Synthese (2007) 159:373–388
DOI 10.1007/s11229-007-9235-0

How to build a brain: from function to implementation

Chris Eliasmith

Published online: 2 September 2007

Abstract To have a fully integrated understanding of neurobiological systems, we must address two fundamental questions: 1. *What* do brains do (what is their function)? and 2. *How* do brains do whatever it is that they do (how is that function implemented)? I begin by arguing that these questions are necessarily inter-related. Thus, addressing one without consideration of an answer to the other, as is often done, is a mistake. I then describe what I take to be the best available approach to addressing both questions. Specifically, to address 2, I adopt the Neural Engineering Framework (NEF) of Eliasmith & Anderson [*Neural engineering: Computation representation and dynamics in neurobiological systems*. Cambridge, MA: MIT Press, 2003] which identifies implementational principles for neural models. To address 1, I suggest that adopting statistical modeling methods for perception and action will be functionally sufficient for capturing biological behavior. I show how these two answers will be mutually constraining, since the process of model selection for the statistical method in this approach can be informed by known anatomical and physiological properties of the brain, captured by the NEF. Similarly, the application of the NEF must be informed by functional hypotheses, captured by the statistical modeling approach.

Keywords Neural architecture · Functional integration · Neurophilosophy · Cognitive architecture · Statistical models · Mental representation · Neural networks

1 Introduction

Theoretical approaches to cognitive science (which I take to include both psychology and neuroscience) often attempt to construct *models* of human or animal behavior.

C. Eliasmith (✉)
Department of Philosophy, University of Waterloo, Waterloo, ON, Canada N2L 3G1
e-mail: celiasmith@uwaterloo.ca

These neurocognitive models are unique in science in that there are often two distinct modeling relations of relevance to their construction. Usually, when developing a theoretical description of a physical system, a scientist needs to concern himself or herself solely with the most effective way to quantify the observed behavior of the system. This is true, for instance, when modeling mechanical, chemical, environmental, geological and other such physical systems. This characterization, however, does not accurately describe the task undertaken by theorists in cognitive science. This is because cognitive modeling essentially entails a kind of "meta-modeling"—modeling a system itself taken to be modeling its environment. The system that the neurocognitive theorist is attempting to describe is taken to have its own internal model (or representation) of the world. As a result, building models in the cognitive sciences means it is essential to address two modeling relations; that between our description and the physical system, and that between a physical system itself and the world.

When considering questions of functional integration, both of these modeling relations are important to consider. And, I believe that these two relations can be captured by answers to the following two questions:

1. What do brains do (what is the relation between the system and its environment)?
2. How do brains accomplish their functions (what is the relation between physically measurable variables of the system and our quantitative description of their interactions across various levels of detail)?

By merely supposing that there are two modeling relations addressed by cognitive theories, we have delineated a reply to the first question: brains build and employ (adaptive, partial, approximate, etc.) models of the world. Of course, this is not a satisfactory answer to that question because it is far too vague and so we must answer it more detail. I will outline what I take to be a promising approach to characterizing the appropriate class of models in Sect. 3.[1]

Notice, however, that addressing the system-world relation (i.e., taking cognitive systems to model the world) cannot complete our theoretical characterization of the system. After all, we have not yet said anything about the "usual" modeling relation, that between the physical system and our mathematical description of it. As a result, it is crucial for theoreticians in cognitive science to address this relation, captured by question two. That is, it is essential to explicitly describe how the physical brain could implement and use the model we take it to be constructing, given our answer to question one.

Supposing that there are two distinct questions which must be addressed by cognitive models, what is the relation between them? Before describing their specific relation, it is first important to establish whether they are related at all. Here, I argue that they are intimately related. Furthermore, it is fair to say that the vast majority of work in psychology and neuroscience has been focused on one question or the other—seldom concurrently addressing both. I would also suggest that it is not unfair to assert that psychology has focused on the first question, whereas neuroscience has

[1] The question of whether or not biological systems model the world is beyond my current scope. I presume that they do, and refer interested readers to relevant psychological and neuroscientific work (e.g., Johnson-Laird 1983; McIntyre et al. 2001; Wolpert et al. 1998).

attended largely to the second. For example, the vast majority of psychological models have not worried about biological realism, assuming that it can be added subsequently (including classical (e.g., SOAR Newell 1990), connectionist (e.g., NETtalk Sejnowski and Rosenberg 1986), and dynamical (e.g., MOT Busemeyer and Townsend 1993) work).[2] Similarly, the vast majority of work in theoretical neuroscience has characterized implementation issues (e.g., information transfer, Rieke et al. 1997; "fine-tuning" of neural integrators, Koulakov et al. 2002; attractor networks, Amit 1989, etc.). Admittedly, since there must be some function that is implemented by a neural system, the work in theoretical neuroscience cannot completely avoid the issue of function (just as the work in psychology cannot completely avoid issues of implementation). Nevertheless, the functions theoretical neuroscientists have focused on tend to be simple, low-dimensional, and considered in isolation (i.e., not as part of a larger, functionally integrated system, or internal model).

