

Penultimate version for *Oxford Handbook of Philosophy and Neuroscience*, John Bickle (ed).

Neurocomputational models: Theory, application, philosophical consequences

Chris Eliasmith, University of Waterloo

Abstract: Neural coding, neural computation, and control (or dynamical systems) theory have recently been integrated to support descriptions of neurobiological systems at many different levels of detail. I discuss the Neural Engineering Framework (NEF), which realizes this integration, and highlight some of the philosophical consequences of its development. To do so, I present the three principles of the NEF, and then describe an example of their application to the development of a neural model of part of the rat navigation system. Relying partly on the insights provided by this model, I describe the implications of the NEF for traditional philosophical problems including: 1) mental representation and semantics; 2) the unity of science (i.e., theory reduction); and 3) appropriate theory construction in the behavioural sciences.

Keywords: theoretical neuroscience, computational neuroscience, neural coding, neural modeling, neural computation, representation, neuroscience, semantics, rodent navigation, path integration, unity of science

1. Introduction

Theoretical (or computational) neuroscience has come to play a role in neuroscience akin to that played by theoretical physics in the physical sciences. However, unlike theoretical physics, theoretical neuroscience is not characterized by a few well-studied basic theories (e.g., string theory, loop quantum gravity, etc.). Instead, as can be seen by perusing the textbooks in the field, theoretical neuroscience is largely a collection of disparate methods, models, and mathematical techniques that relate to neurobiological systems (see e.g., Dayan & Abbott, 2001; Koch, 1998). For any given neural system of interest, some subset of these methods is chosen (or new ones developed) and they are applied in the analysis and/or simulation of that system. In short, while theoretical neuroscience has helped provide a quantified understanding of neural systems, it has done so in a largely unsystematic manner.

As a result of this diversity of techniques, and the accompanying variety of assumptions, it is difficult to discern what philosophically interesting conclusions can be drawn about neural systems *in general*. As a result, rather than focusing on the entire range of techniques used by theoretical neuroscientists, in this article I describe a systematic approach to studying neural systems which has collected and extended a set of consistent methods that are highly general. These methods have come to be called the Neural Engineering Framework (NEF), and can be summarized by three basic principles, which I describe next. While these principles have been extended in more recent work (Tripp & Eliasmith, in press), here I present their original formulation (Eliasmith & Anderson, 2003) which is simpler and does not detract from subsequent discussion. An indication of the generality of these three principles is the wide variety of neural systems they have been used to characterize. These include the barn owl

auditory system (Fischer, 2005), the rodent navigation system (Conklin & Eliasmith, 2005), escape and swimming control in zebrafish (Kuo & Eliasmith, 2005), the translational vestibular ocular reflex in monkeys (Eliasmith, Westover, & Anderson, 2002), working memory systems (Singh & Eliasmith, 2006), and language-based deductive inference (Eliasmith, 2004). These models span sensory, motor and cognitive systems across the phylogenetic tree. This broad range of applicability, which is a consequence of the generality of the NEF, makes subsequent philosophical consequences of greater interest (see section 4).

2. The Neural Engineering Framework (NEF)

2.1 Introduction

The NEF draws heavily on past work in theoretical neuroscience, integrating work on neural coding, population representation, and neural dynamics to enable the construction of large-scale biologically plausible neural simulations. The three principles that form the basis of the framework are:

1. *Representation*: Neural representations are defined by a combination of non-linear encoding and optimal linear decoding.
2. *Transformation*: Transformations of neural representations are functions of the variables that are represented by a population.
3. *Dynamics*: Neural dynamics are characterized by considering neural representations as control theoretic state variables.

These principles are quantitatively defined by Eliasmith and Anderson (2003) so as to: a) apply to a wide variety of single cell dynamics; b) incorporate linear and nonlinear transformations; c) permit linear, nonlinear and time-varying dynamics; and d) support the representation of scalars, vectors, functions, or any combinations of these. In addition, the principles are formulated so as to preserve our current understanding of the biophysical limitations of neural systems (e.g., the presence of significant noise, the intrinsic dynamics of neurons, largely linear somatic interactions of dendritic currents, etc.). In the next four subsections I describe each principle in more detail, and discuss how they provide a unified view of the function of neural systems for the behavioural sciences.

2.2 Representation

Introduction

The notion of ‘representation’ is broadly employed in neuroscience.ⁱ In general, if a neuron “fires” relatively rapidly when an animal is presented with a certain set of stimuli, the neuron is said to “represent” the property that the set of stimuli share (see e.g., Felleman & Van Essen, 1991). This kind of experiment has been performed on mammals since Hubel and Wiesel’s (1962) classic research in which they identified cortical cells selective to the orientation and size of a bar in a cat’s visual field. The slightly earlier “bug detector” experiments of Lettvin et al. (1959/1988), perhaps better known to philosophers, take a similar approach. In the bug detector experiments, retinal ganglion cells were found that respond to small, black, fly-sized dots in a frog’s visual field. These were referred to as “bug detectors” because they fired rapidly when such dots were present and fired less rapidly when they were not. More recently, this method has been used to find “face-selective cells” (i.e., cells that respond strongly to faces in

particular orientations) in monkey visual cortex (Desimone, 1991). In all of these cases, what is deemed important for representation is how actively a neuron responds to some known stimuli.

The two central difficulties with this use of the term ‘representation’ are that it assumes 1) *single* neurons are the basic carriers of content, and 2) content can be determined by what has been called the “naïve causal theory” – the view that a brain state represents whatever causes it to be active – which is well-known to be highly problematic (Dretske, 1988). Little thought has been given in neuroscience to trying to establish a principled means of determining what appropriate representational vehicles are, or how they might be related to the stimuli they are taken to represent. Why, for instance, should we assume that cells that selectively fire in the presence of faces actually represent faces? If the system is unable to use the information carried by such a cell to detect faces, or if the neuron is only partly informative of the presence of a face, or if as yet untested non-face stimuli can cause the cell to be active, such content claims will be misleading.

Work in theoretical neuroscience has been more careful regarding such claims. In particular, researchers examining neural coding are often careful not to assume that the stimuli presented to an animal is automatically, or fully represented, despite observed correlations (Rieke, et al., 1997). As a result, one of the most significant conceptual contributions of theoretical neuroscience to a neuroscientific understanding of representation is an emphasis on *decoding*. As mentioned, characterizing the responses of neurons to stimuli in the environment has been the mainstay of neuroscience. This, however, describes only an *encoding* process. That is, the process of responding to some physical environmental variable through the generation of neural action potentials, or “spike trains.” By adopting an information theoretic view of representation, theoretical neuroscience holds that to truly understand what information is

preserved through the encoding process, we must be able to demonstrate that we (or the system) can at least in principle decode the spike train to give the originally encoded signal. As a result, to fully define representations, we must characterize both encoding and decoding.

In addition, theoretical neuroscientists have distinguished two aspects of representation; temporal representation, and population representation. The former deals with how neurons represent time-varying signals. The latter deals with issues of distributed representation. That is, how a single cell's response contributes to a complex representation over a large group of neurons (i.e., allowing content claims to encompass more than single cells). In the next two subsections, I describe the theoretical characterization of encoding and decoding over time and neural populations employed by the NEF.

