegant implementation of vector subtraction in the FEFs. Instead of *subtracting* the dimensions of $FP \rightarrow A$ from those of $FP \rightarrow B$, a computational mechanism can simply *add* to the latter the dimensions of the vector coarsely coded by neurons whose presaccadic activity codes for a saccade exactly opposite of the one just executed $-(FP \rightarrow A)$. These are the neurons that will be active postsaccadically following the first saccade (presumably via the efference copy of the presaccadic message projected back to FEFs from superior colliculi). Vector summation is a common component of many neurocomputational models. Using the example in Fig. 5, the dimensions of the second saccade ($A \rightarrow B$) results from this computation:

$$\begin{aligned} A \rightarrow B &= (FP \rightarrow B) + -(FP \rightarrow A) \\ &= \langle 12^\circ, 8^\circ \rangle + \langle -3^\circ, -10^\circ \rangle \\ &= \langle 9^\circ, -2^\circ \rangle. \end{aligned}$$

Cellular summation of the retinotopic next target location from fixation point and the oculomotor dimensions of the postsaccadic activity in these dual-response FEF neurons is sufficient to compute the oculomotor dimensions of the next saccade.

This geometric relationship between pre- and postsaccadic movement fields in FEF neurons, and the neural implementation of vector subtraction that it suggests, is potentially an important neurobiological discovery. Response fields have been documented experimentally in sensory, motor, and memorial neurons. It is possible that these other types exhibit *postsensation*, *-movement*, or *-memorial* activity with interesting geometrical relationships to their *presensation*, *-movement*, or *-memorial* fields. A host of neurocomputational insights might lie dormant in the information coded by neuron’s response properties *after* their initial (pre-whatever) activity. This idea has potential impact beyond the specific application we are about to make of it.

GENERALIZING FROM THIS IMPLEMENTATION: A NEUROCOMPUTATIONAL NETWORK THAT COMPUTES MULTIPLE-STEP SEQUENCES

We have constructed a neurocomputational network that generates sequential outputs based on the neural implementation of vector subtraction in the primate FEFs. Before presenting it, we hope to circumvent a potential misunderstanding. Saccade command generation is neither cognitive nor (typically) conscious. How then can a neurocomputational model of this process help us uncover neural mechanisms of

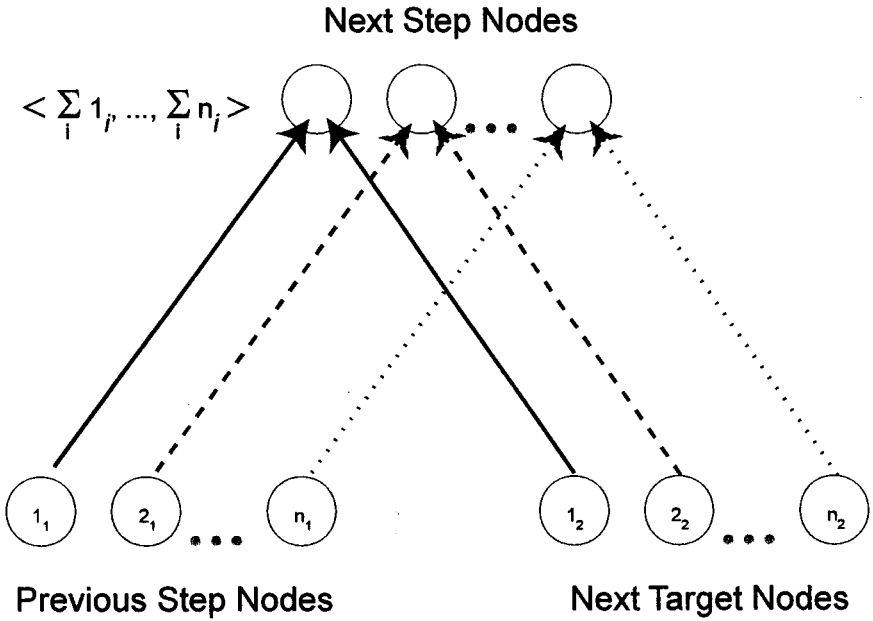


FIG. 6. Vector Subtraction Core generalized from the known neural implementation of vector subtraction in primate FEFs.

either? Our answer is that we are using this cell-physiological implementation of vector subtraction as a “model” in the old-fashioned scientific sense. The primate FEFs are part of a neuronal system that produces outputs analogous in important *structural* fashion to some higher cognitive and conscious processes. Multiple-step sequences of saccade commands have the four sequential features discussed under “Introduction” and “Features of Sequences Computed by Vector Subtraction”. We understand the neural computation that generates these sequences in the primate FEFs and the cell properties and connectivities that implement it there. We now construct a neurocomputational model, based as far as possible on the anatomical and physiological details, that computes the more elaborate sequences comprising some cognitive and conscious processes. We will have to introduce capacities and components beyond those sufficient to compute saccade command sequences and these will require biological justification. We provide this justification and show how our general model provides *testable* hypotheses about the neural mechanisms of some cognitive and conscious processes. Our model assumes the vector-space account of cognitive representation and content presented under “Theoretical Background: Vector-Space Interpretations of Neural Network Activity.” A cognitive process is a sequential path through a vector space from one (hyper-) point to others. Location of a (hyper-) point in the appropriate vector space constitutes a state’s content.

Since our full model has multiple components, we present it in stages with detailed examples of the computations. The first component is the Vector Subtraction Core, which mimics the neural implementation in primate FEFs (see Fig. 6). The Core contains three sets of processing nodes: (i) Next Target nodes, coding for the location

of the next target in the appropriate vector space; (ii) Previous Step nodes, coding for the dimensions of the previous step in the path through the vector space; and (iii) Next Step nodes, coding for the dimensions of the next step (from the last location to the next target). There is one node in each set for each dimension of the appropriate vector space. Activity in the Previous Step nodes represents values exactly opposite those coding for the dimensions of the step just executed, in keeping with the neural implementation of vector subtraction suggested by FEF postsaccadic activity. The dimensions of the next step are the vector sum of the previous step and next target.

