This is a penultimate draft. To appear in P. Thagard (ed.), Philosophy of Psychology and Cognitive Science. Handbook of Philosophy of Science (Vol. 4). Amsterdam: Elsevier.

Computational neuroscience

Chris Eliasmith

Abstract: Taking 'computational neuroscience' to denote theoretical neuroscience, I describe accounts of representation, neural computation, and cognitive function consistent with recent advances in this field. During this discussion I provide an overview of neural coding and survey its possible implications for more traditional theories of the meaning of mental representations. Subsequently, I describe how ideas central to computational neuroscience (neural coding, neural computation, and dynamical systems/control theory) have been integrated to support descriptions of neurobiological systems at many different levels. Finally, I argue, and provide examples to show, that computational neuroscience is an essential ingredient for a complete picture of perceptual, motor, and cognitive function. Throughout the discussion I emphasize how computational neuroscience is important for tackling problems of interest to cognitive scientists.

Keywords: computational neuroscience, neural coding, brain function, neural modeling, cognitive modeling, computation, representation, neuroscience, neuropsychology, semantics, theoretical psychology, theoretical neuroscience

1 Overview

'Computational neuroscience' has come to denote two significantly different disciplines. On the one hand computational neuroscience is sometimes considered a part of 'bioinformatics' or 'computational biology,' a field that is concerned with the application of computers to the collection, organization, and analysis of biological data. Perhaps the best-known example of this kind of work is the recent human genome project, where the sequencing of genes was largely done automatically using computers. On the other hand, 'computational neuroscience' is taken to refer to what has also been called 'theoretical neuroscience'. In this sense, computational neuroscience is the application of theories relating to computation and information processing to neurobiological systems. This is the sense of computational neuroscience with which I will be concerned in the remainder of this article.

I take the distinction between computational and experimental neuroscience to be analogous to that between theoretical and experiment of physics. But, as the difference in terminology suggests, computational neuroscience relies heavily on computer simulation and modeling and so is not, perhaps, 'purely' theoretical. To date, this most often means that particular neural systems have been subjected to detailed mathematical analyses which are often tested through numerical simulations. For example, the tools of nonlinear systems have been applied to systems of coupled oscillators in an attempt to understand the neural mechanisms underlying lamprey and leech locomotion (Wilson 1999, ch. 13). Such systems are then often simulated to determine or demonstrate the effects of certain parameter regimes. These same sorts of mathematical tools are more often, and similarly, applied to a much lower level of description of neural systems: the characterization of the dynamics of action potential generation in single cells (Wilson 1999, ch. 5-10). In large part this is because there has been a great deal of success (ever

since Hodgkin and Huxley's (1952) famous mathematical description of the giant squid axon) at applying such tools to understanding the dynamics of ion conductance in single cells. More recently, researchers have begun to construct detailed, biologically plausible models of large systems of interconnected neurons. Building such complex networks often makes mathematical analysis intractable, demonstrating the indispensability of computers for simulating and analyzing these models. Much of this work can be seen as a direct descendant of connectionism.

As recently as 15 years ago, computational neuroscience was referred to as a "new and fledgling" discipline (Koch and Segev, xi). However, computational neuroscience has historical roots that reach as far back as those of connectionism, at least to the beginning of the last century. It was, nevertheless, about 15 years ago when computational neuroscience split from its connectionist roots. This was about the same time at which the field of 'neural computing' split from its connectionist roots as well. In fact, a similar split is evident in cognitive science, with artificial intelligence (AI) and cognitive psychology (which has largely kept the 'cognitive science' label) focusing on different research problems. In both cases, one branch (AI/neural computing) has become mainly concerned with technical applications of the theory, interested mainly in solving difficult engineering problems, while the other branch (cognitive science/computational neuroscience) has become mainly concerned with the application of the theory to understanding natural systems.

So, in contrast with neural computing, in computational neuroscience a premium is placed on biological realism. That is, models are not constructed merely to realize some function (however useful that function might be from an engineering perspective), but rather to realize a particular function *as it is realized* in some biological system. Computational neuroscientists are not interested in how one might implement working memory in a collection of identical, interconnected, computational nodes, but rather in how working memory is actually implemented by real, heterogeneous neurons in a living brain. It is already well-known that there are multiple ways in which collections of neuron-like units can be made to store memories. The pressing question for computational neuroscientists is: Which of those ways is the one relevant for neurobiology? So, the most obvious connection between computational neuroscience and cognitive science (is their shared interest in trying to come up with naturalistic explanations of the behaviour of complex natural systems.

Nevertheless, the central methodologies of these approaches are vastly different. Computational neuroscientists take as paradigmatic data from experiments in single cell physiology (e.g., intercellular or extracellular microelectrode recordings, dendritic patch clamps, multielectrodes) and neuroanatomy (e.g., cellular and tract staining, immunocytochemistry, tract tracing, deoxyglucose uptake). Cognitive scientists, in contrast, tend to consider data from a wide variety of multi-subject psychology studies (e.g., studies that measure reaction time, gross motor behaviour, linguistic behaviour, etc.). More recently, neuroscience-related techniques like functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and event related potentials (ERPs) have began to inform cognitive science. The biggest difference between the standard methodologies and cognitive science and computational neuroscience is one of scale. Computational neuroscientists have focused on the microscopic and cognitive

scientists have focused on the macroscopic. This has largely been because, until recently, the kind of mathematical descriptions provided by computational neuroscientists have not been "scaled up" to be informed by the data at the macroscopic level.

Because of these methodological differences, cognitive science and computational neuroscience also tend to focus on different kinds of behaviours as targets of explanation. Cognitive scientists, not surprisingly, focus on largely cognitive phenomena including: reasoning, object recognition, imagery, and language processing. Computational neuroscientists focus on more universally biological phenomena including: motor control, low-level perception (e.g., proprioception, retinal and early visual processing), and self-organization (neuron-level learning). As a result, implementational issues (i.e., issues like how many neurons do you have in working memory systems; what signal to noise ratio is supported by neural systems; how interconnected are the neurons; what are the time constants of dendritic potentials observed in working memory; etc.) have proven much more significant to computational neuroscientists.

From the perspective of a cognitive scientist, it might seem that a preoccupation with low-level implementational issues is a good way to waste your time just trying to get the irrelevant details right – details that have nothing to do with cognition (Fodor and Pylyshyn 1988). But perhaps computational neuroscientists agree with Mies van der Rohe who suggested that "God is in the details". That is, perhaps the designs discovered by mother nature are much more effective, robust, efficient, and so on, than those invented by a clever engineer. Perhaps a functional decomposition that is sensitive to the same design constraints as mother nature will look very different from one that is not. If so, time might be well spent looking at the details, no matter how devilishly obscure they may be. After all, we have proof that mother nature is able to construct sophisticated, robust, and extremely complex systems – systems exquisitely designed to adapt to rapidly changing, dynamic environments. Unfortunately, the same cannot be said of human designed systems. Similarly, we know that mother nature's design can scale up appropriately, we do not know that for our own designs. So, since we seem to be at the beginning of an era where the tools for reverse-engineering to brain are becoming available, it would be wise to exploit those tools.