While a divide and conquer approach may often be reasonable in dealing with a system as complex as the brain, such an approach is seldom, if ever, successful when pursued in isolation; that is, without equal consideration given to the synthesis of the parts (Bechtel and Richardson 1993). Thus, it is a mistake to solely consider function and implementation as distinct as seems to be the status quo in psychology and neuroscience. Rather, it is essential to address the theoretical issue of large-scale, biologically plausible functional integration in a unified manner. It is difficult to overstate how difficult this challenge is. As a result, in this paper my goal is not to provide a polished solution to this challenge, but rather to suggest, and provide a reasonable amount of detail about, approaches which I think hold the most promise for successfully meeting this challenge.

The remainder of this essay is structured as follows: I first introduce what I take to be a promising method for relating function to biologically realistic implementation (i.e., an answer to question two); I then introduce what I take to be a promising approach to addressing biologically relevant functions (i.e., an answer to question one); next, I address how these two methods can be integrated to provided the kind of unified approach to understanding cognitive behaviors that I have just suggested is essential. I conclude with a brief discussion of the implications of this view.

2 Implementation and the NEF

In recent years, Charles Anderson and I have championed an approach to building large-scale biologically plausible models called the Neural Engineering Framework (NEF, Eliasmith and Anderson 2003). It consists of three basic principles, quantitatively characterized in the Appendix:[3]

[2] This is true even though degrees of biological inspiration may partly distinguish kinds of psychological model.

[3] While these principles have been extended in more recent work (Tripp and Eliasmith 2007), here I present their original formulation (Eliasmith and Anderson 2003) which is simpler and does not detract from subsequent discussion.

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.

Neural representation (principle 1) is thus characterized by: (1) the (nonlinear) neuron tuning curve, which typically captures the relation of the response of a given cell to a stimulus (e.g., Gaussian tuning to the angle of a bar in the receptive field); and (2) a theoretically defined neural decoder. This decoder is not directly empirically observable, unlike the tuning curve, but rather captures what information is extractable from the given response of the cell. Notably, the decoder still has empirical consequences (namely, the size of synaptic weights), though these are only accessible in the context of a circuit. For instance, if a circuit was needed to estimate the angle of an encoded bar, the responses of all neurons sensitive to encoded bar angles could be pooled, weighted by their decoders and the receiving neurons' encoders (i.e., synaptic weight $\approx$ decoders $\times$ receiving_encoders), and then the subsequent population could be interpreted as representing the 'bar angle' scalar. This simple kind of representation can be similarly used to represent vectors, functions, vector fields or other kinds of mathematical objects.

Technically, the representation circuit described in the previous paragraph is the simplest possible transformation (principle 2): the identity function. That is, the scalar 'bar angle' is simply reproduced from one population to the next. If, rather than finding decoders which decode the information encoded in the original population, we find decoders that decode some function of that information (e.g., two times 'bar angle,' i.e., $f(x) = 2x$), we can similarly define neural connection weights that effect this transformation. The same is true for nonlinear functions as well (e.g., $f(x) = x^2$). In short, we can estimate any function by computing the appropriate linear decoders to extract that function from the encoded information. This holds regardless of the kind of mathematical object that is being represented.

Finally, the dynamics principle (principle 3) brings the first two together, and adds the crucial dimension of time to the circuits. Essentially this principle allows the representations of principle 1 to be combined with the transformations of principle 2 to define sophisticated dynamical circuits. For instance, if we take the simple representation circuit described earlier, which computes the identity function, and make the sending and receiving populations the same, we have constructed a 'neural integrator.' This recurrent circuit will act like a memory (given a state, it will constantly try to preserve that state over time, decoding now what was encoded at the previous time step, i.e., constantly recomputing the identity function). In short, this circuit defines a simple dynamical system, in terms of the representation defined by principle 1, using the transformation defined by principle 2. In fact, the integrator just described has been used by a number of authors to explain the function of the nuclei prepositus hypoglossi in the brain stem, that controls horizontal eye position (Koulakov et al. 2002; Seung 1996). The general relationship between the three principles and a spiking neural population is depicted in Fig. 1.