Temporal representation

Perhaps the best understood aspect of how neural systems represent time-varying signals is the encoding process. In some ways, this should not be too surprising since the focus of neuroscience in general has been on encoding. This is likely because the encoding process can be largely characterized by focusing on single cells. So, the highly successful work on quantifying the dynamics of action potential generation in single cells – including mathematical descriptions of voltage sensitive ion channels of various kinds (Hodgkin & Huxley, 1952), the use of cable equations to describe dendritic and axonal morphology (Rall, 1957; Rall, 1962), and the introduction of canonical models of a large class of neurons (Hoppensteadt & Izhikevich, 2003) – supports a highly mechanistic understanding of encoding. However, fully describing the encoding process also necessitates the identification of the particular, perhaps external, parameters that a neuron may be sensitive to (partially in virtue of its relation to other neurons in

the brain). These more holistic considerations are implicitly captured by the ubiquitously reported neuron “tuning curves.” Improving our understanding of the encoding process is largely an empirical undertaking, one which has a long, successful history in both experimental and theoretical neuroscience. However, this is not true of temporal decoding.

There are two main kinds of theory of temporal decoding in neuroscience. These are referred to as the “rate code” view and the “timing code” view. Generally speaking, rate code theories are those that assume that information about temporal changes in the stimulus is carried by the average rate of firing of the neuron responding to that stimulus. In contrast, timing code theories assume that information about the stimulus is carried by the approximate distance between neighbouring spikes in the spike train generated by the neuron responding to the stimulus.

Despite a wide ranging debate over which code is used by neural systems, when carefully considering rate codes and timing codes, it becomes evident that they are variations on the same theme. Both codes assume that we choose some time window and count how many spikes fall in that window. In the case of rate codes the time window is usually about 100 milliseconds, and in the case of timing codes the size of the window varies depending on the distance between spikes. It should not be too surprising, then, that methods have been developed for understanding temporal decoding that vary smoothly between rate codes and timing codes (Rieke et al., 1997). So, in the end, the distinction between rate codes and timing codes is not a significant one for understanding temporal representation.

These methods are surprisingly simple because they are linear (i.e., rely only on weighted sums). Suppose that we are trying to understand the representational role of two neurons. To do so, we present this ‘population’ with a signal and then record the spikes that it produces in

response to the signal. These spikes are the result of some (well-characterized) highly nonlinear encoding process. As discussed earlier, to properly characterize the representational properties of this neuron, we should be able to use those spikes to reconstruct the original signal. However, to do this we need to identify a decoder that takes those spikes as input and produces an estimate of the original signal. We can begin by assuming that the particular position of a given spike in the spike train does not change the meaning of that spike; i.e., the decoder should be the same for all spikes. Essentially, every time a spike occurs, we place a copy of the decoder at the occurrence time of the spike. We can then sum all of the decoders to get our estimate of the original input signal (see figure 1). There are well-tested techniques for finding optimal decoders of this sort. As a testament to the effectiveness of these assumptions, it has been shown that this kind of decoding captures nearly all of the information that *could* be available in the spike trains of real neurons (Rieke et al. 1997, pp. 170-176).

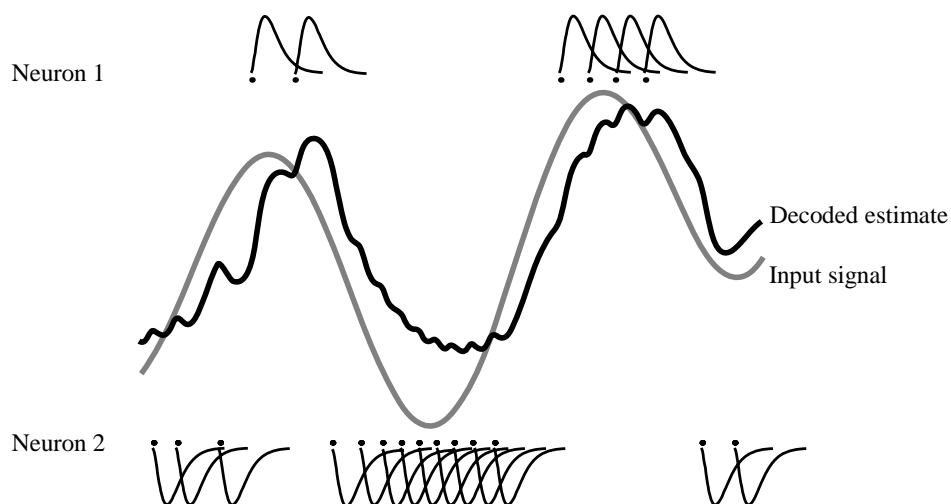


Figure 1: Temporal decoding. This diagram depicts linear decoding of a neural spike train (dots) using stereotypical decoders (skewed bell-shaped curves) on an input signal (grey line). The result of the decoding from two neurons (black line) is a reasonable estimate of the input signal.

This estimate can be indefinitely improved with more neurons. (Adapted from Eliasmith and Anderson, 2003.)

A limitation of this understanding of temporal representation is that it is not clear how *our* ability to decode the information in a spike train relates to how that spike train is actually *used* by the organism. The NEF addresses this issue by identifying the postsynaptic currents (PSCs) observable in the dendrites of receiving neurons with these temporal decoders (Eliasmith and Anderson 2003, ch. 4). While these decoders are no longer optimal, they are biologically plausible (unlike the non-causal optimal decoders found with past methods). And, increasing the number of neurons in the representation can make up for any loss in fidelity of the represented signal. An example of this kind of decoding for two neurons is shown in figure 2a. That figure demonstrates how a rapidly fluctuating signal can be decoded from a neural spike train by a receiving neuron using a timing code (it is a timing code because significant jitter in the position of the spikes would greatly change the estimate). An example of decoding a much slower signal, which is encoded using something more like a rate code, is shown in figure 2b.

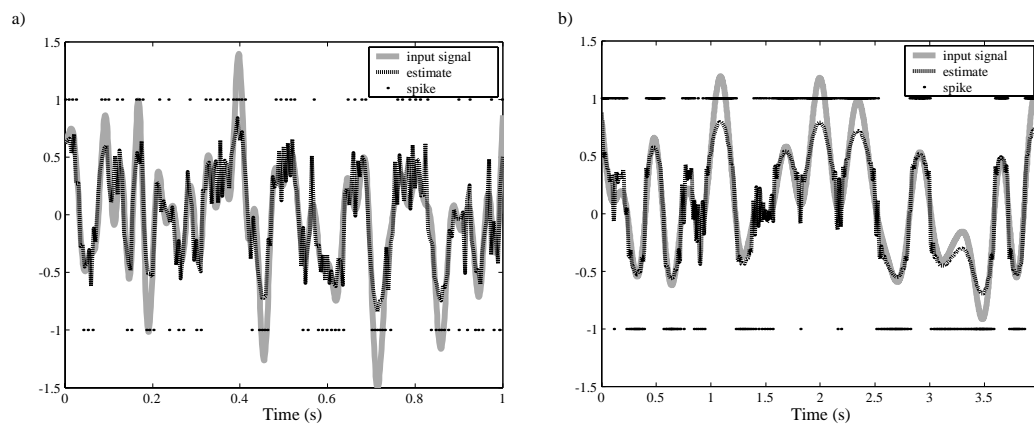


Figure 2: Biologically plausible timing and rate coding. a) A high frequency signal effectively decoded using postsynaptic currents (PSCs) as the decoders. This demonstrates a timing code. b) A low frequency signal (note the difference in time scale) similarly decoded. This demonstrates a typical rate code. (Adapted from Eliasmith and Anderson, 2003.)

Together, these diagrams show that linear temporal decoding with PSCs is biologically plausible and can capture both rate and timing codes, depending on the demands of the situation. As a result, the NEF incorporates a method of characterizing temporal coding in neurons that is unique in its combination of biological plausibility and applicability to understanding the representation of signals at a variety of time scales.