In the example of Fig. 5, each set in the Vector Subtraction Core will contain two nodes, connected appropriately (in the terminology of Fig. 6, $n = 2$). To compute the dimensions of $A \rightarrow B$, activity in the Next Target nodes will represent values of 12 and 8, respectively. Activity in the Previous Step nodes will represent values of -3 and -10 , respectively. Summing these values yields activity in the Next Step nodes representing a movement in the two-dimensional vector space with dimensions $\langle 9, -2 \rangle$.

We cannot limit executable sequences by our model to only two steps, since the sequences comprising cognitive processes and conscious streams typically involve many more. We saw under “Featurer of Sequences Computed by Vector Subtraction” that vector subtraction iterates with information about previous vectors in the sequence. We thus need to supplement the Vector Subtraction Core with a “working memory” that stores and utilizes dimensions of earlier steps in the sequence. A Working Memory Store structured like that illustrated in Fig. 7 provides this. After the dimensions of the previous step occur in the Previous Step nodes, they move to the limited-capacity Working Memory Store. From there they continue as input to the Next Step nodes until they exhaust their time limit in working memory. The number of layers in the Working Memory Store represents the temporal dimensions of working memory for the task at hand. When its time is exhausted, information about a previous step will either be lost or forwarded to longer term memory stores.

Extending the example of Fig. 5 illustrates this new component (see Fig. 9 below.) Consider a sequence of four steps through a two-dimensional space: from FP $\langle 0, 0 \rangle$ to A $\langle 3, 10 \rangle$ to B $\langle 12, 8 \rangle$ to C $\langle 5, -4 \rangle$ to D $\langle -2, -2 \rangle$. In the step $FP \rightarrow A$, the Next Step nodes only receive input from the Next Target nodes ($\langle 3, 10 \rangle$). The system executes a step from the origin through the state space with those dimensions, coming to rest on A. The opposite value of each Next Step node ($\langle -3, -10 \rangle$) then occupies the appropriate Previous Step node (presumably via an “efference copy” mechanism). Location of the next target B from FP ($\langle 12, 8 \rangle$) occupies the Next Target nodes. We saw above that the Next Step nodes compute the dimensions $\langle 9, -2 \rangle$ of the step $A \rightarrow B$. Values that just occupied the Previous Step nodes ($\langle -3, -10 \rangle$) are transferred to the first level of the Working Memory Store. Opposite values of the step just executed ($\langle -9, 2 \rangle$) now occupy the Previous Step nodes. Location of the next target C from FP ($\langle 5, -4 \rangle$) occupies the Next Target nodes. Next Step nodes compute the dimensions of the next step $B \rightarrow C$:

$$\langle -3 + -9 + 5, -10 + 2 + -4 \rangle = \langle -7, -12 \rangle$$

The process repeats; $\langle -3, -10 \rangle$ and $\langle -9, 2 \rangle$ occupy the second and first layers (respectively) of the Working Memory Store and $\langle 7, 12 \rangle$ occupies the Previous Step

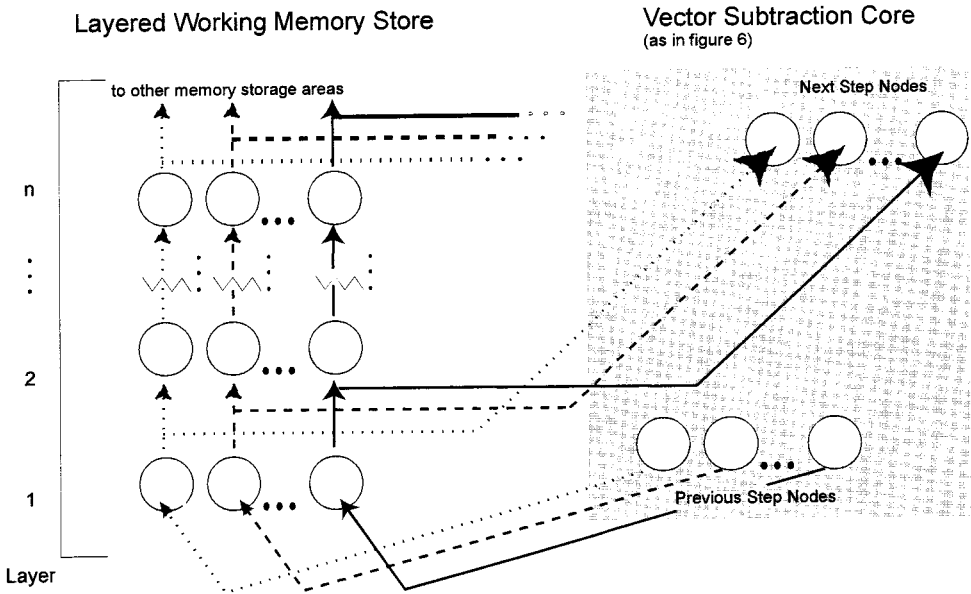


FIG. 7. Working Memory Store. (Vector Subtraction Core remains as pictured in Fig. 6. Only its components relevant to the functioning of the Working Memory Store are diagrammed here.) Layering reflects temporal storage capacity limits in working memory for the cognitive task and neural region subserving it. Dimensions of previous steps in the sequence continue to project to Next Step nodes in the Vector Subtraction Core until their time in working memory is exhausted.

nodes. Then $\langle -2, -2 \rangle$ (location of the next target D from FP) occupies the Next Target nodes. Next Step nodes compute the dimensions of $C \rightarrow D$:

$$\langle -3 + -9 + 7 + -2, -10 + 2 + 12 + -2 \rangle = \langle -7, 2 \rangle.$$

The number of possible steps in a sequence is limited only by the capacity of the working memory store (size and time limit). Presumably this will be specific to each particular type of cognitive processing system using iterated vector subtraction as its neurocomputational mechanism.

Notice that extended sequences computed in this fashion have all four features we seek to understand neurocomputationally. They occur in orderly fashion, moving sequentially from one unified representation or idea to another. The “unified representations or ideas” are points in high-dimensional vector spaces and their “orderly progression” is the path through the space connecting them. The contents of later representations and steps in the sequence (the dimensions of the path to points in the space) depend upon those of earlier representations and steps and the later “target” states. It takes a step with very different dimensions to get to D from C in the example of Fig. 9 than it takes to get to D directly from FP. Finally, multiple steps in a sequence can be computed in advance and then executed ballistically *so long as the system has the capacity to anticipate future locations*. This is the sort of anticipatory cognizing that normal adult humans with intact frontal lobes are good at. (We address this point in the next section). If our model of vector subtraction in the appropriate

high-dimensional space is the computation implemented neurally in regions subserving some higher cognitive and conscious processes, the resulting sequences of states will possess all four features we have emphasized.

