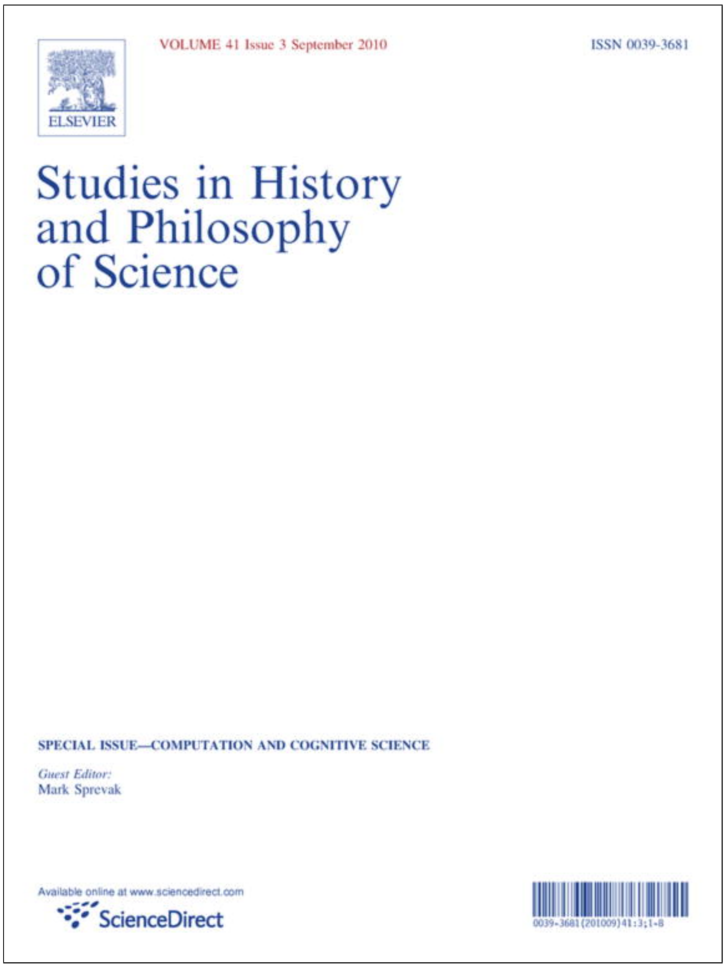

VOLUME 41 Issue 3 September 2010

ISSN 0039-3681

ELSEVIER

Studies in History and Philosophy of Science

SPECIAL ISSUE—COMPUTATION AND COGNITIVE SCIENCE

Guest Editor:
Mark Sprevak

Available online at www.sciencedirect.com
ScienceDirect

0039-3681(201009)41:3;1-8

Studies in History and Philosophy of Science 41 (2010) 313–320

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

How we ought to describe computation in the brain

Chris Eliasmith

Department of Philosophy, Department of Systems Design Engineering, University of Waterloo, Ontario, N2L 3G1, Canada

ARTICLE INFO

Keywords:
Neural computation
Control theory
Neural representation
Cognitive architecture
Cognitive models
Theoretical neuroscience

ABSTRACT

I argue that of the four kinds of quantitative description relevant for understanding brain function, a control theoretic approach is most appealing. This argument proceeds by comparing computational, dynamical, statistical and control theoretic approaches, and identifying criteria for a good description of brain function. These criteria include providing useful decompositions, simple state mappings, and the ability to account for variability. The criteria are justified by their importance in providing unified accounts of multi-level mechanisms that support intervention. Evaluation of the four kinds of description with respect to these criteria supports the claim that control theoretic characterizations of brain function are the kind of quantitative description we ought to provide.

1. Introduction

This essay is structured such that each heading is a specific claim related to quantitative descriptions of brain function. Any subheadings under a given heading are intended to provide additional considerations or details in support of the heading. While this does not provide for typical, smooth, reading of the paper, it serves to make the argument clearer and can shorten reading time, as the content of any "obviously true" heading can be skipped.

The word 'computation' is used in a liberal and definitional sense. I am using the liberal sense in the title (the sense typical of cognitive science usage, which means something like a 'transformation of representations'). However, I am using the definitional sense, from computational theory (i.e. Turing Machine equivalence) in the remainder of the essay. I will generally replace 'computation in the brain' in the first sense with 'a quantitative description of brain function' for clarity.