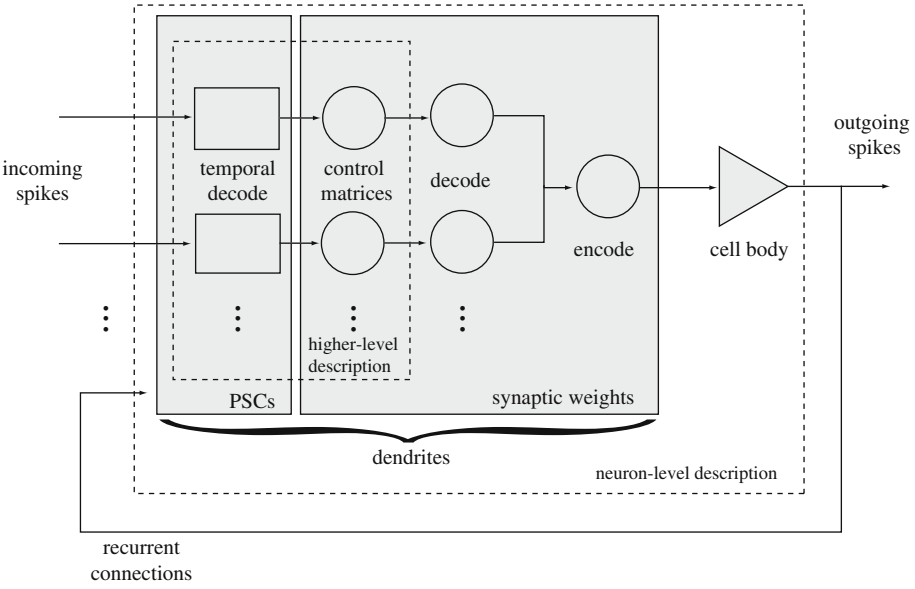

Fig. 1 A generic neural subsystem. The outer dotted line encompasses elements of the neuron-level description, including PSCs, synaptic weights, and the neural nonlinearity in the soma. The inner dotted line encompasses elements of the control theoretic descriptions at the higher-level. The gray boxes identify experimentally measurable elements of neural systems. The elements inside those boxes denote the theoretically relevant components of the description. For a formal description of these elements, see the Appendix (adapted from Eliasmith 2003)

In short, the NEF principles: (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.).

There are number of sources for a detailed discussion of these principles and their application in addition to their original formulation (see, e.g., Eliasmith 2005; Tripp and Eliasmith 2007). For the purposes of this paper, what is most relevant is that this approach has been widely applied to constructing novel, large-scale, biologically realistic models. These include models of the barn owl auditory system (Fischer 2005), the rodent navigation system (Conklin and Eliasmith 2005), escape and swimming control in zebrafish (Kuo and Eliasmith 2005), the translational vestibular ocular reflex in monkeys (Eliasmith et al. 2002), working memory systems (Singh and Eliasmith 2006), and language-based deductive inference (Eliasmith 2004). Notably, these models span sensory, motor and cognitive systems across the phylogenetic tree. Furthermore, the majority of these models have resulted in testable experimental predictions, some of which have been used to drive further experiment (see, e.g., Fischer et al. in press).

The broad applicability and success of this approach warrants the suggestion that it captures some fundamental aspects of the relevant constraints on neural implementation. Currently, there are no obvious competitors to the NEF as a general approach for constructing large-scale mechanistic models the brain to the level of individual, spiking neurons. As a result, it is natural to suggest this is our current best answer to the second question posed earlier: the principles of the NEF describe how functions are implemented in the brain.

However, it should be clear from looking at these principles that they do not answer questions regarding neural *function*. Instead, they define a kind of "neural compiler." Compilers, familiar from computer science, are methods of translation, not hypotheses about function. Of course, the important point about translation is that expressions in one language may take widely varying resources to re-express in another. Consequently, the mathematical expressions that are natural for describing certain functions may take an unacceptable number of neural resources to implement. This, then, puts significant and important constraints on what functions brains actually implement. It does not, however, tell us what those functions may be.

3 Function and statistical modeling

Statistical modeling has historical roots in data collection and analysis for characterizing political states (hence 'stat-istics'), and mathematical roots in probability and error theory. In these contexts, it is practical considerations of the world that drives the use of statistics. That is, descriptive statistics are used to describe the state of the world, including noise and variability in the measured quantities, and inferential statistics is used to pick out the important patterns in those often noisy, measured quantities. In other words, the tools of statistics have been developed to effectively describe complex relationships given real-world data. This, I take it, is a similar problem to that faced by biological systems.

What is important for cognitive science, and somewhat foreign to traditional approaches to the subject, is the centrality of uncertainty, ambiguity, and randomness in this understanding of biological function. Biological systems are not designed to be absolutely certain of the identity of an object ("there is a dog 3 feet away"), but rather they are designed to be certain enough of its identity to allow appropriate response ("that is probably a dog, and is too close for comfort"). To capture this deft handling of uncertainty, perceptual processes can be understood as constructing statistical models of perceptual data, which are used to infer likely states of the world given that data.

We can begin to formalize this characterization by supposing there is some 'data,' y, which is the contribution of world states to neural activity. The purpose of perceptual systems is to construct and use a statistical model $p(y)$ to be able to predict the data (and hence usefully characterize the world states). Because this ideal data distribution will be enormously complex (as it is the probability of all possible data at all times), it is natural to consider a parameterized model (where the dimensionality of the parameters is much smaller than the dimensionality of the data). The biological system thus