However, there is a disanalogy between saccade commanding and some common types of cognitive and conscious processing that might make vector subtraction appropriate for the former but inappropriate for the latter. In saccade generation, as in motor production generally, the neural system computing movement dimensions possesses information about the goal (“targets”) of upcoming movements. The problem that such systems solve is computing and communicating appropriate dimensions to effectors that achieve the prespecified goals (“hit the targets”). The nature of this problem is what makes a neurocomputational operation like vector subtraction appropriate. But in cognition, the problem itself sometimes is *to entertain* a representational state with appropriate content. Representing the appropriate “target state” is the whole point of the process. Vector subtraction is inappropriate for these processes because it requires that the goal (future “targets”) already be represented. (In our neurocomputational model, this representation is the set of values in the Next Target Nodes.) Hence vector subtraction is unnecessary for this type of cognitive problem. Whatever computation is supplying the values to the Next Target Nodes is doing the work.²

This worry forces us to make more explicit the intended explanatory scope of our neurocomputational model. There are common cognitive and conscious processes that lack at least one of the four specific sequential features our model addresses. We do not advertise vector subtraction implemented neurally as a model for all aspects or all types of sequential cognitive and conscious processes. The key is the third condition: later steps in the sequence depend upon the contents and dimensions of earlier steps and of prespecified upcoming “target” states. We are here concerned with a resource for discovering the neural mechanisms of cognitive processes where the subsequent “goals” are already available to the system. The problem such systems confront is that multiple paths through the vector space connect up the later targets. The cognitive system must compute paths with different dimensions to connect the later targets in the correct continuous sequence depending on their contents (i.e., locations in the vector space) and those of the earlier steps taken.

For many cognitive tasks, we possess information about later “target” states and the problem *is* to compute and initiate the sequence that connects them appropriately. Many (though of course not all) problem-solving tasks have this feature, as do many involving linguistic production (“hitting the target” of the ordered verbal sequence that expresses our verbal intention) and some involving linguistic comprehension (I know the “goal” of the utterer’s utterance, but I must unpack the multicomponent sequential verbal string to understand how it “hits this target”). The cognitive aspects of motor processing often have this character and so does an important type of attention (as we see later in this section). We saw in the quotes from William James (under “Introduction”) that a common kind of episode in our conscious streams has all four characteristics we emphasize here. The conscious state of hearing the sequence, “thunder-breaking-on-silence-and-contrasting-with-it,” *sounds different* than that of

² Thanks to Rick Grush for pointing out this disanalogy to us.

“thunder-continuous-with-previous-thunder,” even when the auditory representations of the final thunder-states (taken individually) are the same (i.e., are located in the same region of auditory vector space). Our neurocomputational model provides a structural account of these experienced differences in terms of the different paths through the auditory vector space that wind up at the same final location. Vector subtraction computes the different path dimensions: the different path dimensions are the differences we experience phenomenologically in the two cases.

Clearly, there are types of cognitive and conscious processes that lack prespecified “targets”: brainstorming, other types of problem solving, target specification and goal formation, and the like. For these cognitive tasks and conscious episodes, the problem is simply to generate the appropriate representation (content). These types fall beyond the explanatory scope of our neurocomputational model presented here. Does this “limit” the importance of our model? Not at all, unless one begins by assuming (implausibly) that some single computational essence underlies all types of cognitive and conscious processes.

So far all our illustrations of vector subtraction used a two-dimensional Cartesian space. But as we saw under “Theoretical Background: Vector-Space Interpretations of Neural Network Activity,” vector-space accounts of cognitive representations involve high-dimensional spaces. This is not a problem for our model. Vector subtraction behaves in n dimensions exactly as it behaves in 2 dimensions. Where $A = \langle v_1, v_2, \dots, v_n \rangle$ and $B = \langle v_{1'}, v_{2'}, \dots, v_{n'} \rangle$, $A - B = \langle v_1 - v_{1'}, v_2 - v_{2'}, \dots, v_n - v_{n'} \rangle$.

Consider one further addition to our general model. Cognitive or conscious processes sometimes get redirected mid-sequence to their points of origin. Visual attention provides a nice example. You are monitoring a portion of visual space and get distracted by a peripheral stimulus. You turn away from your initial focus and begin exploring the distraction. Perhaps you set up a new origin for a systematic exploration. But if it is important to keep your attention focused on the initial origin, your move to the periphery prompts a “return to origin” message. (Driving an auto amid attractive billboards hopefully should be a familiar example for most readers.) This feature is a hallmark of attention across all perceptual and representational modalities. It also occurs during all the cognitive and conscious processes our model seeks to address. Sometimes these cognitive and conscious sequences are brought back abruptly to their origin when a subject suddenly realizes that something has gone awry. Sometimes the “return” begins before we are consciously aware of it.

How might we implement this aspect of cognitive and conscious processing into our neurocomputational model? Figure 8 illustrates one possibility. A variable threshold on the “Return to Origin” command is tied to activity at some layer in the Working Memory Store. Once information about previous steps in the sequence reaches the layer at which the threshold is set—that is, once a piece of information has been present in working memory for some prespecified length of time—the Return mechanism inhibits new input from the Next Target to the Next Step nodes in the Vector Subtraction Core. In effect, this sends a $\langle 0, \dots, 0 \rangle$ vector, along with the values in the Previous Step and Working Memory nodes, to the Next Step nodes. The Next Step nodes compute a vector that returns the system to its original FP.