In brief, the argument I present here is:

1. There are four relevant kinds of quantitative description of brain function: computational, dynamical, statistical, and control theoretic
2. We ought to provide the best quantitative description of brain function
3. A good description of brain function provides for simple state mappings, and useful decompositions that account for variability
4. A good description in the brain sciences explains by positing mechanisms that support interventions
5. Computation theoretic descriptions do not meet these criteria well
6. Conclusion 1: therefore, computation theoretic descriptions are not good descriptions (from 3–5)
7. Control theoretic descriptions meet these criteria better than any of the other alternatives
8. Therefore, control theoretic descriptions are the best descriptions (from 1, 7)
9. Conclusion 2: therefore, control theoretic descriptions are the kind of quantitative description we ought to provide (from 2, 8)

One clarification is important: conclusion 2 does not rule out the other descriptions as useful. Rather, it suggests that other descriptions are essentially heuristics for temporarily stating the description. That is, ultimately, other descriptions should be

E-mail address: celiasmith@uwaterloo.ca

0039-3681/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsa.2010.07.001

translated into a unifying description of brain function stated with control theoretic constructs.

2. There are four kinds of quantitative description of brain function

I begin with some considerations regarding how quantitative descriptions relate to physical systems in general, and then turn to which quantitative descriptions are relevant for understanding brain function.

2.1. Different quantitative descriptions are better for different classes of phenomena

I do not worry about how quantitative descriptions are individuated (i.e. why statistical descriptions are different from dynamical descriptions).

2.1.1. Physical systems can have multiple quantitative descriptions

In most cases, what we identify as a physical system (e.g. a gas, a computer chip) can be described using different quantitative descriptions (e.g. statistical or Newtonian mechanics, computational theory or circuit theory). If we are trying to argue for *the best* description of some physical system, we must have a means of picking between these possible descriptions.

2.1.2. Quantitative descriptions have a natural class of physical phenomena that they describe

Notably, many descriptions are of the same mathematical class (e.g. both computational and circuit descriptions are algebraic), so it is not their mathematical properties that distinguish them. Instead, it is the mapping between the mathematics and the physical world that classifies the different kinds of quantitative descriptions. So, in circuit theory, variables are measurable properties like resistance, current, and voltage, whereas in computational theory variables are easily distinguished system states, like low/high voltage, or open/closed (mechanical) gates.

In essence, this is why such descriptions are quantitative descriptions *of* something: there is a defined mapping from the description to physical states. Mappings are natural (i.e. simple, straightforward, easy for us to understand) for the class of phenomena that they are explicitly defined over (and to the extent those definitions are specific). For instance, circuit descriptions are natural over the class of voltages, currents, and so on—they are neither overly specific (i.e. picking out material properties) nor overly abstract (i.e. picking out non-electrical properties like fluid flow).

These considerations result in the unsurprising conclusion that quantitative descriptions are natural for the class of physical systems that they are explicitly defined to be descriptions of.

2.1.3. Quantitative descriptions are implementation independent, but to differing degrees

As is again evident from the computation versus circuit descriptions, some quantitative descriptions (e.g. circuit theory) apply only to a subclass of others (e.g. computational theory). As a result, computational theory is more implementation independent than circuit theory. Notice also that circuit theory is independent of many specific material properties of potential circuit elements, for which chemical descriptions may be most natural.

2.1.4. The goodness of a description varies depending on the phenomena of interest

I have more to say on what constitutes a good description in Section 3. These considerations can be preliminary given an agreed characterization of goodness.

If the agreed notion of goodness is partly psychological (e.g. relies on simplicity), and the natural class for a description is too (e.g. also relying on simplicity), then the goodness of a description will vary depending on the natural class of phenomena in question. A description will be best for the phenomena which fall most directly in its natural class.

Just to be clear, this principle does not result in unbridled relativism: so long as we have a consistent measure of goodness across all phenomena, there will be one description which is best for a given class.

2.2. There are four kinds of quantitative description relevant to brain function

Here, I briefly describe each approach, indicate the class of systems it is most natural for, and describe its type of implementation independence.

2.2.1. Computational

Computational descriptions adopt computational theory, which characterizes systems using Turing languages. Such languages are able to describe any Turing Machine (TM) computable function. I take this to have historically been the dominant approach in cognitive science.

2.2.1.1. The natural physical phenomena for computational descriptions are those that are easily discretizable. What I have called Turing languages assume a mapping between the description in the language and distinct physical states. The paradigm case of this is the high/low voltages of silicon transistors mapped to 1s and 0s in the description. In general, any physical system that has easily distinguished (i.e. discrete in both space and time) states can be well-described by such languages. Often such systems are engineered.

2.2.1.2. Computational descriptions are highly implementation independent. Turing Machines are a powerful computational description precisely because they are completely implementation independent. Much has been made of this by functionalists in cognitive science. Notably, this independence means that certainty of the state value is generally assumed (i.e. that it is either 1 or 0). In short, randomness or noise is typically ignored.

2.2.2. Dynamical

Dynamical systems theory, as a mathematical theory, is extremely general (and arguably equivalent to control theory). However, in the context of cognitive systems, a number of researchers have championed the 'dynamical systems theory of mind', which I refer to as DST. DST uses the mathematical theory but adds additional assumptions when applying it to cognitive systems. Given the equivalence between the mathematical theory of dynamical systems and control descriptions, I will discuss DST unless otherwise noted.