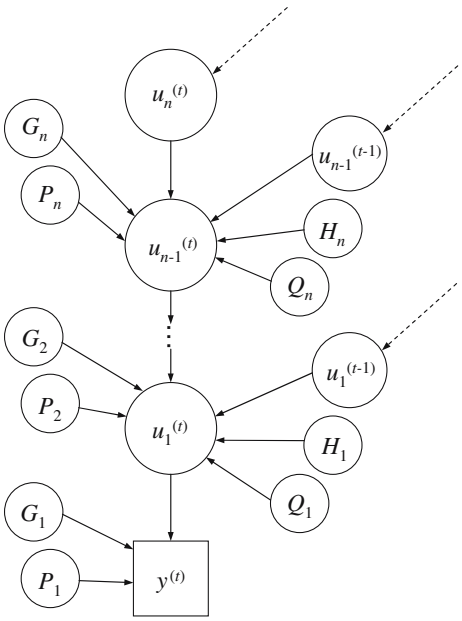

Fig. 2 A hierarchical statistical model. Parameters, indexed over time, t, and layer, i, include $u_i(t)$ (the hidden cause, i.e., neural activity), G_i and H_i (generative and predictive matrices, i.e., synaptic weights), P_i and Q_i (precision matrices, i.e., intra-layer synaptic weights on neurons computing error terms). The dependence relationships of the model parameters/hidden causes are defined by arrows. Dotted lines indicate additional dependencies from model parameters at future or past times

must estimate the distribution of the parameters in order to reason about (i.e., predict) the data. To estimate this distribution, the system needs data. As a result, a kind of bootstrapping process, i.e., learning, is necessary to construct this model. In practice, however, the parameterized model $p(y, \Phi)$, is also too complex to estimate directly. Instead, it has been found that a lower bound on this model can be defined, and model estimation by maximizing this lower bound is feasible (usually given various further assumptions). This method is variously designated Variational Bayes (VB), Maximization of Free Energy, or Product Distribution (PD) theory (Friston 2003; Hinton and van Camp 1993; Wolpert 2004).

Notably, these methods for optimal model inference do not specify the structure of the model itself. However, it has become clear that for many perceptual problems, a hierarchical model—often noted to resemble the hierarchical structure of the brain—is very effective. Essentially, a higher level in the hierarchy attempts to build a statistical model of the level below it. Taken together, the levels define a model of the original input data (see Fig. 2). This kind of hierarchical structure naturally allows the progressive generation of more complex features at higher levels, and progressively captures higher order correlations in the data. Furthermore, application of these bound maximization methods to such a model leads to relations defined between hierarchical levels that are reminiscent of the variety of neural connectivity observed in cortex: that is, feedforward, feedback, and recurrent (interlayer) connections are all essential.

The power of these methods for generating effective statistical models is impressive (Beal 1988). They have been applied to solve a number of standard pattern recognition problems, improving on other state-of-the-art methods (Hinton and Salakhutdinov 2006). However, there are two central issues regarding their application to biological

systems that remain important challenges. The first is the incorporation of time, and the second is an extension to motor control.

While some recent work has incorporated time (Brand and Hertzmann 2000; e.g., Taylor et al. 2007), there is no detailed, theoretically well-founded approach to adding temporal information to such statistical models that is biologically plausible. The Taylor et al. (2007) approach simply treats past times as additional fixed inputs to a two layer model. The Brand and Hertzmann (2000) work models motion as a Hidden Markov Model (HMM; i.e., discrete state transitions) with no attempt at biological plausibility. Our work has extended the bound maximization methods with a hierarchical model to include time, but made the unrealistic assumption that time steps are discrete and independent (Martens and Eliasmith 2007). It is thus reasonable to conclude that statistical models can be adapted to modeling temporal correlations, but current approaches are at early stages of development, especially in the context of biologically plausible constraints.

Less has been done to explicitly relate statistical models to motor control (although see Todorov 2006). As Todorov (2006) describes in detail, it is nevertheless natural for this kind of perceptual approach to extend to stochastic optimal control. Early on, Kalman (1960) showed that a simple optimal estimator, now known as the Kalman filter (KF), is a mathematical dual to the linear-quadratic regulator (LQR). Todorov (2006) has generalized this result to maximum a posteriori (MAP) smoothing and deterministic optimal control for continuous state systems (of which KF/LQR duality is a special case). In short, the best ways of interpreting incoming information via perception, are deeply the same as the best ways of controlling outgoing information via motor action. So the notion that there are a few, specifiable computational principles governing neural function seems plausible. In other words, given this very recent result, it seems clear that there is the enticingly close quantifiable relationship between perception and action that we would hope for. This recognition holds great promise as a means of constructing a general, unified theory of brain function. In sum, perceptual models are reasonably well-established theoretical approaches, motor control problems can be shown to be dual to those approaches, and more 'cognitive' functions (e.g., decision making) will be the result of the interface between the perceptual and motor models.

However, there are a wide variety of challenges faced by this view. As Todorov (2006) notes in his concluding section, important research directions that are left open by his theoretical result relating motor and perceptual models include: motor learning and adaptation, neural implementation of optimal control, and hierarchical/distributed control. It is interesting to note that the perceptual duals of two of these concerns have already been addressed by the statistical models I introduced earlier (as such perceptual models are both learned and hierarchical). What remains left open by both the motor and perceptual approaches to characterizing brain function that I have recommended here is *implementation*.