Figure 9 illustrates the computations of this component. Suppose the variable “Re-

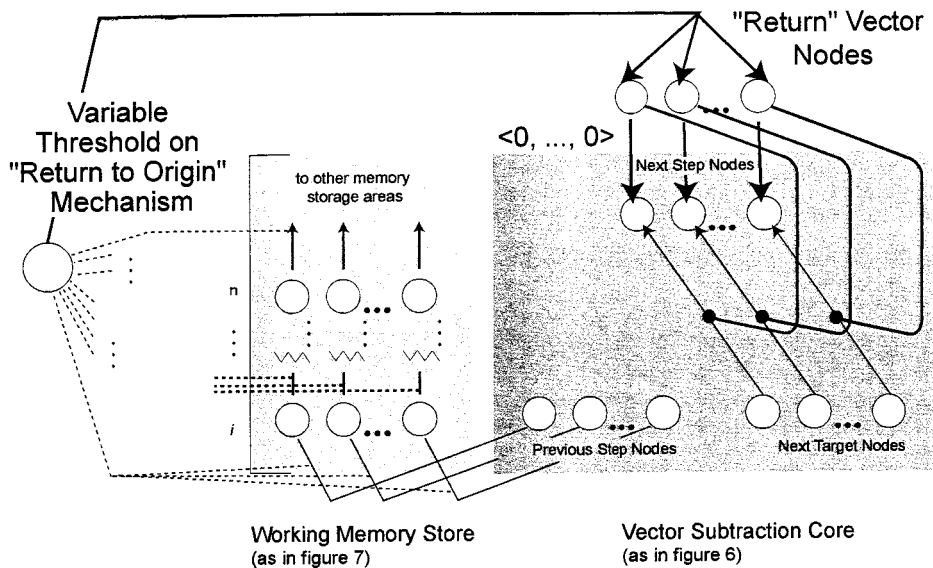


FIG. 8. Return to Origin mechanism controlled by a Variable Threshold. (Vector Subtraction Core and Working Memory Store remain as pictured in Figs. 6 and 7. Only their components relevant to the functioning of the Return to Origin mechanism are diagrammed here.) Variable Threshold is set to a particular layer of the Working Memory Store. Once activated, the resulting computation yields dimensions in the Next Step nodes that return the network to its origin (FP) in the vector space.

turn" threshold is set to the second layer of the Working Memory Store. Activity in that layer activates the Return mechanism. In the illustrated example, this activity occurs right before the network computes the step $C \rightarrow D$. The network will have executed the steps $FP \rightarrow A$, $A \rightarrow B$, and $B \rightarrow C$ as described five paragraphs above. Activating the Return component inhibits the Next Target nodes coding for the location of D. Hence the values that reach the Next Step nodes are $\langle -3, -10 \rangle$, $\langle -9, 2 \rangle$ (from Working Memory nodes), $\langle 7, 12 \rangle$ (now occupying the Previous Step nodes), and $\langle 0, 0 \rangle$ (the effect of inhibiting the Next Target nodes). Next Step nodes compute vector with dimensions

$$\langle -3 + -9 + 7 + 0, -10 + 2 + 12 + 0 \rangle = \langle -5, 4 \rangle.$$

This is exactly the vector necessary to return to FP from C. Had the variable Return mechanism been set higher up the Working Memory Store, processing at this stage would have computed the dimensions of $C \rightarrow D$, as explained above. This "return to origin" process is characteristic of a common type of *attention*. We can thus add this cognitive (and sometimes conscious) process to the list compiled earlier in this section of the explanatory scope of our neurocomputational model.

We derived features of the Vector Subtraction Core directly from the known neurobiology of saccade command activity in primate FEFs. Our model needed additional components to produce sequential processing matching the complexity of some higher cognitive and conscious processes: the structured Working Memory Store and the Return to Origin mechanism. We introduced these in this section for purely com-

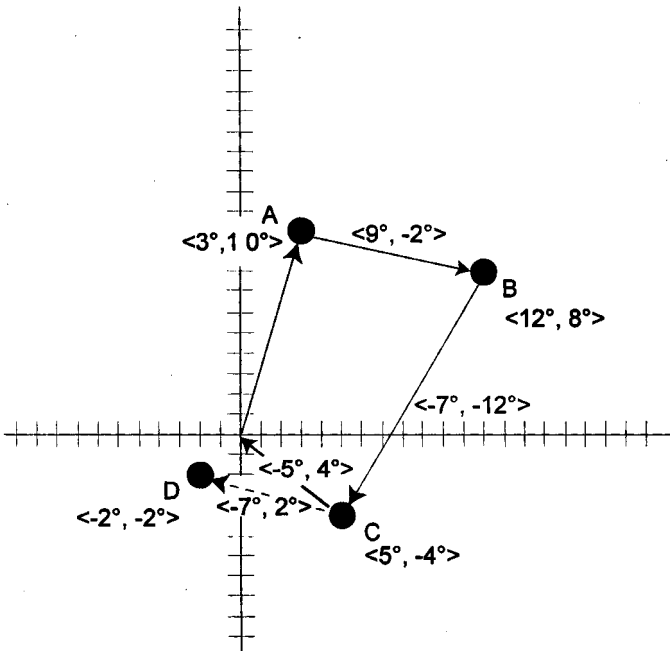


FIG. 9. Vector subtraction computing multiple steps (sequences of representational states) through a vector space in the neurocomputational model supplemented with the Working Memory Store and Return to Origin mechanism. Computations described in the text.

putational reasons. This is not sufficient for our *neurocomputational* aims, however. Is there any biological evidence suggesting that these additional components are present in brains—especially in regions known to be involved in higher cognition and consciousness? And is there any evidence that vector subtraction itself is implemented in these regions?

NEUROBIOLOGICAL PLAUSIBILITY OF THE GENERAL MODEL

We begin with the Working Memory Store. First postulated in cognitive psychology, *working memory* is the capacity to hold an item “transiently in mind” in the service of some cognitive task (Baddeley, 1986; Baars, 1997). Examples range from recalling a telephone number while dialing to imagining a series of chess moves. “Blackboard of the mind” and “global workspace” are common metaphors. Neurobiologically, primate working memory for some types of information appears to be localized to distinct regions of prefrontal cortex (Goldman-Rakic, 1996). Working memory for visuospatial information depends upon activity in the dorsolateral prefrontal convexity. Working memory for features of objects or faces depends upon activity in inferior prefrontal cortex. Working memory for semantic and other verbal processes depends upon activity in anterior prefrontal cortex.