The examples discussed to this point are only time-varying scalar values. To support representations of sufficient complexity to handle the vast variety of behaviours exhibited by neurobiological systems, it is essential to understand how large groups of neurons can cooperate to effectively encode complex, real-world objects.

Population representation

In a well-known series of experiments, Apostolos Georgopoulos and colleagues explored the idea that the representation of physical variables in the cortex could be understood as a weighted sum of the individual neuron responses (Georgopoulos et al., 1986; Georgopoulos et al., 1989). By recording from a population of neurons in motor cortex, they demonstrated that a good prediction of a monkey's arm movement could be made by multiplying neuron firing rates by their preferred direction of movement and summing the result over the population.

Essentially, Georgopoulos discovered a decoding method for extracting information carried by the neural firing rates that captured how this information was used by the motor system.

It is generally agreed that Georgopoulos provided a demonstration of how to decode a scalar variable (arm angle) encoded by a population of neurons. This kind of decoding, we should notice, is identical to that described in the temporal case. It is a simple linear decoding where the temporal decoder is replaced by a population one (i.e., preferred direction). However, the particular decoding chosen by Georgopoulos is far from optimal. Nevertheless, it is a simple matter to determine the optimal linear decoder (Salinas & Abbott, 1994). Furthermore, it is easy to generalize this kind of understanding of neural representation to more complex mathematical objects than scalars (Eliasmith & Anderson, 2003). For instance, instead of understanding neurons in motor cortex as encoding a one-dimensional scalar (i.e., direction), we can take them to be encoding a two-dimensional vector (i.e., direction and distance of arm movements). Indeed, there is evidence that the neurons Georgopoulos originally recorded from carry information about both of these dimensions (Moran & Schwartz, 1999; Schwartz, 1994; Todorov, 2000). However, the discussion so far does not make it evident how to capture the wide variety of neural responses observed by experimental neuroscientists. For instance, one of the most common kinds of tuning curves observed in cortex is a Gaussian-shaped ‘bump’ (sometimes called ‘cosine tuning’) around some preferred stimulus. For example, in lateral intraparietal cortex (LIP), neurons have these bump-like responses centered around positions of objects in the visual field (Andersen et al., 1985; Platt & Glimcher, 1998). On first glance, it may be natural to see them as encoding a scalar value which indicates the current estimate of the position of an object in the visual field. However, there is evidence to suggest that these representations are more sophisticated. For instance, the representation in this area can encode multiple object

positions simultaneously, and can have differing heights of bumps at those positions (Platt & Glimcher, 1997; Sereno & Maunsell, 1998). As a result, a more natural characterization of the representation in this area is as a *function*. That is, the activity of the neurons encodes a function whose height at a location is determined by the presence of various features at that location, such as brightness, shape, etc.

Conveniently, function representation can be understood analogously to scalar and vector representation. Rather than a preferred direction vector in some parameter space, we can take neurons to have preferred functions. This would (approximately) be the function that best matched the neuron's tuning curve over the parameter space (e.g., object position and shape). It is then possible to find the optimal linear functional decoder for estimating some set of functions that the neural population can represent (Eliasmith and Anderson, 2003). With the introduction of function representation, the NEF shows how essentially any definable mathematical object can be represented over a population of neurons.

To this point, I have described both population representation and temporal representation independently. However, since both descriptions are cases of nonlinear encoding and linear decoding, it is a simple matter to combine these two kinds of representation. That is, rather than having a separate temporal decoder and a separate population decoder, we can define a single population-temporal decoder which can be used to decode a spiking, population-wide encoding of some mathematical object that captures the properties to be represented. This, then, completes a computational neuroscientific description of the representational vehicles employed by neurobiological systems.

In sum, this discussion demonstrates the wide variety of kinds of mathematical objects that can be represented in a neurobiologically plausible way with the NEF. This degree of

generality suggests that the representational assumptions embodied in Principle 1 (section 2) are very broadly applicable to neurobiological systems. And, because this characterization of neural representation is simply variations on a common theme (i.e., nonlinear encoding and linear decoding), the NEF serves to unify our understanding of representation in neurobiological systems.

2.3 Computation

Conveniently, the NEF characterizes neural computation in the same way as neural representation. That is, in terms of a nonlinear encoding and a linear decoding. The difference is that representation consists of estimating the identity function, where as computation, more broadly, consists of estimating arbitrary linear or nonlinear functions of the encoded variable. That is, when representing a variable, the system is concerned with decoding the value of that variable as it is encoded into neural spikes trains. The NEF labels decoders for this purpose “representational decoders.” However, it is possible to identify decoders for computing any function of the encoded input; the NEF labels such decoders “transformational decoders.” For example, if we define the representation in LIP to be a representation of the position of an object, we can find representational decoders that estimate the actual position given the neural firing rates. However, we can also use exactly the same encoded information to estimate where the object would be if it was translated 5° to the right. For this we could identify a transformational decoder. This particular example is merely linear transformation of the encoded information, and so is not especially interesting. However, exactly the same methods can be used to find the transformational decoders for estimating nonlinear computations as well (e.g., perhaps the system needs to compute the square of the position of the object).

This account of computation is successful largely because of the nonlinearities in the neural encoding of the available information. When decoding, we can either attempt to eliminate these nonlinearities by appropriately weighting the responses of the population (as with representational decoding), or we can emphasize the nonlinearities necessary to compute the function we need (as with transformational decoding). In either case we can get a good estimate of the appropriate function, and we can improve that estimate by including more neurons in the population encoding the information.

2.4 Dynamics

Given the previous characterizations of representation and computation, it is possible to build neurally realistic circuits that take time-varying signals as input and compute arbitrary functions of those signals. However, these techniques, as they stand, apply only to feedforward computations. As is well-known, recurrence, or backward projections, are ubiquitous in neural systems. This kind of complex interconnectivity suggests that feedforward computation is not sufficient for understanding neurobiological function. As a result, theoretical neuroscientists need a means of characterizing the sophisticated, possibly recurrent, internal dynamics of the representations they take to be present in neural populations.

The third principle of the NEF incorporates the suggestion that neural dynamics can best be understood by taking neural representations to be control theoretic state variables. Control theory is a set of mathematical techniques developed in the 1960s to analyze and synthesize complex, analog, physical systems (Kalman, 1960a; Kalman, 1960b). For linear, time-invariant (LTI) systems, control theory provides a canonical way of expressing, optimizing, and analyzing the set of possible behaviours of the system. More complex dynamics, such as nonlinear and

time-varying dynamics, can also be expressed using control theory, although analysis of the systems is no longer guaranteed to be analytically tractable.

The standard state-space form for control theoretic descriptions of physical systems is a set of differential equations defined over variables called the “state variables” (figure 3a). For any system so described, the current value of the state variables and the set of differential equations governing their dynamics completely determines the future behaviour of the system. In neural systems, the set of differential equations can be taken to describe how the representation in a neural population changes over time. The value of the variables at any particular time is determined by the (spiking) neural representation at that time, and the governing equations are determined by the connection weights between that population and any others providing input to it (possibly including that population itself).

Notably, the standard control system depicted in figure 3a assumes that the dynamics of the physical system being described can be characterized as integration (hence the transfer function being an integral). However, neurons have their dynamics determined by intrinsic properties (e.g., ion channel speed, membrane capacitance, etc.), and do not naturally support integration. As a result, it is necessary to translate the standard control theoretic equations into a form appropriate for neural systems (figure 3b). Fortunately, this translation can be done in the general case (Eliasmith and Anderson 2003, ch. 8). Such a translation allows any standard control theoretic description of a system to be written in an equivalent ‘neural’ control theoretic form. The ability to affect such a translation can prove a great benefit to theorists. In particular, it allows mobilizing the vast theoretical resources of control theory when hypothesizing about some observed biological function: a function that may already be well-characterized by control theory.