2.2.2.1. The natural physical phenomena for DST dynamical descriptions are simple phenomena governed by physical laws. Simplicity is a stated assumption of DST theorists in cognitive science: van Gelder & Port (1995) argue that DST theorists must 'provide a *low-dimensional* model that provides a scientifically tractable description of the same qualitative dynamics as is exhibited by the high-dimensional system (the brain)' (ibid., p. 35). This constraint of low-dimensionality is a severe one, and limits the complexity of such descriptions to simple systems. However, such systems, being continuous, are strictly speaking more computationally powerful than TMs.

2.2.2.2. DST dynamical descriptions are implementation independent. In DST, the low-dimensional descriptions are implementation independent, because they rely on 'lumped parameters'. Such parameters are high-level, non-physical parameters necessary to match dynamics, generally uninformed by implementational constraints.

2.2.3. Statistical

Statistical descriptions describe the probability of various measurable states of the system given other known states of the system. Such models usually have as their central goal the prediction of data.

2.2.3.1. The natural physical phenomena for statistical descriptions are complex phenomena with unknown mechanisms. Complex systems, in virtue of their complexity, often have many unknown or undescribed interactions between system components. As a result, known initial conditions often map to a wide variety of subsequent states. Statistical models are ideal for describing systems of this kind when prediction is of the utmost importance (e.g. in data analysis). Notably, this method, which avoids explicitly positing mechanisms, typically has the cost of making novel interventions difficult to predict.

2.2.3.2. Statistical descriptions are highly implementation independent. Statistical models focus on describing the regularities in the data and hence are silent with respect to the particular physical implementation. In essence, these descriptions would not change if the implementation changes and statistical properties do not. Another way of describing this feature of statistical models is by noting that the model is often highly specific to a given data set. This is consistent with implementation independence because although implementation and the values of measured system states (data) are usually tightly coupled, they do not *have* to be.

2.2.4. Control theoretic

Control theoretic descriptions describe the dynamics of a system through its state space. Usually, the notions of 'controller' and 'plant' are used to describe the system.

2.2.4.1. The natural physical phenomena for control theoretic descriptions are those with directed dynamics. Because of the distinction between plants and controllers, control theoretic descriptions typically apply to systems in which one part of the system directs the dynamics of another part of the system. Control theory uses the general tools of mathematical dynamical systems theory.

2.2.4.2. Control theoretic descriptions vary between implementation independent and implementation specific. Standard (i.e. general mathematical) dynamical analyses are performed in a manner which tends to remove the physical uniqueness of the problem (e.g. through non-dimensionalization, and using normal form analysis). In this respect, many such descriptions are intentionally implementation independent. Nevertheless, they are so by design, not by the nature of the description. The original equations, which are often parametrically tied to specific physical instantiations (e.g. a mechanical circuit, or a circuit in silicon, etc.), can also be used as a system description. In such cases, the description is highly implementation specific.

Thus, control theory can describe implementation independent controllers, while at the same time being able to describe particular implementations of those controllers in a given medium.

2.3. Some of these quantitative descriptions are strictly equivalent

2.3.1. Control theoretic descriptions are equivalent to dynamical and statistical descriptions

This is no more than the mathematical observation that all of these approaches employ methods defined over the reals, and have no evident restrictions on the functions that they can compute in that domain (the restriction of unity integrals on statistical descriptions is not serious as it still leaves an uncountable number of functions). This makes all of these descriptions strictly more powerful than TMs.

2.3.2. Computational descriptions are strictly weaker than the other options

TM languages are strictly weaker than those defined over the continuum (Siegelmann, 1995). Finite state automata (FSAs) are weaker still.

2.3.3. Brain function can be described by any of these candidates

As I have argued in detail elsewhere (Eliasmith, 2000), given the ubiquitous presence of noise in the brain, only a finite amount of information can be passed between the outside world and brain states, or between brain states. As a result, TMs are sufficient for describing the information processing properties of the brain. In fact, FSAs are also sufficient, as the difference between FSAs and TMs is that TMs have infinite resources (tape and time). Brains clearly do not share that luxury. So, FSAs can describe all brain-computable functions.

Since FSAs can describe all brain-computable functions, and since all of the considered descriptions are strictly more powerful than FSAs, all of the considered descriptions can describe brain function. Hence, brain function is the kind of phenomena that has multiple descriptions (see Sect. 2.1.1), so *we must turn to other criteria to determine which is the best.*

2.4. There are no other relevant candidates

There are no other candidates for two reasons: (1) possible candidates are equivalent to those that have been discussed; or (2) possible candidates have not been shown to be generally useful in the description of cognitive systems. Examples of the first type are most logics (equivalent to computational descriptions) and quantum theory (equivalent to continuous descriptions). Examples of the second type include quantum theory (despite the supposed theories of consciousness, no quantum explanations of cognitive system function for even simple tasks have been offered) and hybrid descriptions. I am not aware of serious, non-interim, hybrid descriptions of cognitive systems. In general, hybrid descriptions are temporary because they violate the assumption that a general description should be unified. Detailed discussion of the justification of placing weight on unification is beyond the scope of this paper. Suffice to note that if this is not assumed, a set of non-arbitrary rules for determining when to use which (sub)description of the hybrid must be offered. I assume throughout that unification is a defining feature of a good description.