Our understanding of the neural basis of working memory has progressed greatly with the recent discovery of “memory fields” in prefrontal neurons. Using an oculo-

motor delayed response task, Funahashi et al. (1989) recorded activity in dorsolateral prefrontal neurons during the delay. Activity profiles resembled the receptive fields of sensory neurons and the movement fields of neurons in motor pathways. A given neuron was maximally active during the delay before saccades to one "preferred" remembered location, less active before saccades to locations nearby, and remained at baseline response rate before saccades to all other locations. Funahashi et al. (1989) commonly recorded increased activity during delays for as long as 12–15 s, but never longer than 20 s. Delay-period activity to preferred memory locations expanded or contracted as the delay period was lengthened or shortened. These are features one would expect of a neural mechanism for transient working memory, as it reflects information held temporarily "on line." Information to be retained longer would depend on other, more stable mechanisms (e.g., synaptic long-term potentiation).

Goldman-Rakic (1996) and her colleagues defined these neurons' "memory fields" as the location of the targets that evoked responses during the delay period. The concept generalizes to any object, dimension, or event that prompts a systematic neuronal response during the delay period of a working memory task. (There can be neuronal working memory fields for types of objects, properties, and processes.) Subsequent neuroanatomical investigations have revealed the laminar structure of sensory, memory, and motor control neurons in prefrontal cortical columns dedicated to particular memorial items (Funahashi et al., 1990). Local circuits between pyramidal and inhibitory interneurons provide the beginnings of a theory of memory field formation (Wilson et al., 1994; Goldman-Rakic, 1995). Results from a delayed anti-saccade task confirm that activity during the delay period is linked to the location of the remembered target, not the direction of the saccade about to be initiated (Funahashi et al., 1993). (These are memory fields, not movement fields.) Even the biochemical basis of memory field formation is emerging (Williams & Goldman-Rakic, 1995).

The general concept of neuronal "working memory fields" specific to types of information is theoretically very rich. It provides a neurobiological basis for the Working Memory Store of our general neurocomputational model. There need not be (and almost certainly is not) a unitary working memory "neuronal center." The anatomical locations and physiological properties of "memory field" neurons will differ for particular cognitive tasks and types of information. Neurons with specific memory fields will be in one region for one type of cognitive task, another region for another. Activity in working memory neurons with specific memory fields, connected appropriately with the neuronal populations whose activities realize components of our Vector Subtraction Core for a particular cognitive or conscious task, carries information about previous steps through the appropriate high-dimensional vector space.

Next consider the $\langle 0, \dots, 0 \rangle$ vector message of our Return to Origin mechanism. A recent electrophysiological study of "suppression sites" in the FEFs provides a biologically plausible way that this component could be implemented. Burman and Bruce (1997) studied a class of FEF neurons to which low-grade electric stimulation did not elicit saccades in either visually guided or memory tasks. Instead, it consistently delayed and rendered inaccurate saccades made immediately after stimulation ceased. Their anatomical location and foveal receptive fields suggested to Burman and Bruce that suppression site activity codes for a saccade of 0° (amplitude and

direction). This is exactly the content of the message that our Return to Origin mechanism delivers to the Next Step nodes, replacing the input from the Next Target nodes (see again Fig. 8). This interpretation of FEF suppression sites gains plausibility from the specific effects of microstimulation on the amplitude of subsequent saccades (in both visually guided and memory tasks). Subsequent saccades following microstimulation of suppression sites were consistently hypometric (short of target), often by up to 25%. Interpreting suppression site output as coding for a 0° saccade command explains this result. The hypometric saccade results from an averaging of the accurate saccade command (coded in the presaccadic activity of the appropriate FEF neurons) and the electrically induced 0° command from the suppression sites.

This “suppression site” concept is also rich for biologically plausible neurocomputational theorizing. Generalized to higher dimensional vector spaces, suppression site activity in neural regions subserving higher cognitive functions would code for a $\langle 0, \dots, 0 \rangle$ message in the vector space. This is exactly the message that the Return to Origin component of our general neurocomputational model provides. To implement this component, suppression sites in specific systems must be connected appropriately with the neuronal populations implementing the Next Step nodes and the Working Memory Store. But these types of connectivities apparently exist in primate FEFs. Electrophysiological studies suggest that FEF suppression sites *compete with and inhibit* presaccadic neuronal activity (Goldberg et al., 1986; Dias & Bruce, 1993, 1994). Thus the microanatomical architecture of the FEFs might even provide a neurally realistic model for *how* the Return to Origin nodes (suppression sites) inhibit the Next Target nodes and send a $\langle 0, \dots, 0 \rangle$ message to the Next Step nodes.

The variable threshold on our Return to Origin mechanism is responsive to the *motivational significance* of the initial origin and the prespecified future targets. Evidence from lesion studies, neural imaging, and anatomical tract tracing suggests that in the primate saccade command system, cell properties within and connectivities between *anterior cingulate cortex* (ACC) and frontal areas (including FEFs and dorsolateral prefrontal area 46) realize this feature (Posner & Deheane, 1994; Mesulam, 1998). Anatomically, ACC is ideally located to integrate cognitive and affective information and to communicate the results throughout the brain. Known connectivities and processing between anterior cingulate cortex and FEFs might even provide a neurocomputational model of how the variable threshold is implemented in other neural regions. We have recently published a neurocomputational model of cingulo-frontal and intracingulate circuitry and have generalized it into a biologically plausible account of how motivation and affect effect higher cognitive and conscious processes (Bernstein, Stiehl, & Bickle, in press). The variable threshold in the model reported here emerges as a part of a more comprehensive Significance Activation Mechanism developed there. The computational operation of the latter is inspired directly by the known cell properties and connectivities of this primate cingulo-frontal circuit.

We saw under “Generalizing from This Implementation: A Neurocomputational Network That Computes Multi-Step Sequences” that our model needs the capacity to anticipate future targets in the appropriate vector spaces. This capacity provides for the “ballistic” feature of some cognitive and conscious processes. Neuropsychological evidence has suggested for some time that frontal cortex is crucial for “antici-

patory" planning and execution in primates (Kolb & Whishaw, 1996). For example, following localized unilateral frontal lobectomies patients are often poor at programming and producing sequences of facial movements, though they can recall and reproduce individual movements of the sequence (Kolb & Milner, 1981). Frontal-lobe patients are also poor at formulating and carrying out multistep problem-solving strategies. On the "dinner party" task, these patients are extremely inefficient, constantly break the rules, and often fail to complete a majority of the errands. Yet they are not impaired in explaining the individual errands or the rules (Shallice & Burgess, 1991). In the terminology of our model, a strategy is an anticipatory path through the appropriate high-dimensional solution space. Hyper-points or subvolumes in this space represent the contents of commands to solve subtasks, implemented in activity patterns across appropriate neuronal populations. The location in the frontal cortex of cells with properties and connectivities implementing our model's components and the kinds of deficits resulting from frontal damage together provide rich and connected evidence for the biological plausibility of its required anticipatory feature for generating "ballistic" processes.