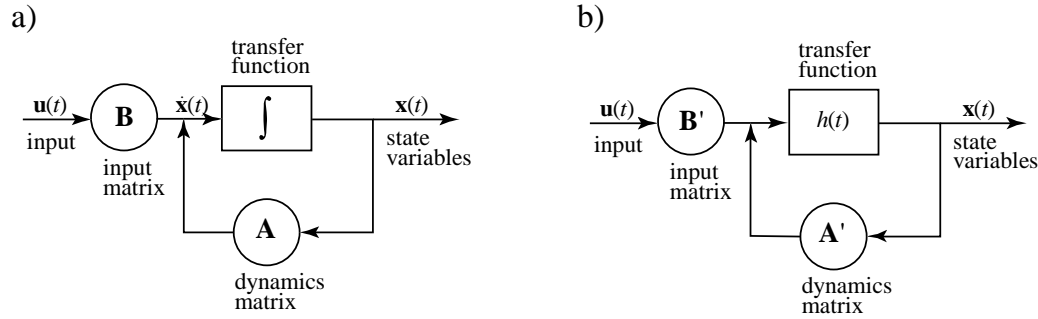


Figure 3: A diagram of the dynamics equation for LTI control theoretic descriptions of a) a standard physical system, $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$ and b) a neural system, $\mathbf{x}(t) = h(t) * (\mathbf{A}'\mathbf{x}(t) + \mathbf{B}'\mathbf{u}(t))$. The input signal, $\mathbf{u}(t)$, can be modified by the parameters in the input matrix, \mathbf{B} , before being added to any recurrent signal which is modified by the parameters in the dynamics matrix, \mathbf{A} . The result is then passed through the transfer function which defines the dynamics of the state variable, $\mathbf{x}(t)$. In a), the canonical form, the transfer function is integration. In b), the neural form, the transfer function is determined by intrinsic neural dynamics. Fortunately, given the canonical form and the transfer function, $h(t)$, \mathbf{A}' and \mathbf{B}' can be determined for any \mathbf{A} and \mathbf{B} , for any linear, nonlinear or time-varying system.

2.5 Synthesis

The previous sections have defined the three principles of the NEF for characterizing neurobiological systems. However, it may not yet be clear how these principles interact, and, more importantly, how they are intended to map onto the observable properties of real neural systems.

Figure 4 depicts how these principles can be integrated in order to characterize the functioning of neurobiological systems at various levels of description. Specifically, figure 4 shows the components of a generic neural subsystem, including temporal decoders, population

decoders, control matrices, encoders, and the spiking neural nonlinearity. A series of such subsystems can be connected in order to describe larger neural systems, since both the inputs and outputs of the subsystems are neural spikes.

Additionally, figure 4 depicts what it means to suggest that neural representations are control theoretic state variables. The state variables are defined by the temporal and population decoders and encoders, the dynamics of the control system are defined by the control matrices, and any functions that must be computed in order to implement the control system can be estimated by replacing the appropriate representational decoders with transformational decoders. This figure also captures how the theoretical elements of this description map onto real neural systems. In particular, the control matrices, decoders, and encoders can be used to analytically compute the connection weights necessary to implement the desired control system in the neural population. The temporal decoders, as noted earlier, are mapped onto the postsynaptic currents (PSCs) produced in dendrites as a result of incoming neural spikes. Finally, the weighted dendritic currents arriving at the soma (cell body) of the neuron determine the output of the neural nonlinearity, i.e., the timing of neural spikes produced by neurons in this population.ⁱⁱ

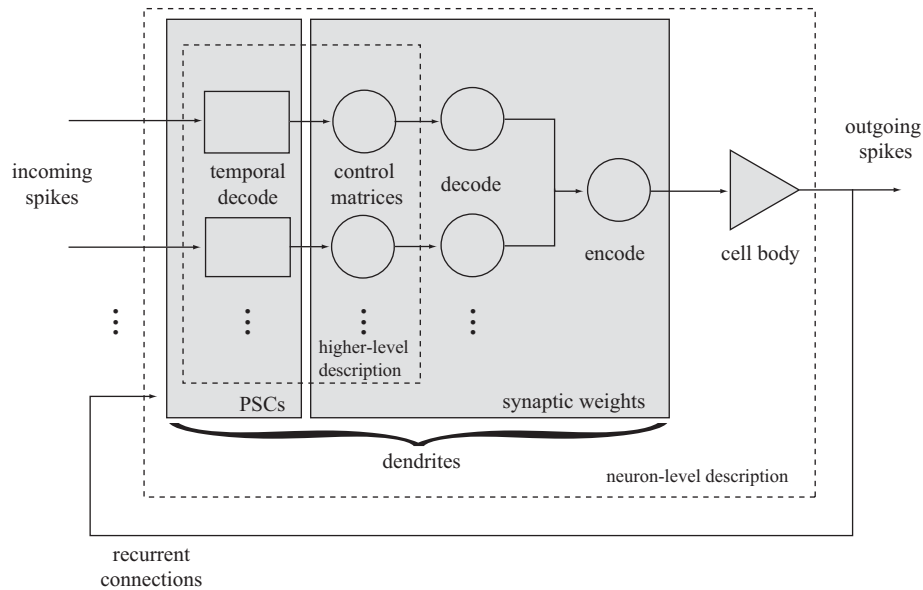


Figure 4: A generic neural subsystem. The outer dotted line encompasses elements of the neuron-level description, including PSCs, synaptic weights, and the neural nonlinearity in the soma. The inner dotted line encompasses elements of the control theoretic descriptions at the higher-level. The grey boxes identify experimentally measurable elements of neural systems. The elements inside those boxes denote the theoretically relevant components of the description. See text for details (adapted from C. Eliasmith, 2003).

Notably, the theoretical elements in this description are not identical to physically measurable properties of neural systems. As a result, there is a sense in which neural systems themselves never internally decode the representations they employ. This is because decoding, encoding, and the dynamics determined by the control matrices are all included in the synaptic weights, and so their individual effects are not measurable. Nevertheless, if our assumptions regarding representation or dynamics of the system are incorrect, the model which embodies these assumptions will make incorrect predictions regarding the responses of individual neurons. So, while we cannot directly measure decoders, we can justify their inclusion in a description of

neural systems insofar as the description is a successful one at predicting the properties we can measure (e.g., spike patterns and tuning curves). This, of course, is a typical means of justifying the introduction of theoretical entities in science. Encouragingly, this approach has been successfully applied to simulating and predicting the behaviour of a large number of neurobiological systems, including sensory, cognitive, and motor systems, as previously enumerated at the end of section 1.

3. Rat Navigation

To better ground the subsequent philosophical claims regarding the NEF, and to show an application of these principles, in this section I describe a detailed model of part of the rat navigation system first presented in Conklin and Eliasmith (2005).

The behaviour of interest for this model is called path integration. It has been observed that rats are able to return directly to a starting location in an environment after having searched the environment in a somewhat random path (Alyan & McNaughton, 1999; Tolman, 1948; see figure 5a). Notably, in these experiments, the only available cues for the rat are self-motion (i.e., only idiothetic cues). As a result, it has been hypothesized that the rat constantly updates an internal representation of its location in an environment relative to its starting point. Based on numerous neurophysiological investigations, it has been demonstrated that the representation can be thought of as a 'bump' of activity which is centered on the rat's current estimate of its location (see figure 5b).