A related concern is that perhaps different descriptions are applicable at different 'levels' of explanation. So while no one theory is hybrid, our over-arching theory of cognition will involve very different kinds of characterization, depending on the explanation of interest. There are a number of ways to allay this concern. First, any commitment to a multiple levels view must be clear on what we mean by 'levels', and also systematically determine when one should switch descriptive levels. My suspicion is that such an approach is unlikely to meet with success. A more satisfying approach is to determine how a unified description can *relate* these intuitive levels to one another. Second, it is perfectly reasonable to introduce simplifying assumptions within a theory if those assumptions are taken to be appropriate in a particular kind of explanation. This is what happens, perhaps, when we employ Newtonian mechanics, with the knowledge that it is not as accurate as Einsteinian mechanics. Nevertheless, such simplification is justified only insofar as we know how to relate the simplified and the more accurate descriptions. I take it that an analogous constraint should apply to cognitive theories. This is something that I do not consider in detail here, but I am confident that the cited examples suggest that control theory is amenable to this kind of incorporation of simplifying assumptions across a wide variety of 'levels'.

It is worth repeating that these considerations do not rule out interesting descriptions that are not control theoretic. Perhaps, for instance, a computational decomposition at a high-level gets something essentially right about cognitive systems. The point is that such isolated successes are not sufficient for adopting that mode of description in general. Such descriptions should be replaced, if possible, by descriptions that do not isolate their successes, but apply widely. Much of the argument below suggests that control theoretic descriptions are in the best position to realize that possibility—and the specific examples suggest such control theoretic descriptions, to some extent, already have.

3. A good description of brain function provides for simple state mappings and decompositions, and accounts for variability

Because each of the considered quantitative descriptions is both general and sufficiently powerful enough to describe brain function, we must adjudicate their applicability by other criteria. Here I argue for several criteria that constitute a good description of brain function. This is not intended to be 'good in all possible senses' (see Sect. 4.3), but rather 'good for successful cognitive scientific theories'.

3.1. A good description provides a simple mapping from data to description states

In order for a description to be useful, it must be practical to map between the empirical data and the states identified by the description. The simpler such a mapping is, the better (because more practical) the description becomes. The mapping is more practical because it is evident how new information can be integrated with, or challenges, the currently accepted description.

If it is difficult, or impossible, to determine how new evidence bears on a quantitative theoretical description, then that quantitative theoretical description itself is deficient. This property, of course, is to be evaluated relative to the merits of rival quantitative theories.

3.2. A good description provides a clear decomposition of the system

Ideally, a quantitative description should act as a guide to decomposing the system. Because of the complexity of neural systems, decomposition is an essential explanatory strategy (Bechtel & Richardson, 1993). The more specific and effective the decomposition for explanatory progress, the better the description.

3.3. A good description accounts for variability

If the type of system to which the description applies is a highly variable class (as is the case of neural systems), then descriptions able to explicitly incorporate this variability will be better than those that do not. Characterizing the precise form and effect of the variability is crucial in the case of complex systems, which generally have significant amounts of unexplained (sometimes unexplainable) behaviour. In addition, the precise nature of the variability can be highly sensitive to implementational constraints. Thus, descriptions sensitive to implementation are often better able to explain the relevant variability than those that are not.

4. A good description explains by positing mechanisms that support interventions

A good description is one that satisfies scientific goals. In cognitive and brain science, I take those goals to include explanation, prediction, and identification of mechanisms in order to reproduce and intervene in the complex behaviours of neurobiological systems. A good descriptive strategy will be applicable at many levels of grain and be able to relate (i.e. unify) the relevant levels.

4.1. Cognitive science aims at explaining and predicting behaviour

Cognitive science has a focus on explaining the underpinnings of behaviour. While the appropriate level at which such a description needs to be given has been a matter of much debate, the aim itself has not been. As a result, any good description must be both explanatory and predictive of behaviour.

4.2. Explanatory means mechanistic

In the case of cognitive and brain sciences, useful explanations are those that appeal to subpersonal mechanisms. This is because it is precisely such explanations which provide a basis for both intervention in behaviour and the artificial reproduction of those behaviours. These mechanisms must be specific enough to allow for intervention. That is, the mechanisms must be specified in a way that relates to the measurable and manipulable properties of the system.