Some exploitation of these tools has already taken place (indeed, information theory has been used since its inception to understand neural systems). However, most work in computational neuroscience has focused on improving our understanding of single cells. And, while the success here has been impressive – there are general methods for modeling both the electrical and morphological properties of individual neurons – the most remarkable behaviours demonstrated by natural neurobiological systems stem from the careful organization of billions of such cells. It is only when we begin considering large groups of neurons that we approach the point of contact between neuroscience and cognitive science. So, major challenges lie in trying to apply mathematical tools to neural systems of a sufficient size to tackle issues of interest to those concerned with cognitive function (Eliasmith 2003).

In this article I sketch some methodological, metaphysical, and conceptual issues that arise when we begin to consider the integration of higher-level (cognitive) and lower-level (neural) descriptions of mental phenomena. More specifically, I discuss accounts of representation, computation, dynamics, and cognition from a computational

neuroscience perspective. This discussion, I suggest, shows that computational neuroscience is indispensable for generating a complete naturalistic characterization of mental/cognitive function.

2 REPRESENTATION

The concept of a 'mental representation' is a central one in cognitive science. As a result, much time has been spent trying to better understand, analyze, and operationalize the notion of representation. Debates in cognitive science on what kinds of representations are used for what cognitive tasks (Kosslyn 1994), how to determine representational content (Churchland 1989; Fodor 1990; Fodor and Lepore 1992), and even on the relevance of representation to cognition (van Gelder 1998) are currently ongoing. As a result, it is essential to address what, if anything, computational neuroscience has to say about representation.

One useful distinction out of the philosophical discussion of representation is that between the contents and the vehicles of representations (Fodor 1981; Cummins 1989). The contents, or meanings, a representation refer to the semantic value of a representation for an animal. 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 syntactic structures, or physical realization of objects that play the role of representations (i.e., carrying a content) in a cognitive system. Currently, computational neuroscience can contribute most to improving our understanding of representational vehicles. Nevertheless, I discuss some of the consequences for our understanding of semantics as well.

Unfortunately, in neuroscience the concept of representation itself has received much less consideration than in the cognitive sciences, despite an equally wide usage.[1] As a result, the use of the term representation in neuroscience is often problematic. 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 and Van Essen 1991). This kind of experiment has been performed since Hubel and Wiesel's (1962) classic experiments in which they identified cortical cells selective to the orientation and size of a bar in a cat's visual field. The 'bug detector' experiments of Lettvin et al. (1988/1959), 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. 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.

[1] A search of PubMed for the term 'representation OR representations OR represent' returns well over 130000 hits. So far this year, in the Journal of Neuroscience alone, 8% of publications (42/543) mention representation in the title or abstract.

The difficulty is that such usage assumes that single neurons are the carriers of content, and that content can be determined by what has been called the 'naïve causal theory'. So, there are no principled means of determining what the representational vehicles are and how they might be related, and a naïve causal theory is well-known to be highly problematic (Dretske 1988).

Nevertheless, much of the work in *computational* neuroscience can help refine the notion of representation such that it is clearly relevant for understanding neurobiological systems and avoids the difficulties of these assumptions. This is partly because, unlike the typical treatment of representation in cognitive science, considerations of representation in computational neuroscience cannot avoid implementation constraints on representation. This proves to be a strength because such constraints highlight significant aspects of representation that have otherwise been missed (e.g. time, precision, etc.).

One of the most significant conceptual contributions of computational neuroscience to a neuroscientific understanding of representation is its 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 encoding some physical environmental variable into neural spike trains. By adopting an information theoretic view of representation, computational neuroscience holds that if we truly understand the encoding process, we must be able to demonstrate that we can decode that spike train to give us the originally encoded signal. As a result, to fully define representations, we must understand both encoding and decoding.[2]

As well, computational neuroscience has focused on two distinct 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 we should understand the contribution of a single cell's response to a complex representation over a large group of neurons. In the next two sections, I describe computational neuroscientific characterizations of encoding and decoding over time and neural populations.

2.1 Temporal representation

What is often considered one of the major flaws of classical cognitive science can simply not be ignored when considering neural systems: the importance of time. The extent and impact of this flaw for classical cognitive science has been vociferously argued by proponents of dynamicism (Port and van Gelder 1995; van Gelder 1998). They have suggested that both symbolic and connectionist models often completely ignore time constraints, or include them only after the fact. This, they suggest, is completely unrealistic. Real systems are significantly constrained by the highly dynamic environments in which they are embedded. To ignore time is to ignore one of the most salient driving forces behind impressiveness of cognitive systems, the limitations of cognitive systems, and the evolution of cognitive systems.

---

[2] Note that this does not demand that the decoding take place explicitly within an animal.

Indeed, neuroscientists, who are interested in exploring real world neural systems, have always been confronted with the essentially temporal nature of neurobiological systems. This is clearly reflected in much of the standard vocabulary of neuroscience: neuroscientists speak of spike times, firing rates, stimulus onset, persistent activity, adaptation, membrane time constants, and so on. So, unlike the often static considerations of representation of symbolicists and connectionists, neuroscientists have always considered time-varying representations. In this respect, any neuroscientific endeavour, be it experimental or theoretical, would be remiss if it ignored time. So it is no surprise that the representation of time-varying signals is central to computational neuroscience.

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 characterized with respect to a single cell. 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 and Huxley 1952), the use of cable equations to describe dendritic and axonal morphology (Rall 1957; Rall 1962), and the introduction of canonical models a large classes of neurons (Hoppensteadt and Izhikevich 2003) – supports a highly mechanistic understanding of encoding. While fully describing the encoding process also necessitates the identification of the particular parameters a neuron may be sensitive to (partially in virtue of its relation to other neurons in the brain), this can largely be inferred from the ubiquitous reporting of neuron tuning curves in a neuroscientific literature. So, improving our understanding of the encoding process is largely an empirical undertaking. 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 (Shadlen and Newsome 1994; Shadlen and Newsome 1995; Buracas et al. 1998). In contrast, timing code theories assume that information about the stimulus is carried by the distance between neighbouring spikes in the spike train generated by the neuron responding to the stimulus (Softky and Koch 1993; de ruyter van Steveninck et al. 1997; Rieke et al. 1997).