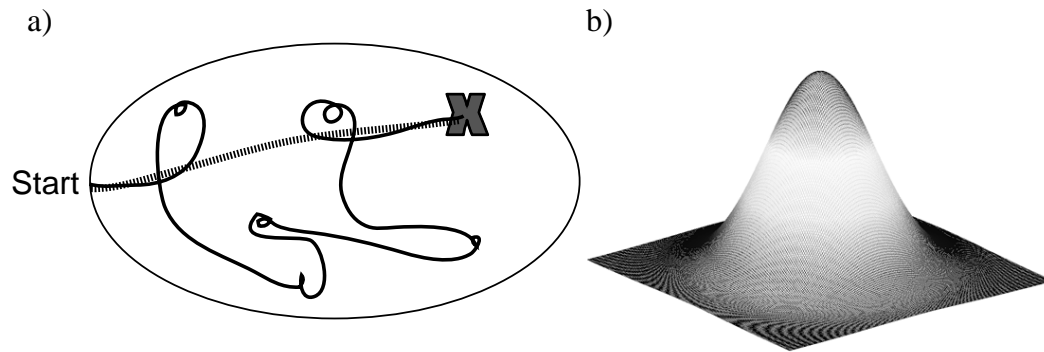


Figure 5. a) Path integration in the rat. This is a schematic illustration of the rat's behaviour while searching for a target ('X'). Not knowing the target location, it searches the environment (solid line) somewhat haphazardly. However, upon locating the target, it is able to plot a course (dashed line) directly back to its starting position. b) Internal representation of location in the rat. This is an idealization of the internal neural representation in a rat, if it were located in the middle of the environment. This type of plot is generated by topographically arranging the receptive fields of cells in the relevant areas (subiculum, parasubiculum, and superficial layers of the entorhinal cortex). The bump indicates the firing rates of neurons in a population so arranged, and the centre of the bump indicates the current estimate of the rat's location.

The challenge from a modeling point of view is to take the known physiological and anatomical properties of neurons in these areas and suggest an arrangement of such neurons that could implement a path integration mechanism that reproduces this observed behaviour. Adopting the NEF, Conklin and I addressed this challenge with a detailed neural model that captures a variety of behavioural and neural observations and provides novel predictions.

In our paper, we begin by characterizing this bump of activity as a two-dimensional bell-shaped function representation. The nonlinear encoding is determined by the response properties of neurons observed in these brain areas (Sharp), and the neural model we employ (a leaky integrate-and-fire (LIF) neuron).

Next, we suggest a high-level mechanism for performing path integration using only self-motion velocity commands available from the vestibular system. The details of the mechanism are not essential. However, it is notable that this suggestion is in the form of a dynamical/control system whose state variables are used to represent the two-dimensional bump. In short, we define a stable function attractor which will ‘hold’ a bump at any location without input (i.e., when the rat is not moving), and then slide the bump to a topographically appropriate position given self-motion information. This specification is done independently of neural implementation.

We then embed the representation into the suggested control system as suggested by principle 3. The necessary representational transformations to implement the system in neurons are accomplished by drawing on the second, computational principle of the NEF. In short, we calculate the feedforward and feedback connections between the relevant populations of neurons by incorporating the appropriate dynamics, decoding and encoding matrices into the neural weights. Interestingly, we find that the resulting weight matrices have a centre-surround organization, consistent with observed connectivity patterns in these parts of cortex.

Most significantly, we provide a number of simulations of the resulting model (see figure 6).

These simulations are the result of the activity of 4000 spiking neurons. Notably, there is small drift error (11%) in completing a circular path (figure 6a). This demonstrates that the model has no biases in any particular direction while updating the representation. This also compares very

favourably to the best past model of path integration, which had an error of approximately 100% using a network of 300 000 neurons to traverse a circular path (Samsonovich & McNaughton, 1997).

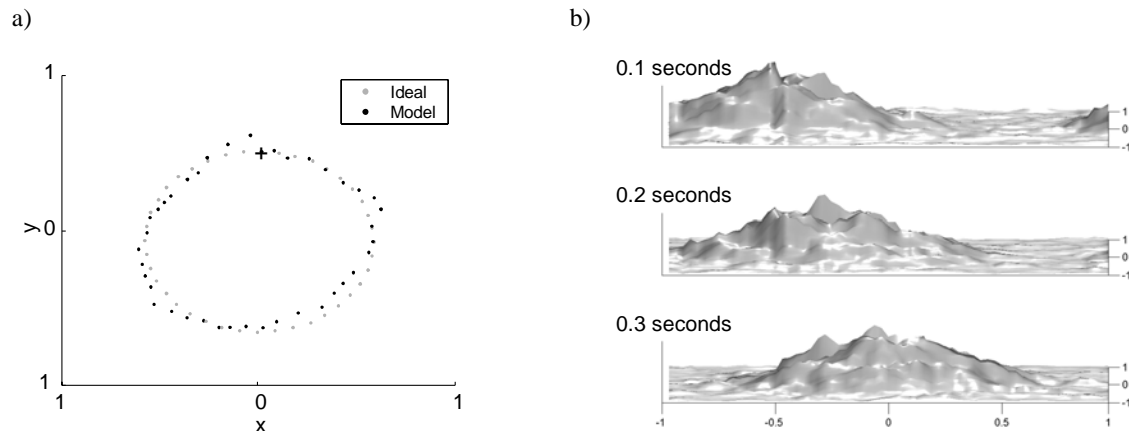


Figure 6. a) The performance of the path model on a circular path (i.e., heading in every direction). The black dots indicate the centre of the bump moving clockwise at samples taken every 50ms. b) Three points in time of the bump moving in a linear path from left to right. The bump is a result of counting spikes in a sliding window to determine firing rate. It is smoothed with a 5-point moving average for legibility. (Adapted from Conklin and Eliasmith, 2005.)

A number of other results of the model are of particular interest for subsequent discussion:

Phase precession. When a theta rhythmⁱⁱⁱ is introduced into the model as global excitation, two phasic phenomena observed in rats are also observed in the model: phase precession (acceleration and deceleration of the representation in phase with theta); and phasic bump width changes.

Types of tuning curves. The model shows the variety of tuning curves observed *in vivo*, including directionally selective cells in opposing directions, and non-selective cells.

Tuning curve resemblance. When tuning curves are generated from model data using the same methods as for real cells (i.e., random foraging paths), details of the curves are very similar.

Sensory input for calibration. The model shows the same effects as observed in the rat for weak (smooth, accelerating updating of the represented location) and strong (a ‘jump’ in the location of the population activity) sensory input.

In addition to these replications of available data, the model makes three main predictions. First, the model predicts that all cells will be velocity (as well as position) sensitive, if they are probed correctly. Second, the model predicts that cells in the path integrator (unlike typical ‘place cells’) will have the same relative location in different environments. Third, path integration should work regardless of the head direction of the rat (e.g., if the rat’s head is pointing in a direction other than its direction of motion, this will not affect path integration). This last prediction is in contradiction with past models, and so is of particular interest.

4. Philosophical consequences

There is much to be said about the consequences of the NEF for our understanding of neural systems: that aspect of the framework has been extensively discussed in neuroscientific journals. In contrast, much less has been written about the philosophical consequences of this means of characterizing neural systems (although see Eliasmith, 2003). In this section I briefly comment on several philosophical issues to which the NEF is relevant. These include the unity

of science (i.e., theory reduction), theory construction in the behavioural sciences, and mental representation.