4 Functional integration

To this point, I have highlighted what I take to be promising answers to both the 'how' and the 'what' questions: the NEF captures *how* the brain computes; the statistical approach captures *what* the brain computes. Both the NEF and statistical approach are good candidates for supporting functionally integrated descriptions of neural systems because of their generality. The NEF generalizes across representation of mathematical objects (scalar, vectors, functions, etc.), kinds of computation (linear, non-linear), cell models (rectified linear, leaky-integrate and fire, conductance based, etc.), and kinds of dynamics (linear, time-varying, non-linear, etc.). The statistical approach generalizes across perceptual processes (object recognition, location estimation, multi-modal integration, etc.) and motor processes (path planning, feedback control, locomotion, target tracking, etc.). Given these considerations, I am willing to make the claim that between these two approaches, there is no obvious gap in our ability to answer, in principle, the 'how' and 'what' questions completely.

Nevertheless, given my suggestion (in Sect. 1) that answers to these two questions must be unified, there remains more to be said regarding how the two approaches interact. The broadest answer is simply that implementational constraints delimit possible function (which is why your desktop computer is not a truly universal Turing machine), and that functional specification is essential for realizing an implementation. So, in practice, the integration of 'how' and 'what' considerations is bound to be an iterative, bootstrapping process.

One example of this kind of integration is the utility of the NEF for model selection. One of the greatest challenges with any statistical modeling is model selection. Once a model is described, there are well-established and effective methods for parameter tuning and inference. However, defining the model itself, i.e., picking parameters, making distributional and independence assumptions, etc., has few systematic constraints. This challenge arises largely from the generality of the approach. Any set of relationships can be modeled statistically—but clearly the brain is tuned to picking up and acting on a particular set of relationships. In defining a model (e.g., picking the number of hierarchical levels, making assumptions about the forms of priors, etc.), we implicitly limit the relationships that can be described by the system. As a result, one good test of the plausibility of a particular model is determining whether or not it can be neurally implemented. If we define a model which demands more neurons than available in the brain, or demands a higher precision of representation than available, or demands a limitless memory, we cannot take the model to be a reasonable choice for characterizing neural function. In other words, the lack of constraints available in the statistical modeling approach can be supplemented with the systematic constraints on neural implementation identified by the NEF. A statistical way of thinking about this integration is that the NEF provides a prior on possible models to be considered. The three principles specify the form of 'reasonable' implementations of any statistical model the brain may construct of the world.

The NEF itself is constrained not only by functional specification, determined by the statistical model that is proposed, but also by available data. This allows for a bridging of the often large gap between detailed anatomical and physiological evidence and a high-level functional description. The NEF relies on information about tuning curves,

projections between neural populations, single cell dynamics, etc. when helping to specify a particular simulation. Integrating these approaches means that this information can also determine how a statistical model might be realized in neural tissue. Furthermore, high-level physiological data, like that available from fMRI and ERP, can be compared to activity generated by a large scale NEF simulation of a given set of brain areas (see e.g., Eliasmith 2004). In short, the NEF can serve as a conduit through which large-scale integrative functional hypotheses meet experimental evidence from a wide variety of neuroscientific methods.

Together, the NEF and statistical approach identify and integrate what are often referred to as 'top-down' (functional) and 'bottom-up' (neurophysiological) constraints. As a result, the generality of the methods allow for 'whole brain' modeling (more accurately, many system modeling). In fact, precisely this ability to support models that address a wide variety of neural function within a single model raise important challenges for the use of these combined approaches: in short, the price of generality is complexity. Consider, for instance, how we might model a task such as reaching for a moving object. To perform this task, the system must track (and hence predict) where the target will be given its current position. This entails extracting motion information. Motion information is available from a wide variety of stimuli, and hence the inclusion of motion information in the model needs to be extracted from a wide variety of stimuli. In other words, we need a fairly sophisticated visual system in the model—one which we construct by specifying a hierarchical statistical model (perhaps only a few levels are necessary) that is then tuned by many example stimuli. The representations available from this model then must be used to generate predictions, in a specific stimulus context, of how the object will move (thus implementing a state estimator). These predictions would then need to be used to determine a statistically optimal control signal to guide the (many-degree of freedom) motion of an arm, which must also be modeled by the system. Each of these aspects of the model could be implemented by determining the kinds and distribution of tuning curves evident in the relevant perceptual and motor areas, identifying appropriate single cell models, and specifying the necessary transformations to implement the needed mappings.

Were this simulation to be built successfully on a first attempt, it would be a significant advance over currently available simulations. The difficulty lies in the diagnosis and updating of the model in the much more likely case that a first attempt fails. That is, the ability to build large-scale, highly integrated models brings with it a great difficulty in 'debugging.' Unlike compartmentalized computer code, such a simulation is likely to need high-dimensional representations which unexpectedly interact (e.g., concurrent representation of current location and prediction of future location), and complex nonlinear transformations that are difficult to predict in a stochastic environment. Of course, this kind of challenge may be appropriate in a highly distributed, multi-functional neural system. As well, the NEF, with its capacity to describe the relation between various levels of representation (e.g., between single neuron and population-level representations) may go some way to making sense of the system for the purposes of debugging. Nevertheless, many of the well-known challenges of designing and debugging analog systems (Sarpeshkar 1998) become prominent with these kinds of models.