4.3. There are other definitions of 'good'

Sections 3 and 4 are the antecedent of a conditional. That is, if we take good descriptions to be of this nature, then we ought to employ control theoretical descriptions of brain function. As such, there is no need to defend this as the best or only definition of 'good'; and I do not intend to. Like any argument, the conclusion is sound insofar as the antecedent is taken to be true. Hopefully however, this definition of 'good' is plain enough to be generally acceptable.

5. Computational descriptions do not satisfy these criteria well

Here I evaluate computational descriptions with respect to the previous criteria for goodness of quantitative description of brain function. I suggest that computational descriptions are not especially good. This is only relevant if there is a better description. In section 6, I argue that control theoretic descriptions are better.

5.1. Mappings from data to computational states is complex

5.1.1. Single cell models are more simply described as dynamical systems

In short, this is because the brain does not functionally discretize well. The earliest attempts to suggest possible discretizations include the McCulloch & Pitts (1943) model of the single cell. Their mapping between logic gates and neurons was not intended to be physiologically plausible, and it clearly is not. There have not been other serious attempts to do so.

Perhaps the reason why is that, even if a state table were available for a neuron, it would not be informative as to the nature of the biological mechanisms that underpin that table. While the information transfer characteristics of neurons suggests that about 1–3 bits of information are transferred per action potential (Rieke et al., 1997), the relationship between input and output bits is not naturally described by a model with discrete states and logic-like transitions. Instead, the simplest neuron models take the form of dynamical descriptions (which can be 'translated' into computational ones, but become much more complex in so doing). These descriptions have variables and parameters that map directly onto the physical properties of single cells, such as cell membrane capacitance, membrane resistance, and ion flux.

5.1.2. Neural methods do not provide easily discretized data sets

When we turn to kinds of experimental data other than single cell spike trains, be it measured electrical properties of individual cells, or large portions of cortex (EEG, fMRI, MEG), or observable motor behaviour, the problem seems worse. All of these types of data are generally analyzed as continuous signals; discretizations are simply not apparent. For instance, EEG and similar methods of measuring brain function are analyzed as continuous signals using spectral and temporal decompositions of various kinds.

Another candidate for discretized states is linguistic behaviour. There are two problems with this candidate: (1) language has many 'continuous' kinds of phenomena in addition to words (which are finite), captured by prosody, pragmatics, and so on; (2) descriptions cast at the linguistic level do not provide the kinds of mechanistic descriptions demanded from useful explanations in cognitive science. Many such explanations seem to demand reference to 'sub-personal', non-linguistic, states, and this takes us outside the domain of linguistic behaviour.

In both cases, lack of apparent natural discretizations makes for lack of apparently TM-like state transitions. Hence, the underlying mechanisms are unlikely to be compactly described by TMs.

5.2. Computational decompositions are not applicable to brains

5.2.1. Computational architectures do not provide useful decompositions

Computational descriptions would be useful if they imply a particular, good, way to decompose the system. TM theory provides the distinction between a tape (input/output) and the transition table, but this not useful for decomposing brain function. We must turn to other possible computational architectures for such suggestions.

The most widespread computational decomposition is the von Neumann architecture. However, this architecture assumes that programs, describing the function of the system, are treated identically to the data on which such programs operate. As a result, such programs can be moved from memory to the CPU and back again. Brains do not share this flexibility. Memory and programs/function are tightly intermixed, as in a typical connectionist network. Despite some early attempts to map a von Neumann-like architecture to psychological descriptions of cognitive function (Atkinson & Shiffrin, 1968), the mapping has not proven useful. The cortex is not divisible into 'memory' and 'processor' as von Neumann architectures are.

It has been suggested that brains are parallel computers. However, parallel architectures exploit the same flexible memory usage, and so suffer from the same inability to map simply to brains.

5.3. Computational descriptions do not account for variability

Computational theory was developed in the context of ideal, non-stochastic, state transitions and easily identifiable states. Digital computers are carefully engineered to respect these assumptions, and this makes their behaviour predictable and repeatable. While there are recent developments that address the effects of stochasticity on computable functions, and it has been shown that this does not affect the computational power of the system, such extensions to TMs have not informed the construction of computers. As a result, typical computational descriptions do not account for variability in the systems they describe.

When describing real physical systems, variability—in short, noise—is inescapable. Brains are no exception to this rule. The implementation independent nature of computational descriptions should make it unsurprising that they tend to be insensitive to implementational issues like noise.

5.4. Computational descriptions do not satisfy the criteria

Given the previous considerations, it should be clear that computational descriptions do not satisfy the criteria for being good descriptions. Computational descriptions do not provide useful mechanistic explanations and predictions of neurobiological behaviour. This is because computational descriptions do not identify appropriate kinds of mechanisms to support intervention, which is a consequence of their failure to meet the criteria described in Sections 5.1–5.3. That is, if a description fails to (1) help decompose the system and (2) capture data through simple mechanisms (relative to its competitors), then that description cannot be used for prediction and intervention. Hence, it is not good (or, more accurately, not as good as its competitors).