4.1 Theories, models, and levels

On occasion, cognitive scientists have expressed hostility towards the idea that our understanding of cognitive processes can be improved by a better understanding of how the brain functions (Fodor, 1999; Jackendoff, 2002; Lycan, 1984). They have suggested that knowing where things happen in the brain, or how they are implemented in the brain, is not relevant for answering the more important question of what the cognitive architecture is. These kinds of arguments have driven some to a view of brain sciences that draws sharp distinctions between ‘cognitive’ and ‘biological’ approaches to understanding neural function (Davies, 2000). More generally, this has been taken to suggest that science, in general, is not unified (Fodor, 1974). Often, the reduction of psychology to biology motivates these concerns precisely because the gap between these two modes of characterizing neural systems strikes many as unsurmountable. Despite the adoption of fMRI, PET, ERP and similar research methods by psychologists, there are no well-established techniques for integrating such temporally or spatially ‘broad’ views of neural function with work in the electrophysiology or biochemistry of neurons. Similarly, models that incorporate biologically realistic single neurons tend to focus on low-level perception (e.g., receptive fields, motion, contour sensitivity, etc.), motor control (saccade generation, vestibular ocular reflex, invertebrate locomotion, digestion, etc.), and single-cell learning (e.g., retinal wave effects, receptive field learning, cortical column organization, etc.). Given the principles and applications of the NEF, it seems plausible to suggest that this theoretical approach to neural systems stands to bridge this gap. That is, the NEF integrates

single cell models (via nonlinear encoding) with cognitive mechanisms (via control theoretic descriptions). This integration is highly precise (i.e., each principle is quantitative), relates directly to underlying, measurable physical processes, and scales as computational power permits. The rat model presented earlier provides a somewhat ‘uncognitive’ example of this sort of integration. However, Eliasmith and colleagues (Eliasmith & Conklin, forthcoming; Eliasmith, 2004) present a similarly derived (though much larger) model of language-based deductive inference which effectively characterizes the well-known Wason selection task (Wason, 1966), and exhibits cognitive behaviour found only in human reasoners. Models of these sorts make the brain sciences look much less ‘disunified’ than has been argued in the past.

In fact, I would like to propose that application of the NEF can suggest a particular kind of scientific unification. In particular, we can see from the rat model that the application of the general NEF principles boils down to the continually more precise specification of sets of boundary conditions. Some of these conditions are in the form of a hypothesis regarding the dynamics (e.g., integration), some are in the form of a control mechanism (e.g., the particular matrices defining the path integrator), and others are in the form of empirically measured neuron tuning curves (e.g., the distribution of tuning curves of cells in subiculum). But, in general, as we progress from general theory (NEF) to specific model (rat path integration) we have followed a route of increasing specificity. Were we to continue on that route, that is, were we to match each model neuron with a specific neuron in a specific rat (rather than matching neuron property distributions across rats), we would end up with a highly detailed (perhaps too detailed) model that would be able to make specific predictions regarding the behaviour of a particular rat. Of course, the model as originally presented is not interested in such questions, so the level of specificity stopped much earlier. Nevertheless, this notion that increasingly detailed boundary

conditions can serve to traverse a model-to-theory hierarchy seems a potentially useful description of how the brain sciences may be unified. The unification stems from the fact that, at each step in the specification, the basic principles do not change.

Unlike past models, those derived from the NEF can be placed in such a ‘unification hierarchy.’ This allows for general principles of neural function that might not otherwise be evident to become obvious. For instance, the close relation between rat path integration, mechanisms of horizontal eye control by the nucleus prepositus hypoglossi, the head direction system, and working memory (all of which can be characterized as vector integrators) might have gone unnoticed without this principled underpinning (see Eliasmith, 2005 for details of these and related generalizations).

While much remains to be said about the potential unification or disunification of brain sciences, the NEF provides a first plausible and detailed story about how such a unification might be worked out. As a result, it is in a unique position to support replies to the notion that “the description of mental processes at the cognitive level can be divorced from the description of their physical realization” (Fodor & Pylyshyn, 1988, p. 54).

4.2 Theory construction in the behavioural sciences

I have argued elsewhere that current approaches to cognitive science have their theoretical foundations grounded in metaphor (Eliasmith, 2003). I there suggest that the NEF is unique in its avoidance of metaphor for theoretical insight (though not for explanation). An important consequence of this reliance on metaphor by past approaches that I have not emphasized in past discussions is the adverse effect it has on theory construction in the behavioural sciences. In short, if the theoretical import of a metaphor is not explicitly grounded,

that is, directly related to independently observable quantities, then it will be difficult to construct empirically testable theories. I take notions like ‘empirical testability’ and ‘directly related’ to be matters of degree. However, in both cases, the more the better.

I would like to suggest that the NEF is unique among approaches to the behavioural sciences in that the elements of the theory are (very) directly relatable to observables, and hence the models stemming from the NEF are highly testable. So, for instance, the elements in the path integration model can be mapped directly onto elements in the brain (or, more precisely, the distribution of element properties can be so mapped). That is, model neurons produce spike rates, tuning curves, spike patterns, somatic currents, etc. that can be compared directly to spike rates, tuning curves, spike patterns, somatic currents, etc. in real rat neurons. In contrast, when classical cognitive science assumes the mind is like a computer, and imports notions of data-structure, symbolic representation, etc. into behavioural models based on such a metaphor, we are left wondering how such theoretical entities relate to the physical system that they are supposed to be describing. This is largely because we do not know how to measure (independent of this theory) such entities in a real behaving system. In short, the ability of NEF generated models to be mapped in detail onto independently measurable physical properties makes them more convincing (because more empirically testable), than those generated by classical models. But what of other approaches in the behavioural sciences? Let me briefly consider two other paradigms, dynamicism and connectionism.

The ‘dynamicist’ view in cognitive science emphasizes dynamical descriptions for behavioural systems (Port & van Gelder, 1995). However, there is an extremely important difference between these dynamical descriptions and those generated by the NEF. For dynamicists, the variables over which the differential equations are defined are not explicitly

related to the physical system itself (Eliasmith, 1997; Eliasmith, 2003). So, for example, the “motivation” variable in motivational oscillatory theory (MOT), which is intended to characterize some high-level property of the animal, is never related to any specific physical property of the system (Busemeyer & Townsend, 1993) – and it is entirely unclear how it could be. In contrast, the NEF is explicit on the relation between higher-level neural representations and the activations of single cells. And, it is precisely these representations that serve as the variables over which the dynamics are defined. Again as in the classical case, the explicitness of the mapping, and hence empirical import, of models generated by the NEF is far more impressive.

What about connectionism? The difficulty with connectionism is precisely that, while endorsing ‘brain-like’ models, it is far too abstracted from neurobiological constraints. That is, because only the barest features of neural architectures are preserved in connectionist models, it is difficult to relate the resulting models back to real brains. While I earlier expressed concern with Fodor and Pylyshyn’s distancing of cognitive from neural approaches to the behavioural sciences, they clearly understand the importance of exploiting any available constraints:

Understanding both psychological principles and the way that they are neurophysiologically implemented is much better (and, indeed, more empirically secure) than only understanding one or the other. That is not at issue. The question is whether there is anything to be gained by designing “brain style” models that are uncommitted about how the models map onto brains. (1988, p. 62)

While Fodor and Pylyshyn suspect that there is no interesting interplay between these levels of description, they do realize the import of both. I am suggesting that the NEF helps integrate high-level and low-level approaches, thus providing more empirically secure models that can draw directly from various levels of description in the behavioural sciences. The path integration model is able to make the specific, empirical predictions it does precisely because of the detailed mapping between the model and the neurobiological system it simulates. Notably, these predictions are both behavioural (relating head direction to integration bias), and neural (predicting specific relations between neural tuning curves and the rat's environment). The ability of a single model to successfully address multiple levels of description make it reasonable to think that Fodor and Pylyshyn's dismissal of intertheoretic unity is premature.