Many results in neuroscience are reported under the assumption that rate coding is an appropriate way to characterize temporal representation in neurons. For this reason, it is common to see firing rates, or spike histograms, of neurons over time. To generate these figures, neuroscientists usually determine the mean firing rate of a spike train over a relatively long, sliding time window (usually about 100 milliseconds). However, there are a number of problems with adopting such a view of neural representation. Perhaps the most obvious is that response times of many animals to a given stimulus are much quicker than 100 milliseconds. If the motor system was averaging over 100 milliseconds time window in order to determine what information was in the spike train arriving from sensory systems, such rapid responses would be impossible. In fact, there is a large volume of evidence that behavioural decisions are often made on the basis of one or two

neural spikes arriving a few milliseconds apart (see Rieke et al. 1997, pp. 55-63 for a review). As well, there is evidence that spike trains with exactly the same average firing rate but different placement of spikes within the hundred millisecond window produce very different output from a neuron they are presented to (Segundo et al. 1963). And finally, it has been demonstrated that rate codes cannot transmit information at the rates observed in neurons (Rieke et al. 1997), although the timing code can (MacKay and McCulloch 1952).

Given these difficulties, many computational neuroscientists prefer to understand temporal coding in neurons as dependent on the timing of individual spikes. A standard way to characterize this code is to take the inverse of the inter-spike intervals in a neuron spike train. So, when the inter-spike interval is short, the encoded signal was high, and when the inter-spike interval is long, the encoded signal was low. However, it would be misleading to say that it is the *precise* time of spikes that carries information about the stimulus in most such characterizations. There is good evidence that the precise timing of spikes is not mandatory for the successful transmission of signals with neurons (Bialek et al. 1991). In some ways, this is not a surprising result, since a code which was extremely sensitive to the timing of individual spikes would not be robust to noise.[3]

When we look carefully at the difference between typical rate codes and timing codes, we notice 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 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, as they are linear (i.e. rely only on weighted sums). Suppose we are trying to understand the representation of a particular neuron. To do so, we present it 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 mentioned, if we truly understand representational abilities 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. 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. Furthermore, under the assumption of linear decoding, the process of decoding is very simple. Essentially, every time a spike occurs, we place a copy of the decoder at the occurrence time of the spike. We 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, surprisingly, perhaps, 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).

---

[3] There seems to be a large variety of sources of noise in nervous systems (Eliasmith 2001).

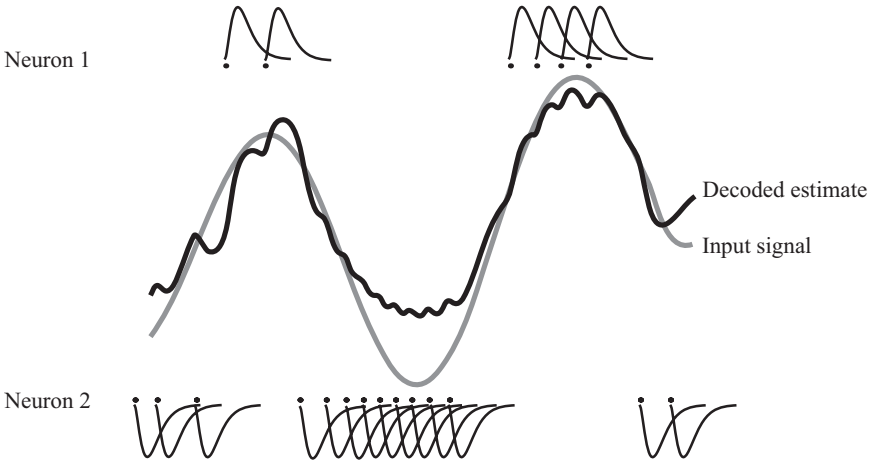

Figure 1: Temporal decoding. This diagram depicts linear decoding of a neural spike train (dots) using stereotypical decoders (skewed bell-shaped curves) on an input signal (grey line). The result of the decoding from two neurons (black line) is a reasonable estimate of the input signal. This estimate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

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. Recently, it has been suggested that the postsynaptic currents observable in the dendrites of receiving neurons can act as temporal decoders (Eliasmith and Anderson 2003, ch. 4). While the amount information lost increases under this assumption, it is biologically plausible (unlike the acausal optimal decoders described earlier), and increasing the number of neurons in the representation can make up for any information lost. An example of this kind of decoding for two neurons is shown in figure 2a. That example shows 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 any 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.

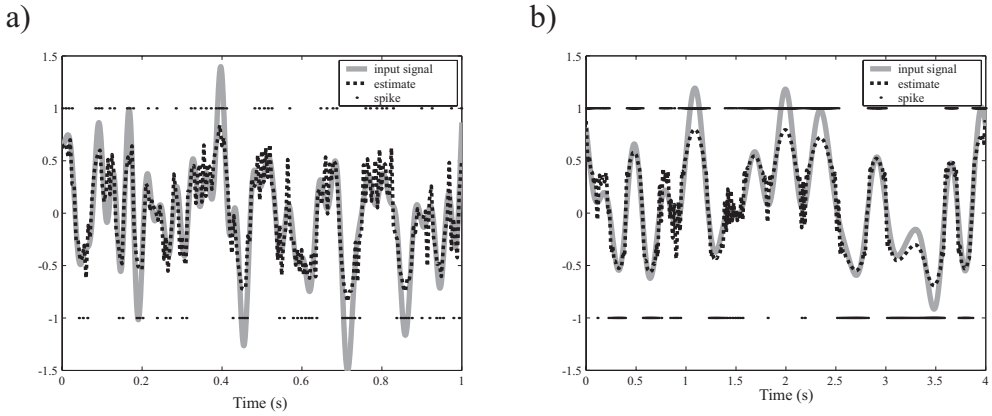

Figure 2: Biologically plausible timing and rate coding. a) A high frequency signal effectively decoded using postsynaptic currents (PSCs) as the decoders. This demonstrates a timing code. b) A low frequency

signal (note the difference in time scale) similarly decoded. This demonstrates a typical rate code. Together, these show that linear temporal decoding with PSCs is biologically plausible and continuous between rate and timing codes, depending on the signals.

This way of understanding temporal coding in neurons is both biologically plausible, and applicable to understanding the representation of signals and a wide variety of time scales. However, the signals being represented are extremely simple. All of these examples are time-varying scalar values. To support representations of sufficient complexity to handle cognitive functions, we need to understand how large groups of neurons can cooperate to effectively encode complex, real-world objects.

2.2 Population representation

In a well-known series of experiments, Ausonio Georgopoulos 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, he 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 reasonably well how this information was used by the actual motor system.

As a result, 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 (Eliasmith and Anderson 2003). Furthermore, it is easy to generalize this kind of understanding of neural representation to more complex mathematical objects. 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 (Schwartz 1994; Moran and Schwartz 1999; Todorov 2000).