Finally, is it plausible to hypothesize that the cell properties and connectivities that actually implement vector subtraction in the primate FEFs are present in other cortical areas involved in higher cognition and consciousness? Cytoarchitecturally, in terms of its cell types, columnar structure, and layering, primate FEFs consist of typical frontal-type cortex with distinctive granular layers (Parent, 1996). In terms of its cellular constitution and organization, FEFs resemble much of the rest of frontal cortex. They also resemble much of dorsal parietal and inferior temporal cortex, which is interesting because the latter are implicated in spatial representation/manipulation and visual object recognition, respectively (Carlson, 1994). These capacities are paradigmatically cognitive. Hence much of cortex known to be involved in higher cognition shares the cytoarchitecture of primate FEFs. The basic cellular resources from which evolution forged a vector subtraction mechanism to compute sequential saccade commands are present in these other regions. This suggests that the necessary cellular structures and connectivities are present there to implement vector subtraction.

Our neurocomputational model offers testable predictions about the cell properties and connectivities in regions subserving an important type of higher cognition and consciousness. If vector subtraction is the neurocomputation generating these processes, anatomical and physiological investigation should reveal:

1. Cells with fields responsive to the dimensions of the objects or events relevant to the particular cognitive process, analogous to receptive fields of sensory neurons and movement fields of neurons in motor pathways. Activity across the population of these cells will code for the location of the "next target" in the high-dimensional vector space constituting the representations involved in the process (perhaps via vector averaging) (Next Target nodes).

2. Cells with "postvector" fields coding for the dimensions opposite those of the last step through the vector space (perhaps via an "efference copy" mechanism) (Previous Step nodes).

3. Appropriate connectivities among the cells in the network to sum the previous step and next target values and to project this message to "effector systems" execut-

ing the task (presumably as a vector average across the population of active neurons) (Next Step nodes).

4. Cells with appropriate memory fields for the dimensions of objects or events involved in the task, appropriate activity dynamics reflecting the temporal dimensions of working memory involved in the task, and appropriate connectivities to the network's neuronal populations implementing the Next Step and Previous Step nodes (Working Memory Store).

5. Cells that code for a $\langle 0, \dots, 0 \rangle$ location in the appropriate vector space, connected in the necessary fashion to the network's neuronal populations implementing the Next Step nodes and Working Memory Store (variable threshold on the Return to Origin mechanism).

For any neural region possessing cell properties and connectivities that meet these conditions, and known to be involved in the type of higher cognitive or conscious process having the four features emphasized in this paper, we have a viable neurobiological hypothesis for *how* that region generates the characteristic sequential features of that process. Our model predicts that neuroanatomical and physiological investigations in "cognitive" and "conscious" neural regions will reveal these cell properties and connectivities. In this way, neurocomputational models like ours suggest empirical, *mechanistic* bridges across the levels that at present separate "mind" from "brain."

IMPLICATIONS FOR PHENOMENOLOGY

If our neurocomputational model describes correctly the mechanisms of some common episodes in our Jamesian conscious streams, it yields interesting consequences for phenomenology. Our discussion of these consequences here is speculative and quick, but they seem intriguing enough to motivate philosophers, psychologists, and other higher level theorists to investigate our model (not just neuroscientists and computational modelers). Our model suggests that sequences of contents through stretches of our conscious streams are much more *preprogrammed* and *semiballistic* than folk psychology assumes. By "preprogrammed" we mean that these sequences of contents—paths connecting the points or subvolumes in the appropriate high-dimensional vector spaces—are computed in advance (up to the limits of working memory). The steps are then run off with little capacity for revision via "reflective feedback." Sequences computed by our model are "adjustable" by the Return to Origin mechanism and so are not completely "ballistic." But the capacities for adjustment are limited. According to our model, we humans have little control over short stretches of the contents that parade through our conscious streams. Once computed and initiated, these sequences run off quite "automatically."

Is there independent evidence for this consequence? Some temporal considerations, often neglected, speak in its favor. As Dennett (1991) stresses, our brains are under strict time pressures. On the input side, "there are perceptual analysis tasks such as speech perception which would be beyond the physical limits of the brain's machinery if it didn't utilize *ingenious anticipatory strategies*" (p. 144; our emphasis). Speech perception is a paradigm higher cognitive task. We can consciously attend to its outcome and some stages of the process. Vector subtraction implemented

neurally would constitute one of the brain's "ingenious anticipatory strategies." Similar time constraints also require anticipatory, preprogrammed, semiballistic processes on the output side. Dennett (1991) writes: "Many acts must occur so fast and with such accurate triggering that the brain *has no time to adjust its control signals in the light of feedback*" (p. 145; our emphasis). The only "control signals" available to the brain must be themselves "preprogrammed" and "semiballistic," like activity in our Return to Origin mechanism. Nature's time constraints force brains to be much more "automatic" and less "reflective" in even their cognitive and conscious processing than folk psychology suggests.

Yet we still have that persistent phenomenological self-image: our conscious, cognitive selves construed in the image of Rodin's "The Thinker." We can call this the "Elaborate Practical Reasoning" image. Sequential contents of our conscious streams result from "all things considered" idea-generation and systematic reflection on their relevance to current cognizing and behavior. We do lots of revising and rethinking about both ends and means, followed by a decision to entertain consciously some sequence of ideas. Doesn't that description fit better with our first-person experiences of conscious cognizing? So aren't the "preprogrammed" and "semiballistic" aspects of our model its crucial flaws, if it seeks to provide a neurocomputational account of consciousness?