The combination of the NEF and statistical modeling approach is uniquely general, and able to directly connect with neurally relevant data. As a result, I have suggested that this marriage of methods is our best current approach for exploring more highly integrated, and larger scale neural models. But, having identified methods that can subserve functional integration brings with them a price: increased design challenges. This is a price that must be paid if we are to gain a deeper understanding of neural systems.

5 Conclusion

Clearly, the project of this paper—to identify a promising route to generating functionally integrated neural models—is only a first and tentative step towards a challenging research goal. Nevertheless it may be of interest to very briefly consider some of the many consequences of adopting this approach. First, the traditional notion of 'representation' does not naturally fit into this account. Instead, representations are 'deeply statistical' (i.e., representations are themselves statistical distributions).

As a result of shifting our understanding of representations in this way, our understanding of inference naturally shifts from logical inference to probabilistic inference as well. This is important for understanding how to design experiments that test hypotheses relying on these kinds of representations and transformations.

Furthermore, given such an integrated approach to modeling will likely demand more sophisticated experimental approaches—approaches that carefully intermix perceptual, cognitive, and motor aspects in their entirety. In short, such a view may help theoreticians get past modeling data, to modeling animals themselves. After all, animals don't control 'button presses,' but rather the complex, fluid, adaptive motion that leads to such a result.

It is beyond the scope of this paper to compare and contrast these three considerations with past paradigms in the behavioral sciences. Nevertheless, these brief suggestions may be enough to peek our curiosity sufficiently to revisit some of our assumptions about how best to describe brain function. After all, if we really want to build a brain, we had better be convinced of the utility of our basic principles.

Appendix

This appendix describes each of the three principles of the NEF quantitatively. For simplicity only the vector case is considered. Notably, function, scalar, and other representational forms are instances of vector representation—scalars being one-dimensional vectors, functions being representable as a vector of coefficients defined over an orthonormal basis (which itself does not need to be represented), and so on.

Neural representation

In the NEF, representation in neural populations is characterized in terms of a nonlinear encoding process and a linear decoding process (Eliasmith and Anderson 2003).

Encoding involves converting a quantity, $\mathbf{x}(t)$, from stimulus space into a spike train:

$$\sum_n \delta(t - t_{in}) = G_i \left[J_i(\mathbf{x}(t)) \right] \tag{1}$$

where $G_i [\cdot]$ ψ is the nonlinear function describing the spiking response model (e.g., leaky integrate-and-fire, Hodgkin-Huxley, or other conductance based models), J_i is the current in the soma of the cell, i indexes the neuron, and n indexes the spikes produced by the neuron. Specifically, the current is given by

$$J_i(\mathbf{x}) = \alpha_i \left\langle \tilde{\phi}_i \cdot \mathbf{x} \right\rangle + J_i^{bias} + \eta_i \tag{2}$$

where $J_i(\mathbf{x})$ is the input current to neuron i, $\mathbf{x}$ is the vector variable of the stimulus space encoded by the neuron, α_i is a gain factor, $\tilde{\phi}_i$ is the preferred direction vector of the neuron in the stimulus space, J_i^{bias} is a bias current that accounts for background activity, and η_i models neural noise. Notably, the dot product, $\left\langle \tilde{\phi}_i \cdot \mathbf{x} \right\rangle$, describes the relation between a high-dimensional physical quantity (e.g., a stimulus) and the resulting scalar signal describing the input current. In short, Eq. 1 captures the nonlinear encoding process from a high-dimensional variable, $\mathbf{x}$, to a one dimensional soma current, J_i, to a train of spikes, $\delta(t - t_{in})$.

To understand how a neural system might use the information encoded into a spike train in this manner, we must characterize a neurally plausible decoding as well. To do so we need to understand how this information can be converted from spike trains back into a relevant quantity in stimulus space. Note that this does not mean that the decoding process takes place explicitly in neurons. Rather, it is a theoretically useful means of characterizing part of the information processing characteristics of neurons. In the NEF we characterize decoding in terms of post-synaptic currents and decoding weights. Somewhat surprisingly, a plausible means of characterizing this decoding is as a *linear* transformation of the spike train. Specifically, we can estimate the original stimulus vector $\mathbf{x}(t)$ by decoding an estimate, $\hat{\mathbf{x}}(t)$, using a linear combination of filters, $h_i(t)$, weighted by decoding weights,ϕ_i:

$$\hat{\mathbf{x}}(t) = \sum_{in} \delta(t - t_{in}) * h_i(t) \phi_i = \sum_{in} h_i(t - t_{in}) \phi_i \tag{3}$$

where '$*$' indicates convolution. These h_i (t) are thus linear decoding filters which, for reasons of biological plausibility, we take to be the postsynaptic currents (PSCs) in the subsequent neuron.