However, scalar and vector representation alone is not sufficient for capturing a wide variety of neural responses. 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 and 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 and Glimcher 1997; Sereno and Maunsell 1998). And, although

this has not been directly tested in LIP, there is evidence that this kind of representation is used to encode information about the uncertainty of the estimate of the parameter being encoded (Britten and Newsome 1998). So, this kind of representation seems best understood as the representation of a function, not a scalar. The width of this function can be used to encode the uncertainty of the representation (wider functions indicate greater uncertainty). And, more complex functions like bimodal Gaussians, could be used to encode position and certainty information about multiple object simultaneously. So, it is natural to suggest that LIP is representing functions, not scalars or vectors.

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 data. An increasing number of neuroscientists and computational neuroscientists have been taking seriously the idea that neural systems must perform some kind of statistical inference in order to operate effectively in such an environment (Eliasmith and Anderson 2003; Kording and Wolpert 2004; Rao 2004; Deneve and Pouget in press). So, it is only to be expected that the representations in neural systems be able to support statistical inference. Clearly, the kind of function representation described for LIP can fulfill this role.

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). 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).

These examples demonstrate the wide variety of kinds of mathematical objects that can be represented in a neurobiologically plausible way. Nevertheless, a question that might remain for cognitive scientists is: is this kind of representation appropriate for understanding *cognitive* behaviour? That is, do we have reason to think that things like functions, vectors, and scalars are sufficient for explaining imagery, object recognition, and language use? After all, the standard kinds of representations used to explain these cognitive phenomena are things like pictures, graphs, and symbolic representations. I would like to suggest that the differences between these kinds of representations are terminological. That is, representations like images *just are* functions, or vectors. Take, for instance, the representation of an image on a computer. A computer represents an image as simply a long binary string. A long binary string is a vector. Similarly, if we consider a non-digitized version of a projected image, we can express the image as changes in light intensities that are a function of spatial position. That is, we can write down a mathematical expression which captures all of the information contained in the projected image. So, despite the particular vocabulary we might use for identifying classes of representations, they contain the same information about objects, spatial relations, color patterning, and so on. Similar kinds of examples can be generated for the representation of objects and language-like symbols (Eliasmith 2003).

So, there is nothing inherently limiting in understanding neurons as able to represent mathematical objects rather than the typical classes of representation identified by cognitive scientists. And, the characterization of a wide variety of representations as

variations on a theme (i.e. nonlinear encoding and linear decoding), rather than distinct classes, greatly unifies our understanding of representation in neurobiological systems.

In some ways, this discussion of representation has a great affinity for the discussion of distributed representation in connectionism. Clearly, each of these kinds of population representation are distributed representations. However, there are some important differences. For instance, connectionists only consider the case of vector representation. For another, they have no principled relation between the representations of individual nodes and 'higher-level' representations over a set of nodes (i.e., each node is an element of the represented vector for connectionists, but not for the distributed representation described above). For example, if we take a neural system to be representing images, we can define a representation of that image as a function (a 'higher-level' representation) embedded in a population of neurons (at the 'lower-level). Connectionists cannot do this. And, related to these two limitations, connectionists always try to learn representations rather than having a hypothesis about what representations are relevant for a particular task and then determining if that hypothesis is consistent with other (computational, dynamic, or intrinsic) properties of the population.

The constraints introduced by trying to learn representations can be very severe. It is well known that the order in which items are presented, the number of occurrences of items in the presentations, and the particular choice of which items are presented all greatly influence the results of learning. Furthermore, while these difficulties affect the typical three layer network, they become unmanageable for more complex kinds of networks. In contrast, understanding neural representation in terms of encoding and decoding avoids such difficulties. Very complex networks, with any number of layers, can have representations and transformations embedded in them using these methods. As well, since the implementation of the representations is a neural one, biologically plausible learning can be introduced at any point in the process, as necessary. So, while both connectionism and a computational neuroscience approach understand representations as distributed, only the latter provides principled methods for moving between levels of description of the representational capacities in a neurobiologically realistic network.

To this point, I have described both population representation and temporal representation independently. However because 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.

2.3 Semantics

As I mentioned earlier, the tools of computational neuroscience are best suited to characterizing representational vehicles rather than representational content. Nevertheless, the characterization of vehicles and contents is not independent (c.f. Cummins 1989): if we think that all vehicles can only support scalars, then only the kinds

of contents that can be carried by scalars can be contents of a system with those vehicles. In general, the ways in which we can characterize vehicles ought to tell us something about the kinds of contents those vehicles can carry.

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; Fodor 1990; Dretske 1995; 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 (Loar 1981; Harman 1982). 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. A common 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 (Field 1977; Block 1986). 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.

In the preceding 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 in (decoding), as well as how it is related to previous goings-on (encoding) are relevant for determining content. So given the characterization of vehicles I have presented, two-factor theories of content seem most plausible.

However, 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 (1992) 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*?" (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 computational neuroscientific characterization of representation presented above, there is a tight relation between the encoding and decoding processes. 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 (as I discuss below). 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. While there remains much to be said regarding the precise relation between encoding and decoding, such considerations suggest that this characterization can avoid the main weakness of past two-factor theories.

3 COMPUTATION

Conveniently, the above characterization of representation leads naturally to an understanding of computation in neural systems. Of course, without an account of computation, any characterization of representation is useless. The vast variety of complex and interesting behaviours observable in animals arise not from simply representing the environment, but performing complicated, and currently ill-understood, computations with these representations.

Before considering neural computation in more detail, it will be beneficial to address a terminological problem that often arises during such discussions. In the past, the notion of 'computation' has been closely allied to processing by serial digital computers. For this reason, many theorists take computational characterizations to be applicable only to machines that have discrete symbolic representations and process them via rules (van Gelder 1995). Unfortunately, this use of the term belies the existence of an entire field of research on analog computation (Uhr 1994; Douglas and Mahowald 1995; Hammerstrom 1995). The reason the strict notion of computation is often adopted, is because it is unclear whether there can be a principled distinction between computational and non-computational mechanisms if analog computation is considered to be a kind of computation. Nevertheless, I will adopt the widely accepted view in computational neuroscience at the notion 'computation' is applicable to neural systems in virtue of the kinds of descriptions we apply to them (Churchland et al. 1990): indeed, this area of research would not be called *computational* neuroscience otherwise.

This terminological issues aside, then, we can define neural computation – just as we did neural representation – in terms of a nonlinear encoding and a (different) linear decoding. Essentially, representation consists of trying to compute the identity function. That is, whatever is encoded into the neural spikes trains is what we are trying to decode when we take neurons to be representing their input. I refer to the decoders used for computing this function as 'representational decoders'. More generally, we can identify decoders for computing any function of the encoded input: I refer to these other decoders as 'transformational decoders'. So, 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 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; Eliasmith and Anderson 2003).

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.

4 DYNAMICS

Given the previous characterizations of representation and computation, it is possible to build neurally realistic circuits that take time time-varying signals as input and compute 'interesting' 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 to understand neurobiological function. As a result, computational neuroscientists need a means of characterizing the sophisticated, possibly recurrent, internal dynamics for the representations they take to be present in neural populations.