In conclusion, the lesson to be learned is that the more explicit and independently testable the mappings of your theories or models, the more success you will have in constructing interesting theories. In the behavioural sciences, the NEF uniquely provides such mappings across traditionally distinct levels of description.

4.3 Mental representation and semantics

One useful distinction that has arisen out of the philosophical discussion of representation is that between the contents (semantics) and the vehicles (syntax) of representations (Cummins, 1989; Fodor, 1981). In general, contents are thought to be determined by the object in the world that the representation picks out, the relation of that representation to other representations, or a combination of both. The vehicles of representations are the physical realization of objects that play the role of representations (i.e., carrying a content) in a system. Theoretical neuroscience

can contribute to improving our understanding of both representational vehicles and their contents.

The characterization of representation in section 2.2 is most clearly about vehicles. The NEF (and previous work on which it draws) tells us how to characterize structures in the brain as able to carry contents as values of scalars, vectors, functions, etc. I suspect this characterization is sufficient for understanding the complete set of vehicles available to neural systems.

However, let me emphasize one particular class that is generally overlooked by the philosophical community. As far as I am aware, there are no discussions in the philosophical literature of representational systems that include the uncertainty of the representations in the representations themselves. Because the NEF can describe how neural systems represent functions, it can describe how neural systems represent probability distributions over possible states of the world – just such a representation of uncertainty. In fact, the rat’s ‘bump’ of activity can be interpreted as just such a representation. Rather than supposing that the centre of the bump is representing the rat’s location, we can more interestingly suppose that the bump represents a distribution of the possible locations of the rat with varying certainties. Thus, a wider bump would indicate more uncertainty (variance) in the representation, and a narrower bump less uncertainty. In either case, the best estimate of the rat’s actual location will be the mean, but understanding the bump in this way increases the amount of information that such a representation carries, and allows the representation to support sophisticated reasoning strategies that employ statistical inference (Eliasmith & Anderson, 2003).

This is not merely idle speculation. There is increasing evidence that precisely this kind of representation is used to encode information about the uncertainty of the estimate of the stimulus being encoded, and that this information is used by the nervous system for (nearly

optimal) statistical inference (Britten & Newsome, 1998; Knill & Pouget, 2004; Kording & Wolpert, 2004; Stocker & Simoncelli, 2006). It seems essential, given the noisy, complex, and uncertain environment in which neurobiological systems reside, for such systems to be able to make decisions with partial, incomplete, or noisy data. So, it is only to be expected that the representations in neural systems can support statistical inference. In sum, an important class of vehicles for understanding minds, those that carry a content and its uncertainty, has been overlooked in philosophical discussions.

Let me now turn to a brief consideration of semantics (see Eliasmith, 2006 for a more detailed discussion of a semantic theory consistent with the NEF). There are three broad classes of semantic theories: causal, conceptual role, and two-factor theories. Causal theories of meaning have as their main thesis that mental representations are about, and thereby mean, what causes them (Dretske, 1981; Dretske, 1995; Fodor, 1990; Fodor, 1998). In the context of the previous discussion this means that the encoding process alone determines meaning. Conceptual role theories hold that the meaning of a term is determined by its overall role in a conceptual scheme (Harman, 1982; Loar, 1981). Under such theories, the meaning of a term is determined by the inferences it causes, the inferences it is the result of, or both. Here, the focus is on the decoding of whatever information happens to be in some neural state.

One theoretical move, to avoid the difficult problems that arise when adopting either a causal theory or a conceptual role theory, is to combine them into a 'two-factor' theory (Block, 1986; Field, 1977). On two-factor theories, causal relations and conceptual role are equally important, independent elements of the meaning of a term: "the two-factor approach can be regarded as making a conjunctive claim for each sentence" (Block 1986, p. 627). So, only two-factor theories explicitly acknowledge both encoding and decoding.

Given the NEF characterization of representational vehicles, a representation is only defined once both the encoding and decoding processes are identified. This means that, contrary to both causal and conceptual role theories of content, both how the information in neural spikes is used (decoding), as well as how it is related to previous goings-on (encoding) are relevant for determining content. In the rat example, the bump indicates the rat's location precisely because it is caused by (or correlated with) the rat's actual location in the world, and because it is used by the rat to determine how to move (e.g., making a bee-line back to the starting location once the goal has been achieved). So given the characterization of vehicles we have seen, two-factor theories of content seem most plausible.

However, as just noted it is assumed by past two-factor theories that the factors are *independent*. This property raises a grave difficulty for such theories. In criticizing Block's theory, Fodor and Lepore remark "We now have to face the nasty question: *What keeps the two factors stuck together?* For example, what prevents there being an expression that has the inferential role appropriate to the content *4 is a prime number* but the truth conditions appropriate to the content *water is wet?*" (1992, p. 170). If, in other words, there is no relation between the two factors (i.e., they are simply a conjunction), it is quite possible that massive misalignments between causal relations and conceptual roles can occur.

However, in the NEF characterization of representation there is a tight relation between the encoding and decoding processes – they are not independent. Broadly speaking, the population-temporal decoders are found in order to estimate some function of the encoded parameter. While for simple representation this function is identity, it need not be in the case of transformational decoding. That is to say, all of the inferences derivable from some particular neural encoding depend on the information carried by that encoding. As a result, if there is no

relation between the 'wetness of water' and '4 being a prime number', it would it be impossible for the latter to be part of the conceptual role of the encoding of the former given the NEF characterization. Such considerations suggest that this characterization can avoid the main weakness of past two-factor theories, though more work must be done to propose a full-fledged theory of representation.

5. Conclusion

Admittedly, each of these discussions of the philosophical consequences of the NEF are deserving of much more careful, article-length treatment. Nevertheless, hopefully these suggestions demonstrate the utility of looking beyond traditional philosophical or psychological approaches to the behavioural sciences for addressing a variety of philosophical problems.

It is also worth emphasizing that there is much subtlety to the NEF itself which has not been addressed here. The framework has been designed to account for the ubiquitous effects of noise, to allow novel methods for the analysis of learning rules, and to incorporate the wide variety of single cell dynamics (e.g., adaptation, bursting, etc.) seen across neurobiological systems. Each of these developments may also serve to highlight new approaches to related philosophical problems.

Acknowledgements

Parts of sections 2 and 4 of this article are based on Eliasmith (2007).