I should note that past successes of computational descriptions (e.g. ACT-R, SOAR, etc.) do not belie this conclusion. The claim at issue is that models relying on computational descriptions cannot provide the *unity* to the descriptions of cognitive phenomena that are ultimately of interest. Computational descriptions identify *some* mechanisms and interventions, but their descriptive assumptions do not capture the broad class of mechanisms and interventions of interest to cognitive scientists in general.

5.5. Aside: brains are 'computers' in some ways

Notably, it would be misleading to say that computational descriptions do not apply to brains, full stop—and this is not what I claim. To clarify this point, in this section I show how brains *fall under computational descriptions* without affecting my conclusion that such descriptions are not the best quantitative models for cognitive science.

5.5.1. Brains have TM descriptions

Given previous considerations regarding noise, it is reasonable to claim that there is some TM description of brain function. Furthermore, like all finite physical implementations of TMs, brains will not be universal TMs. They will only be as computationally powerful as FSAs.

5.5.2. This result is uninteresting both theoretically and practically

5.5.2.1. Theoretically, because of Kolmogorov. Kolmogorov has shown that two implementations of a given TM cannot be usefully considered equivalent unless they are almost identical (or unless one can assume infinite strings). As a result, identifying a TM that is implemented by the brain does not tell you how to reproduce the described function in another implementational setting (Le Cun & Denker, 1992).

5.5.2.2. Practically, because such descriptions are too easy. As Searle has pointed out at some length (Searle, 1990), TM descriptions of physical systems are ubiquitous (he suggests Microsoft Word is implemented by a wall). Searle's point is a little misleading: it is typically very difficult to figure out how to map states from a TM description of Word to microphysical states of a wall in the appropriate way. However, when we do not know *which* function is being computed, as in the case of the brain, we do not have any useful constraints on how to construct the TM description (i.e. we do not have a TM description ready that we have to map to the brain). As a result, it becomes extremely easy to come up with some mapping or other. We have no reason to believe that such mappings are good, relevant, or in any way interesting.

5.5.3. Continuity is irrelevant to the goodness of quantitative descriptions

A number of authors have suggested that continuity is feature of brains that fundamentally distinguishes them from Turing Machines (see e.g. Churchland, 1995; van Gelder, 1998; Piccinini, 2008). In fact, it has been shown that analog computers are theoretically more powerful that TMs (Siegelmann, 1995). However, this result relies on such a computer having complete access to the real number line. Real world machines, however, do not have such access if there is any expectation of computationally irrelevant disturbance (i.e. noise, no matter how small).

In summary, there is no use arguing over whether or not there is some TM description of brain function—there is. However, what we are really interested in is whether it is a *good* description.

6. Control theoretic descriptions are good descriptions

Here, I argue that control theoretic descriptions are good descriptions of brain function at many scales. Specifically, I consider descriptions of single neurons and networks of single neurons, up to and including those responsible for linguistic behaviour.

6.1. Control descriptions provide simple mappings from data to control theory states

By far the best mechanistic description we currently have of single neural cells is as non-linear electrical circuits. The circuits are naturally described by non-linear systems theory, the main mathematical tool of control theory. As a result, control theoretic states are widely accepted as the simplest, most powerful, descriptions of single neuron behaviour.

As we compose single cells into larger networks, it is useful to adopt the language of representation and computation. Eliasmith & Anderson (2003) propose the Neural Engineering Framework (NEF), a detailed theory of how neural systems can be understood. I will not review here the three central principles of this approach, but I will note the following: the third principle provides a direct mapping from the single cell data collected by neurophysiologists to control theory. This mapping consists of a nonlinear encoding, determined directly from the data, and linear decoding that is optimal and mapped directly to the neurophysiology. This low-level neurophysiological mapping allows for prediction of single cell and aggregate data. In short, the mapping is simple between many kinds of neural data and control theoretic states.

These methods have been successfully applied in a wide variety of models, including the barn owl auditory system (Fischer, 2005), the rodent navigation system (Conklin & Eliasmith, 2005), escape and swimming control in zebrafish (Kuo & Eliasmith, 2005), working memory systems (Singh & Eliasmith, 2006), the translational vestibular ocular reflex in monkeys (Eliasmith et al., 2002), and context sensitive linguistic inference (Eliasmith, 2005). This variety suggests the mapping is a useful one for positing and testing general neural mechanisms.

6.2. Control descriptions provide a useful decomposition

6.2.1. Control descriptions distinguish plants and controllers

The central decomposition employed by control descriptions is between a controller and a plant. While both are described by dynamical systems theory, the controller is taken to be a part of the system that varies the input to a plant in order to achieve a desired state (provided to the controller).