We offer three replies to this worry. First, when we consider the phenomenology of, for example, intentional action, we tend to think of the "reflective" cases. But we thereby ignore the myriad "automatic" examples that occupy our consciousness the vast majority of time. Consider my sequential intentional action 5 s ago to type "vast majority." That was a common type of sequential intentional action, and it is one I was conscious of, but it did not result from elaborate practical reasoning. Neither do the *vast majority* of the sequential intentions that occupy our consciousness. Most appear, proceed as steps in short sequences, and are replaced quickly by more of the same. Our experiences of most mid-sequence interruptions of conscious streams are similar. Dennett (1991) aptly describes this neglected phenomenological fact: "Although we are occasionally conscious of performing elaborate practical reasoning . . . these are *relatively rare experiences*. Most of our intentional actions are performed without any such preamble, and a good thing, too, since there wouldn't be time" (p. 252; our emphasis). Reflect back on your stream-of-conscious experiences as you comprehended the last sentence in the Dennett quote. We doubt that much "elaborate practical reasoning" occupied that stretch of your Jamesian stream. Instead, phenomenology reveals a short sequence of conscious ideas that resulted plausibly from "preprogrammed, semiballistic" neural mechanisms. Careful phenomenology uncorrupted by folk-psychological assumptions supports, rather than challenges, our neurocomputational account of short stretches in our conscious streams.

Second, careful phenomenology also reveals the fragmentary and disconnected feature of longer stretches through our Jamesian streams. Psychologist Bernard Baars (1997) has remarked that "most of the time we humans do not engage in [extended] logical or even structured thinking. . . . We can do it, but it is something of a feat" (p. 96). Evidence from social and clinical psychology, based ultimately on carefully

guided introspective reports, shows instead that “we humans devote most of our conscious streams to fantasies, dreams, disconnected thoughts, and debatable beliefs. The stream of consciousness, as William James wrote famously, seems a messy, arbitrary sort of thing, full of stops and starts, hopping and skipping from one half-articulated thought to another” (p. 96). Our limited ability to engage in long stretches of connected conscious thinking is a “recent cultural product” (p. 96). So even the order in which short stretches occur in our conscious streams typically does not result from “elaborate practical reasoning.” Careful phenomenology is once again consistent with implications of our neurocomputational model: a number of networks computing and running off short sequences of representations in semiballistic fashion, constantly in competition with each other for temporary entry into our Jamesian streams.

Third, a remarkable result using positron emission tomography (PET) suggests an explanation for those rare stretches of conscious “elaborate practical reasoning” that is consistent with our model. Paulescu et al. (1993) investigated the neural regions that become active when humans engage in “quiet inner speech”: when they generate an inner monologue without talking aloud (see also Baddeley, 1993, Fig. 2, and Baars, 1997, Fig. 4.) Activity was most pronounced in Broca’s area (encompassing the second and third convolutions in frontal cortex), Wernicke’s area (encompassing the first and second convolutions in temporal cortex), and the left supra-marginal gyrus. These are the classic *speech production and comprehension* “centers”! When we engage in “quiet inner speech,” from the brain’s point of view we are literally *producing and comprehending linguistic utterances*. We are literally *talking to ourselves*.

This result holds fascinating implications for consciousness science. Those relatively rare instances of conscious “elaborate practical reasoning” are instances of “quiet inner speech,” where we “talk to ourselves,” where we “carry on an inner dialogue.” In keeping with Paulescu et al.’s (1993) PET results, in these instances our language production and comprehension areas will be maximally active. But these language regions lack access to most cognitive processing taking place elsewhere in the brain. Speech production and comprehension are subject to tremendous time pressures, and so can only be privy to a limited amount of ongoing neural activity elsewhere. The expressive capacity of linguistic utterances (in bits of information per second) is orders of magnitude less than the processing capacities of parallel neural networks. All of the information processing taking place in a parallel network cannot be crammed into a real-time comprehensible or producible set of sentences. Lots of information gets lost in the input–output transition of language production. Finally, the anatomical pathways between language regions and other cortical areas are limited. Our language areas typically enjoy only limited access to the outputs of information-processing networks in the rest of the cognizing brain. Based on the limited information they receive, in quiet inner speech our language areas construct (and comprehend) *a linguistic story* about what is going on elsewhere in the brain. These constructions are the “elaborate practical reasonings” that occasionally occupy short stretches of our Jamesian conscious streams. Folk psychology mistakenly elevates these stretches to the status of norm for conscious processing and mistakenly assumes that their contents reflect the actual mechanisms of cognition. However,

these internal stories probably are outright fabrications, confabulations that occupy briefly our conscious streams because of the activity levels in the language centers. In these rare and brief bouts of conscious “elaborate practical reasoning, we are literally *lying* to ourselves.

Recall that our neurocomputational model seeks to provide plausible (and testable) hypotheses about the neural mechanisms of linguistic production and comprehension. Linguistic sequences possess all four features addressed by our general model. Temporal constraints require that language processing be both “preprogrammed” and “semiballistic” much of the time. Think of a heated verbal exchange: how quickly you comprehend a semantically complex utterance, then formulate and initiate a reply. Think also about how quickly and “automatically” your “error sensor” (Return to Origin mechanism?!) engages when a sequence is going amiss: often before you are consciously aware of the error. Vector subtraction implemented neurally in the language-processing regions would produce sequential outputs with exactly these features. Even the anatomical locations and connectivities of the language regions are suggestive. Broca’s area lies in the frontal cortex just caudal and inferior to the FEFs (Brodmann’s areas 8 and 6) and to known cortical sites of working memory (areas 45 and 46). Wernicke’s area lies caudal to this, in temporal cortex neighboring frontal cortex. As we saw in “Neurobiological Plausibility of the General Model,” these areas are similar to each other and to primate FEFs at all levels of anatomical and cytoarchitectural analysis: similar cell types and distribution, columnar arrangement, and laminar structure. If the appropriate cell properties and connectivities exist in these areas, vector subtraction is a promising hypothesis about the neural mechanisms of language production and comprehension: both for utterances bound for public exhibition and those retained for “quiet inner speech.” Some of the latter become the relatively rare and short-lived stretches of “elaborate practical reasoning” in our Jamesian conscious streams. Through its potential account of the neural mechanisms underlying language processing, even the phenomenology of these experiences seems explicable by vector subtraction implemented neurally.