References

References

- Alyan, S. H., & McNaughton, G. L. (1999). Hippocampectomized rats are capable of homing by path integration. *Behavioral Neuroscience, 113*, 19-31.
- Andersen, R. A., Essick, G. K., & Siegel, R. M. (1985). The encoding of spatial location by posterior parietal neurons. *Science, 230*, 456-458.
- Block, N. (1986). Advertisement for a semantics for psychology. In P. French, T. Uehling & H. Wettstein (Eds.), *Midwest studies in philosophy* (pp. 615-678). Minneapolis: University of Minnesota Press.
- Britten, K. H., & Newsome, W. T. (1998). Tuning bandwidths for near-threshold stimuli in area MT. *Journal of neurophysiology, 80*(2), 762-770.
- Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review, 100*(3), 432-459.
- Conklin, J., & Eliasmith, C. (2005). An attractor network model of path integration in the rat. *Journal of computational neuroscience, 18*, 183-203.
- Cummins, R. (1989). *Meaning and mental representation*. Cambridge, MA: MIT Press.
- Davies. (2000). Interaction without reduction: The relationship between personal and sub-personal levels of description. *Mind & Society, 1*(2), 87.
- Dayan, P., & Abbott, L. H. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

- Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of cognitive neuroscience*, 3, 1-8.
- Dretske, F. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.
- Dretske, F. (1988). *Explaining behavior*. Cambridge, MA: MIT Press.
- Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.
- Eliasmith, C. (2007). Computational neuroscience. In P. Thagard (Ed.), *Philosophy of psychology and cognitive science*. Amsterdam: Elsevier.
- Eliasmith, C. (2006). Neurosemantics and categories. In C. Lefebvre, & H. Cohen (Eds.), *Handbook of categorization in cognitive science*. Amsterdam: Elsevier.
- Eliasmith, C. (2005). A unified approach to building and controlling spiking attractor networks. *Neural computation*, 17(6), 1276-1314.
- Eliasmith, C. (2004). Learning context sensitive logical inference in a neurobiological simulation. In S. Levy, & R. Gayler (Eds.), *AAAI fall symposium: Compositional connectionism in cognitive science*. AAAI Press, pp. 17-20.
- Eliasmith, C. (2003). Moving beyond metaphors: Understanding the mind for what it is. *Journal of Philosophy*, 100(10), 493-520.
- Eliasmith, C. (2003). Neural engineering: Unraveling the complexities of neural systems. *IEEE Canadian Review*, 43, 13-15.
- Eliasmith, C. (1997). Computation and dynamical models of mind. *Minds and Machines*, 7, 531-541.
- Eliasmith, C., & Conklin, J. (forthcoming). How to build biologically plausible cognitive models: With an application to the Wason card task.

- Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.
- Eliasmith, C., Westover, M. B., & Anderson, C. H. (2002). A general framework for neurobiological modeling: An application to the vestibular system. *Neurocomputing*, 46, 1071-1076.
- Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in primate visual cortex. *Cerebral Cortex*, 1, 1-47.
- Field, H. (1977). Logic, meaning, and conceptual role. *Journal of Philosophy*, 74, 379-409.
- Fischer, B. (2005). A model of the computations leading to a representation of auditory space in the midbrain of the barn owl. (PhD, Washington University in St. Louis).
- Fodor, J. (1999). Let your brain alone. *London Review of Books*, 21(19)
- Fodor, J. (1998). *Concepts: Where cognitive science went wrong*. New York: Oxford University Press.
- Fodor, J. (1990). *A theory of content and other essays*. Cambridge, MA: MIT Press.
- Fodor, J. (1981). *Representations*. Cambridge, MA: MIT Press.
- Fodor, J. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, 28(2), 97.
- Fodor, J., & Lepore, E. (1992). *Holism: A shopper's guide*. Oxford, UK: Basil Blackwell.
- Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3-71.
- Georgopoulos, A. P., Lurito, J. T., Petrides, M., Schwartz, A., & Massey, J. (1989). Mental rotation of the neuronal population vector. *Science*, 243, 234-236.

- Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, *243*, 1416-19.
- Harman, G. (1982). Conceptual role semantics. *Notre Dame Journal of Formal Logic*, *23*, 242-256.
- Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, *117*, 500-544.
- Hoppensteadt, F., & Izhikevich, E. (2003). Canonical neural models In M. Arbib (Ed.), *The handbook of brain theory and neural networks*. Cambridge, MA: MIT Press.
- Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology (London)*, *160*, 106-154.
- Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.
- Kalman, R. E. (1960a). Contributions to the theory of optimal control. *Boletin de la Sociedad Matemática Mexicana*, *5*, 102-119.
- Kalman, R. E. (1960b). A new approach to linear filtering and prediction problems. *ASME Journal of Basic Engineering*, *82*, 35-45.
- Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in neurosciences*, *27*(12), 712-719.
- Koch, C. (1998). *Biophysics of computation: Information processing in single neurons*. Oxford, UK: Oxford University Press.
- Kording, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244-247.

- Kuo, D., & Eliasmith, C. (2005). Integrating behavioral and neural data in a model of zebrafish network interaction. *Biological Cybernetics*, 93(3), 178-187.
- Lettvin, J., Maturana, H., McCulloch, W., & Pitts, W. (1959/1988). What the frog's eye tells the frog's brain. In W. McCulloch (Ed.), *Embodiments of mind*. Cambridge, MA: MIT Press.
- Loar, B. (1981). *Mind and meaning*. London, UK: Cambridge University Press.
- Lycan, W. (1984). *Logical form in natural language*. Cambridge, MA: MIT Press.
- Moran, D. W., & Schwartz, A. B. (1999). Motor cortical representation of speed and direction during reaching. *Journal of Neurophysiology*, 82, 2676-2692.
- Platt, M. L., & Glimcher, G. W. (1998). Response fields of intraparietal neurons quantified with multiple saccadic targets. *Experimental Brain Research*, 121, 65-75.
- Platt, M. L., & Glimcher, G. W. (1997). Responses of intraparietal neurons to saccadic targets and visual distractors. *Journal of Neurophysiology*, 78, 1574-1589.
- Port, R., & van Gelder, T. (1995). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.
- Rall, W. (1962). Theory of physiological properties of dendrites. *Annual New York Academy of Science*, 96, 1071-1092.
- Rall, W. (1957). Membrane time constant of motoneurons. *Science*, 126, 454.
- Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.
- Salinas, E., & Abbott, L. (1994). Vector reconstruction from firing rates. *Journal of computational neuroscience*, 1, 89-107.
- Samsonovich, A., & McNaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor model. *Journal of Neuroscience*, 17, 5900-5920.

- Schwartz, A. B. (1994). Direct cortical representation of drawing. *Science*, 265, 540-542.
- Sereno, A. B., & Maunsell, J. H. R. (1998). Shape selectivity in primate lateral intraparietal cortex. *Nature*, 395, 500-503.
- Sharp, P. E. (1997) Subicular cells generate similar spatial firing patterns in two geometrically and visually distinctive environments: Comparison with hippocampal place cells. *Behavioural and Brain Research*, 85, 71-92.
- Singh, R., & Eliasmith, C. (2006). Higher-dimensional neurons explain the tuning and dynamics of working memory cells. *Journal of Neuroscience*, 26, 3667-3678.
- Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, 9(4), 578-585.
- Todorov, E. (2000). Direct cortical control of muscle activation in voluntary arm movements: A model. *Nature Neuroscience*, 3, 391-398.
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55, 189-208.
- Tripp, B., & Eliasmith, C. (in press). Neural populations can induce reliable postsynaptic currents without observable spike rate changes or precise spike timing. *Cerebral Cortex*,
- Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology*. Harmondsworth: Penguin, pp. 135-151.
-

i A search of PubMed indexed neuroscience journals (53 journals) over the last 5 years for the term 'represent*' returns over 2300 hits.

ii This nonlinearity can be captured by a set of differential equations that describes the dynamics of the channel conductances that control the flow of ions through the cell membrane resulting in action potentials, or it could be a simpler reduced model of neural spiking (like the common leaky integrate-and-fire model).

iii In much of the hippocampal complex, there is a globally observable voltage oscillation at 7-10Hz. This is called the theta rhythm.