Eliasmith and Anderson (2003) suggest that neural dynamics can be best 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 1960; Kalman 1960). 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 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 be able 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). This allows any standard control theoretic description of a system to be written an equivalent 'neural' control theoretic form. This can prove a great benefit to theorists, as they can then draw from the vast resources of control theory when hypothesizing about which neural architectures may be able to realize some function: a function that may already be well-understood by control theorists.

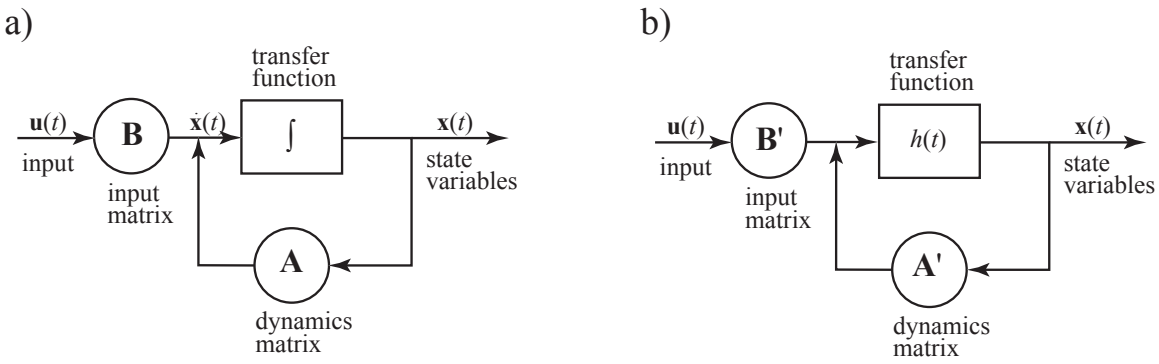

Figure 3: Diagram of the dynamics equation for LTI control theoretic descriptions of a) a standard physical system, $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$ and b) a neural system, $\mathbf{x}(t) = h(t) * \left(\mathbf{A}'\mathbf{x}(t) + \mathbf{B}'\mathbf{u}(t)\right)$. The input signal, $\mathbf{u}(t)$, can be modified by the parameters in the input matrix, **B**, before being added to any recurrent signal which is modified by the parameters in the dynamics matrix, **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, $h(t)$, **A**' and **B**' can be determined for any **A** and **B**.

Notably, adopting this description of the dynamics of neural systems is reminiscent of the 'dynamicist' view in cognitive science (Port and van Gelder 1995). Indeed, both characterizations share an emphasis on the importance of time, and both suppose that sets of differential equations are the best mathematical tools available for quantifying cognitive systems under this assumption. However, there is an extremely important difference between these two views. 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 and Townsend 1993) – and it is entirely unclear how it could be. In contrast, the approach described above 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. This makes the computational neuroscience approach more directly responsive to (and informative for) the experimental results generated by neuroscientists and cognitive neuropsychologists.

5 SYNTHESIS

The previous sections have defined three important principles 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 describe 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.[4]

[4] 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)

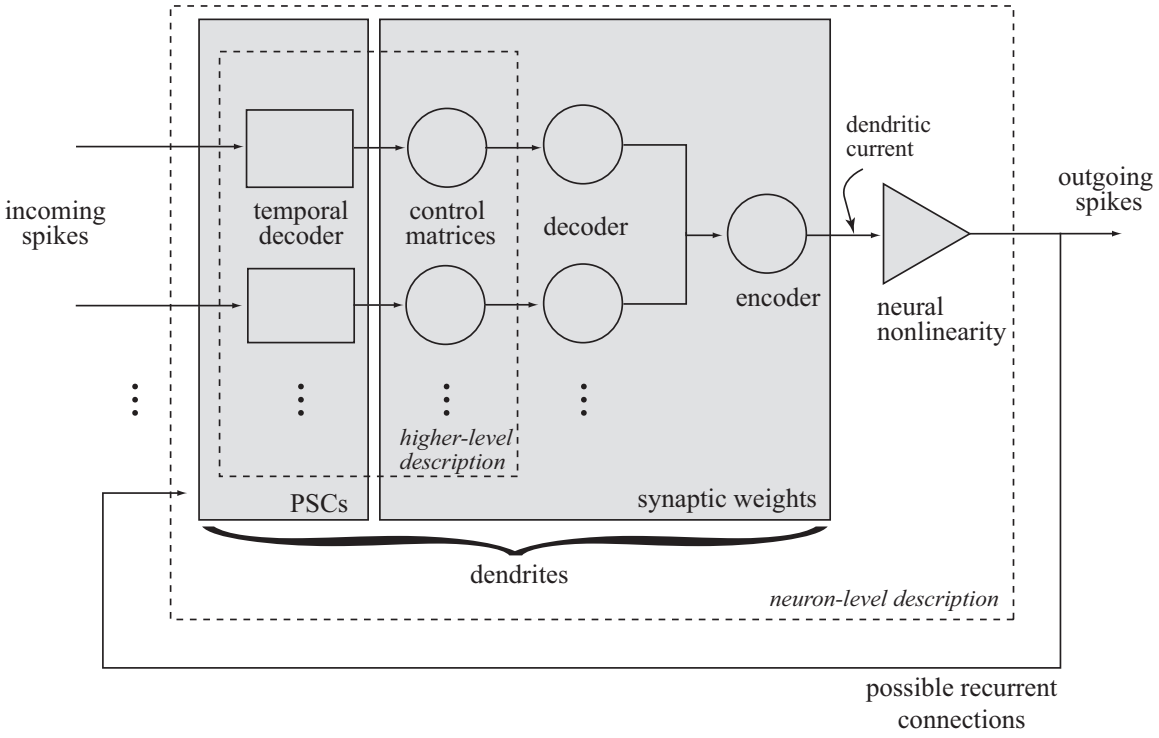

Figure 4: A generic neural subsystem (adapted from Eliasmith 2003). The outer dotted line encompasses elements of the neuron-level description, including PSCs, synaptic weights, and the neural nonlinearity in the soma. The inner dotted line encompasses elements of the control theoretic descriptions at 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.

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 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. 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. This, of course, is a typical means of justifying the introduction of theoretical entities in science.

This approach has been successfully applied to modeling the number of neurobiological systems, including sensory, cognitive, and motor systems (Eliasmith and Anderson 1999; Eliasmith and Anderson 2000; Nenadic et al. 2000; Eliasmith and Anderson 2001). Of greatest interest to cognitive scientists, however, is whether or not this kind of description can contribute to improving our understanding of high-level cognitive function.