To find the ϕ_i weights to determine this estimate, we minimize the mean-squared error,

$$\begin{aligned} E &= \frac{1}{2} \left\langle \left[\mathbf{x}(t) - \hat{\mathbf{x}}(t) \right]^2 \right\rangle_{\mathbf{x},t} \\ &= \frac{1}{2} \left\langle \left[\mathbf{x}(t) - \sum_{in} \left(h_i(t - t_{in}) + \eta_i \right) \phi_i \right]^2 \right\rangle_{\mathbf{x},t,\eta} \end{aligned} \tag{4}$$

where $\langle\cdot\rangle_{\mathbf{x}}$ denotes integration over the range of $\mathbf{x}$, and η_i models the expected noise. By optimizing with Gaussian random noise, we ensure that fine tuning is not a concern, since the decoding weights will be robust to fluctuations. For biological plausibility, this error is solved allowing the linear decoders to be PSCs, hence the minimization is done only over $\mathbf{x}$.

Defining a nonlinear encoding and a linear decoding (over both time and populations of neurons) provides a general means for capturing time-varying neural representation.

Neural computation

As stated in principle 2, neural computation is a special case of neural representation. As a result, we can modify (4) to find optimal linear decoders for a function of the stimulus space, rather than the stimulus space itself, i.e.,

$$E = \frac{1}{2}\left\langle \left[\mathbf{x}(t) - f(\hat{\mathbf{x}}(t))\right]^2 \right\rangle_{\mathbf{x},t}. \tag{5}$$

Solving this equation provides optimal decoders ϕ_i^f which give an estimate of that function, rather than an estimate of the variable itself as in the representation case. This implies that representation is a 'degenerate' computation where the function is merely identity. This approach has been shown to work well for both linear and nonlinear function computation (Eliasmith and Anderson 2003).

Neural dynamics

For generality, we can write the relevant dynamics of a population in control theoretic form, i.e., employing the dynamics state equation that comprises the foundation of modern control theory,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{6}$$

where $\mathbf{A}$ is the dynamics matrix, $\mathbf{B}$ is the input matrix, $\mathbf{u}(t)$ is the input or control vector, and $\mathbf{x}(t)$ is the state vector. In general, these matrices and vectors can describe a wide variety of linear, time-invariant physical systems. Here, we use (6) to capture the hypothesized high-level dynamics of a population of neurons.

Initially, this high-level characterization is divorced from neural-level, implementational considerations. However, it is possible to modify these matrices to render the system neurally plausible. First, we must account for intrinsic neural dynamics by converting this characterization into a neurally relevant one. To do so, we assume a model of PSCs given by $h(t) = \tau^{-1}e^{-t/\tau}$, and can then derive the following relation between (6) and a neurally plausible control theory:

$$\begin{aligned} \mathbf{A}' &= \tau\mathbf{A} + \mathbf{I} \\ \mathbf{B}' &= \tau\mathbf{B}. \end{aligned} \tag{7}$$

So our description of the high-level *neurally plausible* dynamics becomes

$$\mathbf{x}(t) = h(t) * \left[\mathbf{A}'\mathbf{x}(t) + \mathbf{B}'\mathbf{u}(t)\right]. \tag{8}$$

Notably, this transformation is general, and assumes nothing about the form of $\mathbf{A}$ or $\mathbf{B}$. So, given any behavioral system defined in the form of (6), it is possible to construct the neural counterpart by solving for $\mathbf{A}'$ and $\mathbf{B}'$. In fact, despite starting with linear time-invariant systems, these methods can successfully be employed to model a much broader class of dynamical systems. A variety of applications of this method to linear, nonlinear, and time-varying neural systems is described in Eliasmith (2005).

Next, we must incorporate this high-level description of the dynamics with our previous characterization of the neural representation. To do so we combine the dynamics of (8), the encoding of (1), and the population decoding of $\mathbf{x}$ and $\mathbf{u}$ from (3). That is, we take $\hat{\mathbf{x}} = \sum_{jn} h_j(t - t_{jn})\phi_j^{\mathbf{x}}$ and $\hat{\mathbf{u}} = \sum_{kn} h_k(t - t_{kn})\phi_k^{\mathbf{u}}$, which gives

$$\begin{aligned}
\sum_n \delta(t - t_{in}) &= G_i\left[\alpha_i \left\langle \tilde{\phi}_i \mathbf{x}(t)\right\rangle + J_i^{bias}\right] \\
&= G_i\left[\alpha_i \left\langle \tilde{\phi}_i \left[\mathbf{A}'\hat{\mathbf{x}}(t) + \mathbf{B}'\hat{\mathbf{u}}(t)\right]\right\rangle + J_i^{bias}\right] \\
&= G_i\left[\alpha_i \left\langle \tilde{\phi}_i \left[\mathbf{A}' \sum_{jn} h_j(t - t_{jn})\phi_j^{\mathbf{x}} + \mathbf{B}' \sum_{kn} h_k(t - t_{kn})\phi_k^{\mathbf{u}}\right]\right\rangle \right. \\
&\quad \left. + J_i^{bias}\right].
\end{aligned} \tag{9}$$