6.2.2. Motor and perceptual systems decompose well as controllers and plants

Peripheral neural motor systems act like the controller for the body as plant. That is, peripheral motor systems determine the details of muscle contractions given higher level specifications of motor actions. More precisely, there is evidence for a hierarchy of such interactions in the motor system (Grafton & Hamilton, 2007). Thus, this decomposition maps in a straightforward way onto our current understanding of motor control. In addition, in closed-loop control, controllers are assumed to have sensors that feedback the state of the plant, allowing the controller to be more sophisticated. This fits well with the role of the many perceptual systems found in the brain. These systems can be naturally thought of as similarly organized (though dual) to the motor hierarchy (Todorov, 2007).

6.2.3. Neural systems are appropriately described as (hierarchical) directed dynamical systems

As a result, the control theory method for decomposing systems is useful for understanding the kinds of hierarchies observed in the brain. A nested control theoretic description of plant dynamics directed by feedback controllers, at least in broad outline, seems appropriate to describing neural function. Furthermore, this decomposition does justice to the massively interconnected nature of perceptual and motor systems. We cannot yet be certain of the most appropriate decomposition of neural systems, but preliminary evidence at least suggests that ideas from control theory may help provide just such a decomposition.

6.3. Control descriptions incorporate variability

Control theory was developed to describe physical systems. As a result, it accommodates noise: optimizing controllers in the face of noise is a long-standing part of control theory. Both the analytic and synthetic aspects of control theory naturally deal with variability. This suggests that such descriptions are appropriate for noisy systems like the brain. Notably, the NEF has noise as a core concern, and has been used to quantify in detail the relationship between noise (and other variability) and neural properties (Eliasmith & Anderson, 2003).

7. Control theoretic descriptions are the best quantitative descriptions

This section is dedicated to a brief comparison of control theoretic descriptions with the rival candidates.

7.1. Comparison to computational descriptions

Compared to computational descriptions, control theoretic descriptions unify our description of phenomena of interest to cognitive science. For instance, working memory and navigation are typical ‘cognitive’ phenomena. Locomotion and reflexes are more typically ‘sensory-motor’ phenomena. Control theory applies to both. And, to allay concerns that such descriptions do not apply to ‘higher’ cognitive phenomena, the BioSLIE model presented in

Eliasmith (2005) proves the utility of control theoretic descriptions to higher linguistic and inference tasks. The BioSLIE model demonstrates syntactic generalization, and makes predictions regarding learning history and response times in a cognitive task (the Wason card task). The details of this approach are beyond the scope of the present discussion (see Stewart & Eliasmith, in press, for a description).

Thus, while computational descriptions perform poorly as neuron level descriptions, control theoretic descriptions do well at both the neuron level and the cognitive level. It is not simply the case that control theory as discussed here is more specific than rival computational descriptions. For instance, consider Newell's SOAR architecture, which is specific, but clearly suffers from the same decomposition difficulties. One does not need to go into the details of SOAR's computational description to see why the decomposition is a poor one. The distinction between program and memory and the lack of a systematic relationship to neural hardware already suggest (despite the additional specificity) that SOAR will fail *given our stated criteria*. This is not to deny the successes of SOAR, rather it is to suggest that, in the long run, those successes will not withstand deeper theoretical failings.

In sum, control theoretic descriptions more effectively meet the criteria for good quantitative descriptions of neural systems than computational descriptions.

These differences arise in large part due to the fact that computational theory is designed to be implementation independent, whereas control theory is designed to be implementation sensitive. The physical systems that computational theory best applies to are carefully engineered. The physical systems analyzed using control theory need not be. The brain, of course, falls into this latter category. Hence, it should not be surprising that control theory is in a better position to describe neural mechanisms in a manner useful to cognitive science than the descriptions offered by computational theory.

7.2. Comparison to dynamical descriptions

7.2.1. DST dynamics are divorced from implementation

A consequence of the DST insistence on the use of lumped parameters is that such models become extremely difficult to confirm or disconfirm in light of the vast majority of neural data. There is no standard way to map lumped parameters to physically manipulable parameters of the system (usually only observable behaviour is mapped to the model). As a result, there are no constraints on what might or might not be 'lumped' in such models. Hence no standard decomposition strategy is available.

Moreover, such a mapping is rarely offered for *specific* DST models, hence the mechanisms underwriting the observed dynamics in such models are not only *ad hoc*, but obscure. Therefore, despite sharing a mathematical heritage with control theory, DST descriptions are not as appropriate for description of cognitive systems.

7.2.2. Mechanisms are abstract

A related consequence of DST's treatment of lumped parameters is that the mechanisms described by DST are highly abstract. That is, to the extent there are mechanisms, there is no mapping from those mechanisms to the internal physical states of the system. Hence, methods of interacting with the system are not evident from such models. Without being able to predict the effects of interventions, the models become less useful to the brain sciences.