6 COGNITION

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 (Lycan 1984; Fodor 1999). They have suggested that knowing where things happen in the brain, or how they are implemented in the brain, are not relevant for answering the more important question of what the cognitive architecture is. This, however, is no longer the widely-held view that it once was. The immense recent increase in cognitive research employing techniques that measure physiological signals in the brain (such as functional magnetic resonance imaging (fMRI), positron emission tomography (PET), event related potentials (ERPs), etc.), suggest that a more common view is that improving our understanding of neural organization will greatly benefit our understanding of the nature of cognitive function.

Despite the seeming prominence of this view, there are no well-established modeling techniques for bridging the gap between the functioning of single neurons and the functioning of the cognitive systems of which they are a part. 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.). In contrast, cognitive scientists are largely interested in models of cognitive functions like reasoning, memory, language, and so on. So, the question naturally arises as to whether the methods of computational neuroscientists can actually prove useful to cognitive scientists.

For some time now, computational neuroscientists have built sophisticated models of brain areas whose functions map naturally onto certain psychological categories. Perhaps the most successful example of such models are those related to short-term/working memory (Camperi and Wang 1998; Laing and Chow 2001; Brody et al. 2003). In general, these models have helped make the notion of working memory much more precise. Specific implementational constraints have been derived from some of these models, including the kinds of time constants and receptors that are likely necessary to support working memory function. Preliminary results regarding the source of working memory limitations (Trappenberg 2003), and the cost of representational complexity (Eliasmith and Anderson 2003) in such networks have also been generated. Such results have vastly improved our understanding of how working memory might be organized, its strengths, expected limitations, and so on. Such insights should inform our functional decomposition of other high-level cognitive categories. So, in this instance, computational neuroscience serves as a useful means of filling in important details for improving our understanding of a specific cognitive function, and may serve as a springboard for better understanding other cognitive functions.

However, working memory itself does not seem to be particularly representative of what counts as truly cognitive behaviour, since it is found in a wide variety of nonhuman animals. As well, the available models of working memory tend not to be systemic. That is, they are focused on small, highly-constrained neural areas, and do not have complex dynamics or support sophisticated representations. So, it could well be

argued that such examples do not demonstrate that the methods described earlier will be able to tackle interesting problems in cognitive science.

Recently, however, significantly more sophisticated models of clearly high-level cognitive function have been built. As an example, I will consider a model of the Wason card selection task (Wason 1966). This task requires the manipulation of language-like structures and is performed only by humans. Thus it is, if anything is, a high-level cognitive phenomenon. A considerable amount of research in psychology has focussed on exploring this task (Griggs and Cox 1982; Reich and Ruth 1982; Cheng and Holyoak 1985; Cosmides 1989; Gigerenzer and Hug 1992; Chater and Oaksford 1996; Liberman and Klar 1996; Fiddick et al. 2000). The task itself consists of subjects being presented with a conditional statement such as "if P then Q," which they must test for validity given some data. So, for example, suppose that a subject is presented with a rule such as "if a card has a vowel on one side, then it has an even number on the other". They are then give a data set that consists of four cards laying on a table with one of 'A', 'B', '2', or '3' written on the visible side of each card. The subject is then asked to choose the card or cards that must be turned over in order to test this conditional. The correct choice in this example are the 'A' and '3' cards. However, this combination is usually chosen only by a small minority of subjects (5-40%). More often, subjects will choose the 'A' card or the 'A' and '2' cards (80-90%).

In an interesting variation on this task, the rules chosen were made significantly more concrete, or familiar. For instance, the researcher might use the rule, "if someone drinks alcohol then that person is over 18," with the data being analogous to that presented earlier. This shift in the rule's content massively increased success on the card selection task. Much of the discussion regarding the Wason card selection task focuses on attempting to explain the differences between this 'concrete' task and the original 'abstract' task.

The neural-level model generated using the techniques described earlier demonstrates how a specific hypothesis regarding these differences can be shown to be neurally plausible. Specifically, the model shows that (Cheng and Holyoak 1985; Cosmides 1989; Gigerenzer and Hug 1992) domain general, context-sensitive inference can account for the differences in performance in human subjects on these tasks in a manner consistent with known neuroanatomy and neurophysiology. Some highlights of the model are that: 1) it is a large-scale model, consisting of nine separate, interconnected populations; 2) it performs structure-sensitive processing (it uses one of the available vector binding operations (Plate 1994; Plate 1997), for encoding the rules into language-like structures and transforming the rules appropriately); 3) it performs context sensitive processing (i.e., different transformations are applied depending on whether the rule is concrete or abstract); 4) it learns which transformations to perform (i.e., based on reinforcement or the provision of the correct answer, it updates the appropriate transformation in the current context to determine which transformation is appropriate for which context); 5) the model is mapped to known neuroanatomy; and 6) these functions are carried out using populations of spiking neurons that encode complex representations (100 dimensional vectors).

In sum, this model has both cognitive and neural plausibility: that is, it explains cognitive behaviour and does so in a neurally realistic way. Unlike past attempts at

understanding human performance on this task, it can be used to make specific predictions regarding physiological measures of people performing these tasks, as well as behavioural predictions. Also unlike other suggestions, it has been shown that the hypothesized dynamics, representations, and computations can actually be carried out in neural hardware. More generally, this model serves to demonstrate that we can, indeed, begin to bridge the gap between neural and psychological explanations of cognitive systems.

Given such examples, I suspect that the methods of computational neuroscience will not only prove useful, but will be essential for giving a comprehensive explanation of how cognitive systems operate. It is not only the success of such models, but also our changing understanding of the nature of cognitive systems, that will render such accounts indispensable. Let me consider two recent shifts in our understanding of cognition.

First, it has been suggested by a number of psychologists and philosophers, perhaps Gibson (1955) being the most famous, that the distinction between perception/action and cognition may be a misleading one. They suggest that in order to understand the complicated behaviours we see arising from neurobiological systems – no matter how complicated – we must first understand the tight connection between perception and action. Indeed, there has been a significant amount of recent work in dealing with what is often called "embedded" or "embodied" cognition (Varela et al. 1991; Haugeland 1993; Hutchins 1995; Clark 1997; Kelly 2001). This is the view that in order to understand cognition, we must understand an organism as placed within its environment. These theorists thus take the organism, its actions, and its environment as tightly linked: linked through the perceptual/motor loop of the organism itself. For these theorists, thinking of cognitive behaviours as somehow specially internal, or separable from action and perception, is misguided. Related experimental work suggests that much of our cognitive life is thus tightly tied to more 'basic' perceptual and action-oriented processes (Barsalou 1999; Richardson et al. 2001; Matlock in press). Thus, if we can not easily distinguish cognitive processes from non-cognitive ones, then work in computational neuroscience on simple motor and perceptual systems is highly relevant for understanding cognitive function.