It is important to keep in mind that the temporal filtering is only done once (here included in the estimate of the signals), despite the fact that it is include in both (8) and (3). That is, $h(t)$ in these equations both defines the dynamics and defines the decoding of the representations. To put it in a more familiar form, this equation can be written as

$$\begin{aligned}
&G_i\left[\alpha_i \left\langle \tilde{\phi}_i \left[\mathbf{A}' \sum_{jn} h_j(t - t_{jn})\phi_j^{\mathbf{x}} + \mathbf{B}' \sum_{kn} h_k(t - t_{kn})\phi_k^{\mathbf{u}}\right]\right\rangle + J_i^{bias}\right] \\
&= G_i\left[\sum_{jn} \omega_{ij} h_j(t - t_{jn}) + \sum_{kn} \omega_{ik} h_k(t - t_{kn}) + J_i^{bias}\right]
\end{aligned} \tag{10}$$

where $\omega_{ij} = \alpha_i \left\langle \tilde{\phi}_i \mathbf{A}'\phi_j^{\mathbf{x}}\right\rangle$ and $\omega_{ik} = \alpha_i \left\langle \tilde{\phi}_i \mathbf{B}'\phi_k^{\mathbf{u}}\right\rangle$ are the recurrent and input connection weights respectively. These weights will now implement the dynamics defined by the control theoretic structure from (8) in a neurally plausible network.

Taken together, these three sections allow for the construction of large, spiking neural network models that implement a given (linear/nonlinear/time-varying) high

level hypothesis about the function of a neural system, through the time-dependent transformation of neural representations.

References

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks*. New York, NY: Cambridge University Press.

Beal, M. (1998). Variational algorithms for approximate Bayesian inference. Ph.D., University College London.

Bechtel, W., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press.

Brand, M., & Hertzmann, A. (2000). Style machines. In *Proceedings of SIGGRAPH*, pp. 183–192.

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.

Eliasmith, C. (2003). Neural engineering: Unraveling the comlexities of neural systems. *IEEE Canadian Review, 43*, 13–15.

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* (pp. 17–20). AAAI Press.

Eliasmith, C. (2005). A unified approach to building and controlling spiking attractor networks. *Neural Computation, 17*(6), 1276–1314.

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.

Fischer, B. (2005). A model of the computations leading to a representation of auditory space in the midbrain of the barn owl. Ph.D., Washington University in St. Louis.

Fischer, B. J., Pena, J. L., & Konishi, M. (in press). Emergence of multiplicative auditory responses in the midbrain of the barn owl. *Journal of Neurophysiology*.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks, 16*(9), 1325–1352.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504–507.

Hinton, G., & van Camp, D. (1993). Keeping neural networks simple by minimizing the description length of the weights. *ACM COLT '93*.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard Press.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *ASME Journal of Basic Engineering, 82*, 35–45.

Koulakov, A. A., Raghavachari, S., Kepecs, A., & Lisman, J. E. (2002). Model for a robust neural integrator. *Nature Neuroscience, 5*(8), 775–782.

Kuo, D., & Eliasmith, C. (2005). Integrating behavioral and neural data in a model of zebrafish network interaction. *Biological Cybernetics, 93*(3), 178–187.

Martens, J., & Eliasmith, C. (2007). A biologically realistic model of statistical inference applied to random dot motion. *COSYNE 2007,* Salt Lake City. 94.

McIntyre, J., Zago, M., Berthoz, A., & Lacquaniti, F. (2001). Does the brain model Newton's laws? *Nature Neuroscience, 4*(7), 693–694.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Rieke, F., Warland, D., de Ruyter van Steveninick, R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Sarpeshkar, R. (1998). Analog versus digital: Extrapolating from electronics to neurobiology. *Neural Computation, 10*, 1601–1638.

Sejnowski, T. J., & Rosenberg, C. R. (1986). NETtalk: A parallel network that learns to read aloud. *Cognitive Science Quarterly, 14*, 179–211.

Seung, H. S. (1996). How the brain keeps the eyes still. *National Academy of Science USA, Neurobiology, 93*, 13339–13344.

Singh, R., & Eliasmith, C. (2006). Higher-dimensional neurons explain the tuning and dynamics of working memory cells. *Journal of Neuroscience, 26*, 3667–3678.

Taylor, G. W., Hinton, G. E., & Roweis, S. (2007). Modeling human motion using binary latent variables. In B. Schòlkopf, J. C. Platt, & T. Hofmann (Eds.), *Advances in neural information processing systems 19*. Cambridge, MA: MIT Press.

Todorov, E. (2006). Optimal control theory. In K. Doya (Ed.), *Bayesian brain: probabilistic approaches to neural coding* (chapter 12, pp. 269–298). MIT Press.

Tripp, B., & Eliasmith, C. (2007). Neural populations can induce reliable postsynaptic currents without observable spike rate changes or precise spike timing. *Cerebral Cortex*, *17*, 1830–1840.

Wolpert, D. H. (2004). Information theory—the bridge connecting bounded rational game theory and statistical physics. In D. Braha & Y. Bar-Yam (Eds.), *Complex engineering systems*. Perseus Books.

Wolpert, D. M., Goodbody, S. J., & Husain, M. (1998). Maintaining internal representations: The role of the human superior parietal lobe. *Nature Neuroscience, 1*(6), 529–533.