7.3. Comparison to statistical descriptions

Statistical descriptions are probably less familiar to philosophers of science and philosophers of mind than the other competitors. For that reason, let us briefly consider an example of a statistical model for explaining behaviour. Often, the perceived motion of an object is influenced by surrounding information (e.g. contrast). This effect, called the 'Thompson effect' (which explains why people tend to drive faster in the fog), was the target of a modelling effort by Stocker & Simoncelli (2006). In this work, they developed a method for determining the nature of the bias that leads people to sometimes make mistaken judgments about motion. Essentially, they derived a method for determining what individual's prior probability was for velocities based on their performance on a behavioural experiment. Stocker and Simoncelli then demonstrated that this inferred prior did a good job of predicting the subject's performance under a wide variety of motion estimation tasks. Although this model is very good at predicting the subject's performance, it has little to say about the mechanism underlying that performance.

7.3.1. Statistical descriptions do not provide decompositions or mechanisms

The implementation independence of statistical descriptions has similar consequences for statistical descriptions as it has for computational ones, though for slightly different reasons. Statistical descriptions have a clear mapping to the empirical data, as they are usually direct descriptions of the properties of the data. However, such a mapping can vary from experiment to experiment, making the mappings not systematic, and hence failing to suggest underlying mechanisms. Indeed, statistical descriptions are often employed exactly when the operating mechanisms are least clear.

Statistical descriptions do not provide any suggested decompositions either. Statistical models typically adopt any decomposition assumptions as part of the methods used to collect the data. They themselves they do not derive such decompositions, as one might wish.

7.3.2. Statistical descriptions are predictive but not explanatory

The 'data-focused' nature of statistical models is both their strength and their weakness. Because statistical descriptions are most concerned with capturing regularities in the data, they are often very useful for prediction. This is appropriate for some purposes, but it does not suit what I take to be the main goal of cognitive science: *explaining* how neural systems work. As discussed in Section 3.1, without knowledge of mechanisms, descriptions would not be explanatory: they would not support intervention, and hence not be useful.

8. If we want a good description of brain function, we ought to adopt a control theoretic approach

How we ought to understand *computation* in the brain is not as computational theory would demand. Instead, if we let pragmatic considerations drive our descriptions (as we ought, otherwise we cannot choose), control theoretic descriptions are most promising for advancing our understanding of neural systems.

Recall the clarification above: this conclusion does not rule out other descriptions of brain function from qualifying as useful. Rather, it suggests that the other descriptions are ultimately heuristics, stepping stones, along the way to stating the description in a control theoretic manner. Often there are equivalent formulations of a given model within different approaches. However, control theory should be primary in this case: it stands the best chance of providing good, useful, unified descriptions of brain function.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *The Psychology of Learning and Motivation, 2*, 89–195.

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

Churchland, P. (1995). *The engine of reason, the seat of the soul: A philosophical journey into the brain*. Cambridge, MA: MIT Press.

Conklin, J., & Eliasmith, C. (2005). An attractor network model of path integration in the rat. *Journal of Computational Neuroscience, 18*, 183–203.

Eliasmith, C. (2000). Is the brain analog or digital? The solution and its consequences for cognitive science. *Cognitive Science Quarterly, 1*(2), 147–170.

Eliasmith, C. (2005). Cognition with neurons: A large-scale, biologically realistic model of the Wason task. In G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society, Stresa, Italy* (pp. 624–629). Mahwah, NJ: Lawrence Erlbaum Associates.

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. thesis, Washington University in St. Louis.

Grafton, S. T., & Hamilton, A. F. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Human Movement Science, 26*, 590–616.

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

Le Cun, Y., & Denker, J. S. (1992). Natural versus universal probability, complexity, and entropy. *IEEE Workshop on the Physics of Computation*, 122–127.

McCulloch, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 7*, 115–133.

Piccinini, G. (2008). Some neural networks compute, others don't. *Neural Networks, 21*(2–3), 311–321.

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

Searle, J. (1990). Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association, 64*, 21–37.

Siegelmann, H. T. (1995). Computation beyond the Turing limit. *Science, 238*(28), 632–637.

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

Stewart, T., & Eliasmith, C. (in press). Compositionality and biologically plausible models. In W. Hinzen, E. Machery, & M. Werning (Eds.), *Oxford handbook of compositionality*. Oxford: Oxford University Press.

Stocker, A., & Simoncelli, E. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9*, 578–585.

Todorov, E. (2007). Optimal control theory. In K. Doya, S. Ishii, A. Pouget, & R. Rao (Eds.), *Bayesian brain: Probabilistic approaches to neural coding* (pp. 269–298). Cambridge, MA: MIT Press.

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

van Gelder, T., & Port, R. (1995). It's about time: An overview of the dynamical approach to cognition. In R. Port, & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 1–44). Cambridge, MA: MIT Press.