Second, not only has the distinction between cognition and perception/action come under scrutiny in recent years, but so has the distinction between cognition and emotion (Damasio 1994). There is ample neurological evidence that people with damage to emotionally related brain structures (e.g. ventromedial prefrontal cortex) become partly cognitively incompetent. While many cognitive functions are preserved in such patients, tasks such as decision-making, risk assessment and planning become severely impaired (Bechara et al. 2000; Spinella 2003). As a result, the traditional cognitivist view that emotions inhibit our ability to be rational seems implausible. From a computational perspective, this suggests that distinguishing cognitive systems from other systems may not allow us to properly explain what we typically take to be cognitive functions.

Both of these challenges to the idea that cognition is somehow separable from other neural functions suggest the need for a method that is continuous between typically cognitive systems and supposedly non-cognitive ones. On the face of it, this need should not be surprising since all such systems fundamentally use spiking neural cells to perform their functions. As a result, if we have methods for embedding high-level functions into

complex networks of neural cells, we can use those methods to consistently characterize what are intuitively different high-level functions (e.g., cognitive versus non-cognitive). This consistency thus makes it possible to interface models of traditionally non-cognitive systems with those of more cognitive brain systems. As a result, the methods of computational neuroscience, unlike other methods, make it possible to unify a wide variety of neural functions – a unification that seems essential if we are to understand the whole system – so such methods are essential to constructing good explanations in cognitive science.

7 REFERENCES

Andersen, R. A., G. K. Essick, et al. (1985). "The encoding of spatial location by posterior parietal neurons." Science **230**: 456-458.

Barsalou, L. (1999). "Perceptual symbol systems." Behavioral and Brain Sciences **22**: 577-609.

Bechara, A., H. Damasio, et al. (2000). "Emotion, decision making and the orbitofrontal cortex." Cerebral cortex **10**(3): 295-307.

Bialek, W., F. Rieke, et al. (1991). "Reading a neural code." Science **252**: 1854-57.

Block, N. (1986). Advertisement for a semantics for psychology. Midwest Studies in Philosophy. P. French, T. Uehling and H. Wettstein. Minneapolis, University of Minnesota Press. **X:** 615-678.

Britten, K. and W. Newsome (1998). "Tuning bandwidths for near-threshold stimuli in area MT." Journal of Neurophysiology **80**(2): 762-770.

Brody, C. D., R. Romo, et al. (2003). "Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations." Current Opinion in Neurobiology **13**: 204-211.

Buracas, G. T., A. M. Zador, et al. (1998). "Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex." Neuron(20): 959-969.

Busemeyer, J. R. and J. T. Townsend (1993). "Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment." Psychological Review **100**(3): 432-459.

Camperi, M. and X. J. Wang (1998). "A model of visuospatial working memory in prefrontal cortex: recurrent network and cellular bistability." Journal of Computational Neuroscience **5**: 383-405.

Chater, N. and M. Oaksford (1996). "Rational explanation of the selection task." Psychological review **103**: 381-391.

Cheng, P. W. and K. J. Holyoak (1985). Pragmatic reasoning schemas. Cognitive Psychology.

Churchland, P. (1989). A neurocomputational perspective. Cambridge, MA, MIT Press.

Churchland, P., C. Koch, et al. (1990). What is computational neuroscience? Computational Neuroscience. E. Schwartz. Cambridge, M.A., MIT Press.

Clark, A. (1997). Being there: Putting brain, body and world together again. Cambridge, MA, MIT Press.

Cosmides, L. (1989). "The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task." Cognition **31**: 187-276.

Cummins, R. (1989). Meaning and mental representation. Cambridge, MA, MIT Press.

Damasio, A. R. (1994). Descartes' error: Emotion, reason, and the human brain. New York, NY, Grosset/Putnam.

de ruyter van Steveninck, R. R., G. D. Lewen, et al. (1997). "Reproducibility and variability in neural spike trains." Science **275**: 1805-1808.

Deneve, S. and A. Pouget (in press). "Bayesian multisensory and cross-modal spatial links." Journal of Neurophysiology.

Desimone, R. (1991). "Face-selective cells in the temporal cortex of monkeys." Journal of Cognitive Neuroscience **3**: 1-8.

Douglas, R. and M. Mahowald (1995). Silicon neurons. The handbook for brain theory and neural networks. M. Arbib. Cambridge, MA, MIT Press.

Dretske, F. (1981). Knowledge and the flow of information. Cambridge, MA, MIT Press.

Dretske, F. (1988). Explaining behavior. Cambridge, MA, MIT Press.

Dretske, F. (1995). Naturalizing the Mind. Cambridge, MA, MIT Press.

Eliasmith, C. (1997). "Computation and dynamical models of mind." Minds and Machines **7**: 531-541.

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. and C. H. Anderson (1999). "Developing and applying a toolkit from a general neurocomputational framework." Neurocomputing **26**: 1013-1018.

Eliasmith, C. and C. H. Anderson (2000). "Rethinking central pattern generators: A general framework." Neurocomputing **32**: 735-740.

Eliasmith, C. and C. H. Anderson (2001). "Beyond bumps: Spiking networks that store smooth n-dimensional functions." Neurocomputting **38**: 581-586.

Eliasmith, C. and C. H. Anderson (2003). Neural engineering: Computation, representation and dynamics in neurobiological systems. Cambridge, MA, MIT Press.

Felleman, D. J. and D. C. Van Essen (1991). "Distributed hierarchical processing in primate visual cortex." Cerebral Cortex **1**: 1-47.

Fiddick, L., L. Cosmides, et al. (2000). "No interpretation without representation: the role of domain-specific representations and inferences in the Wason selection task." Cognition **77**: 1-79.

Field, H. (1977). "Logic, meaning, and conceptual role." Journal of Philosophy **74**: 379-409.

Fodor, J. (1981). Representations. Cambridge, MA, MIT Press.

Fodor, J. (1990). A theory of content and other essays. Cambridge, MA, MIT Press.

Fodor, J. (1998). Concepts: Where cognitive science went wrong. New York, Oxford University Press.

Fodor, J. (1999). "Diary." London Review of Books **21**(19).

Fodor, J. and E. Lepore (1992). Holism: A shopper's guide. Oxford, UK, Basil Blackwell.

Fodor, J. and Z. Pylyshyn (1988). "Connectionism and cognitive architecture: A critical analysis." Cognition **28**: 3-71.

Georgopoulos, A. P., J. T. Lurito, et al. (1989). "Mental rotation of the neuronal population vector." Science **243**: 234-236.

Georgopoulos, A. P., A. B. Schwartz, et al. (1986). "Neuronal population coding of movement direction." Science **243**(1416-19).

Gibson, J. J. and E. J. Gibson (1955). "Perceptual learning: Differentiation or enrichment ?" Psychological Review **62**: 324-341.

Gigerenzer, G. and K. Hug (1992). "Domain specific reasoning: Social contracts, cheating and prospective change." Cognition **43**: 127-171.