ACKNOWLEDGMENTS

Rick Grush and Rodney Cotterill refereed an earlier draft. Each offered numerous suggestions that led to many improvements. Rodney graciously sent us offprints of related work of his. We will explore some comparisons and contrasts in a future paper. This research was supported initially by a College Research Award in Spring, 1997 to J.B. from the Dean’s Office, College of Arts and Sciences, East Carolina University. We dedicate this article to the memory of our friend, Gary Peterson, late of the Department of Anatomy and Cell Biology at the East Carolina University School of Medicine. Gary, our neuroanatomy consultant, was killed tragically by a drunk driver in November, 1998 while bicycling in the hills north of Durham, NC.

REFERENCES

- Baars, B. (1997). *In the theater of consciousness*. Oxford: Oxford Univ. Press.
- Baddeley, A. (1986). *Working memory*. Oxford: Oxford Univ. Press.
- Baddeley, A. (1993). Verbal and visual subsystems of working memory. *Current Biology*, **3**, 563–565.
- Barbas, H., & Mesulam, M. (1981). Organization of afferent input to subdivisions of area 8 in the rhesus monkey. *Journal of Comparative Neurology*, **200**, 407–431.

- Bernstein, M., Stiehl, S., & Bickle, J. (in press). The effect of motivation on the stream of consciousness: Generalizing to a neurocomputational model from cingulo-parieto-frontal circuits controlling saccadic eye movements. In R. Ellis & N. Newton (Eds.), *The cauldron of consciousness: Motivation, affect, and self-organization*. New York: John Benjamins.
- Bruce, C. J., & Goldberg, M. E. (1985). Primate frontal eye fields. I. Single neurons discharging before saccades. *Journal of Neurophysiology*, **53/3**, 603–635.
- Bruce, C. J., Goldberg, M. E., Stanton, G. B., & Bushnell, M. C. (1985). Primate frontal eye fields. II. Physiological and anatomical correlates of electrically-evoked eye movements. *Journal of Neurophysiology*, **54**, 714–734.
- Burman, D., & Bruce, C. J. (1997). Suppression of task-related saccades by electrical stimulation in the primate's frontal eye fields. *Journal of Neurophysiology*, **77**, 2252–2267.
- Carlson, N. (1994). *Physiology of behavior* (5th ed.). Boston: Allyn and Bacon.
- Churchland, P. M. (1989). *A neurocomputational perspective*. Cambridge, MA: MIT Press.
- Churchland, P. S., & Sejnowski, T. (1992). *The computational brain*. Cambridge, MA: MIT Press.
- Dennett, D. C. (1991). *Consciousness explained*. Boston: Little, Brown.
- Dias, E., & Bruce, C. J. (1993). Frontal eye field activity in conjunction with short latency (“express”) saccades made in a gap paradigm. *Society for Neuroscience Abstracts*, **19**, 426.
- Dias, E., & Bruce, C. J. (1994). A physiological correlate of fixation disengagement in the primate's frontal eye field. *Journal of Neurophysiology*, **73**, 2532–2537.
- Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, **61**, 331–349.
- Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. (1990). Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms. *Journal of Neurophysiology*, **65**, 814–831.
- Funahashi, S., Chafee, M., & Goldman-Rakic, P. (1993). Prefrontal neuronal activity in rhesus monkeys performing an anti-saccade task. *Nature*, **365**, 753–756.
- Goldberg, M. E., & Bruce, C. J. (1990). Primate frontal eye fields. III. Maintenance of a spatially accurate saccade signal. *Journal of Neurophysiology*, **64/2**, 489–508.
- Goldberg, M. E., Eggers, H., & Gouras, P. (1992). The ocular motor system. In E. Kandel, J. Schwartz, & T. Jessell (Eds.), *Principles of neural science* (3rd ed.). New York: Appleton and Lange.
- Goldberg, M. E., Bushnell, M., & Bruce, C. J. (1986). The effect of attentive fixation on eye movements evoked by electrical stimulation of the frontal eye fields. *Experimental Brain Research*, **61**, 579–584.
- Goldman-Rakic, P. (1995). Cellular basis of working memory. *Neuron*, **14**, 477–485.
- Goldman-Rakic, P. (1996). Regional and cellular fractionation of working memory. *Proceedings of the National Academy of Sciences*, **93**, 13473–13480.
- James, W. (1890). *The principles of psychology*. New York: Henry Holt.
- Kolb, B., & Milner, B. (1981). Performance of complex arm and facial movements after focal brain lesions. *Neuropsychologia*, **19**, 514–515.
- Kolb, B., & Whishaw, I. (1996). *Fundamentals of human neuropsychology*. New York: W. H. Freeman.
- Lee, C. W., Rohrer, R., & Sparks, D. L. (1988). Population coding of saccadic eye movements by neurons in the superior colliculus. *Nature*, **332**, 357–360.
- Llinás, R. (1975). The cortex of the cerebellum. *Scientific American*, **232/1**, 56–71.
- Mays, L., & Sparks, D. (1980). Dissociation of visual and saccade-related responses in superior colliculus. *Journal of Neurophysiology*, **43**, 207–232.
- Mesulam, M. (1998). From sensation to cognition. *Brain*, **121**, 1013–1052.
- Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.
- Parent, A. (1996). *Carpenter's human neuroanatomy* (9th ed.). Baltimore: Williams and Wilkins.
- Paulescu, E., Frith, D., & Frackowiak, R. (1993). The neural correlates of the verbal component of working memory. *Nature*, **362**, 342–345.

- Posner, M., & Deheane, S. (1994). Attentional networks. *Trends in Neuroscience*, **17**, 75–79.
- Shallice, T., & Burgess, P. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain*, **114**, 727–741.
- Williams, G., & Goldman-Rakic, P. (1995). Modulation of memory fields by dopamine D1 receptors in prefrontal cortex. *Nature*, **376**, 572–575.
- Wilson, F. A. W., O'Scalaidhe, S., & Goldman-Rakic, P. (1994). Functional synergism between putative γ -aminobutyrate-containing neurons and pyramidal neurons in prefrontal cortex. *Proceedings of the National Academy of Sciences*, **91**, 4009–4013.

Received June 24, 1999