Griggs, R. and R. Cox (1982). "The elusive thematic material effect in Wason's selection task." British Journal of Psychology **73**: 407-420.

Hammerstrom, D. (1995). Digital VLSI for neural networks. The handbook of brain theory and neural networks. M. Arbib. Cambridge, MA, MIT Press.

Harman, G. (1982). "Conceptual role semantics." Notre Dame Journal of Formal Logic **23**: 242-56.

Haugeland, J. (1993). Mind embedded and embodied. Mind and Cognition: An International Symposium, Taipei, Taiwan, Academia Sinica.

Hodgkin, A. L. and A. Huxley (1952). "A quantitative description of membrane current and its application to conduction and excitation in nerves." Journal of Physiology (London) **117**: 500-544.

Hoppensteadt, F. and E. Izhikevich (2003). Canonical neural models. The handbook of brain theory and neural networks. M. Arbib. Cambridge, MA, MIT Press.

Hubel, D. and T. Wiesel (1962). "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." Journal of Physiology (London) **160**: 106-154.

Hutchins, E. (1995). Cognition in the wild. Cambridge, M.A., MIT Press.

Kalman, R. E. (1960). "Contributions to the theory of optimal control." Boletin de la Sociedada Matematica Mexicana **5**: 102-119.

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

Kelly, S. D. (2001). "Demonstrative concepts and experience." Philosophical Review **110**(3): 397-420.

Kording, K. P. and D. M. Wolpert (2004). "Bayesian integration in sensorimotor learning." Nature **427**: 244-247.

Kosslyn, S. (1994). Image and brain: The resolution of the imagery debate. Cambridge, MA, The MIT Press.

Laing, C. R. and C. C. Chow (2001). "Stationary bumps in networks of spiking neurons." Neural Computation **13**: 1473-1494.

Lettvin, J., H. Maturana, et al. (1988/1959). What the frog's eye tells the frog's brain. Embodiments of mind. W. McCulloch. Cambridge, MA, MIT Press.

Liberman, N. and Y. Klar (1996). "Hypothesis testing in Wason's selection task: social exchange cheating detection or task understanding." Cognition **58**: 127-156.

Loar, B. (1981). Mind and meaning. London, UK, Cambridge University Press.

Lycan, W. (1984). Logical form in natural language. Cambridge, MA, MIT Press.

MacKay, D. and W. S. McCulloch (1952). "The limiting information capacity of a neuronal link." Bulletin of Mathematical Biophysics **14**: 127-135.

Matlock, T. (in press). "Fictive motion as cognitive simulation." Memory and Cognition.

Moran, D. W. and A. B. Schwartz (1999). "Motor cortical representation of speed and direction during reaching." Journal of Neurophysiology **82**: 2676-2692.

Nenadic, Z., C. H. Anderson, et al. (2000). Control of arm movement using a population of neurons. Computational neuroscience: Trends in research 2000. J. Bower, Elsevier Press.

Plate, T. (1997). A common framework for distributed representation schemes for computational structure. Connectionist systems for knowledge representation and deduction. F. Maire, R. Hayward and J. Diederich. Brisbane, AU, Queensland University of Technology**:** 15-34.

Plate, T. A. (1994). Distributed representations and nested compositional structure. Computer Science. Toronto, University of Toronto.

Platt, M. L. and G. W. Glimcher (1997). "Responses of intraparietal neurons to saccadic targets and visual distractors." Journal of Neurophysiology **78**: 1574-1589.

Platt, M. L. and G. W. Glimcher (1998). "Response fields of intraparietal neurons quantified with multiple saccadic targets." Experimental Brain Research **121**: 65-75.

Port, R. and T. van Gelder, Eds. (1995). Mind as motion: Explorations in the dynamics of cognition. Cambridge, MA, MIT Press.

Rall, W. (1957). "Membrane time constant of motoneurons." Science **126**: 454.

Rall, W. (1962). "Theory of physiological properties of dendrites." Annual New York Academy of Science **96**: 1071-1092.

Rao, R. (2004). "Bayesian computation in recurrent neural circuits." Neural Computation **16**.

Reich, S. S. and P. Ruth (1982). "Wason's selection task: Verification, falsification and matching." British Journal of Psychology **73**: 395-405.

Richardson, D., M. Spivey, et al. (2001). Language is spatial: Experimental evidence for image schemas of concrete and abstract verbs. Twenty-third Meeting of the Cognitive Science Society.

Rieke, F., D. Warland, et al. (1997). Spikes: Exploring the neural code. Cambridge, MA, MIT Press.

Schwartz, A. B. (1994). "Direct cortical representation of drawing." Science **265**: 540-542.

Segundo, J. P., G. P. Moore, et al. (1963). "Sensitivity of neurones in Aplysia to temporal pattern of arriving impulses." Journal of Experimental Biology **40**: 643-667.

Sereno, A. B. and J. H. R. Maunsell (1998). "Shape selectivity in primate lateral intraparietal cortex." Nature **395**: 500-503.

Shadlen, M. and W. Newsome (1994). "Noise, neural codes and cortical organization." Current Opinion in Neurobiology **4**: 569-579.

Shadlen, M. and W. Newsome (1995). "Is there a signal in the noise?" Current Opinion in Neurobiology **5**: 248-250.

Softky, W. and C. Koch (1993). "The highly irregular firing of cortical cells is inconsistent with the temporal integration of random EPSPs." Journal of Neuroscience **13**: 334-350.

Spinella, M. (2003). "Evolutionary mismatch, neural reward circuits, and pathological gambling." International Journal of Neuroscience **113**(4): 503-512.

Todorov, E. (2000). "Direct cortical control of muscle activation in voluntary arm movements: A model." Nature Neuroscience **3**: 391-398.

Trappenberg, T. P. (2003). "Why is our capacity of working memory so large?" Neural Information Processing-Letters and Reviews **1**(3): 97-101.

Uhr, L. (1994). Digital and analog microcircuit and sub-net structures for connectionist networks. Artificial intelligence and neural networks: Steps toward principled integration. V. Honavar and L. Uhr. Boston, MA, Academic Press**:** 341-370.

van Gelder, T. (1995). "What might cognition be, if not computation?" The Journal of Philosophy **XCI**(7): 345-381.

van Gelder, T. (1998). "The dynamical hypothesis in cognitive science." Behavioral and Brain Sciences **21**(5): 615-665.

Varela, F., E. Thompson, et al. (1991). The embodied mind: Cognitive science and human experience. Cambridge, MA, MIT Press.

Wason, P. C. (1966). Reasoning. New horizons in psychology. B. M. Foss. Harmondsworth, Penguin.

Wilson, H. (1999). Spikes decisions and actions: Dynamical foundations of neuroscience. Oxford, UK, Oxford University Press.