MOVING BEYOND METAPHORS: UNDERSTANDING THE MIND FOR WHAT IT IS

Penultimate draft. Final version is in the Journal of Philosophy, 2003, C(10):493-520.

Chris Eliasmith
University of Waterloo
Department of Philosophy
eliasmith@uwaterloo.ca

1 INTRODUCTION

In the last 50 years, there have been three major approaches to understanding cognitive systems and theorizing about the nature of the mind: symbolicism, connectionism, and dynamicism. Each of these approaches has relied heavily on a preferred metaphor for understanding the mind/brain. Most famously, symbolicism, or classical cognitive science, relies on the "mind as computer" metaphor. Under this view, the mind is the software of the brain. Jerry Fodor, [1] for one, has argued that the impressive theoretical power provided by this metaphor is good reason to suppose that cognitive systems have a symbolic "language of thought" which, like a computer programming language, expresses the rules that the system follows. Fodor claims that this metaphor is essential for providing a useful account of how the mind works.

Similarly, connectionists have relied on a metaphor for providing their account of how the mind works. This metaphor, however, is much more subtle than the symbolicist one; connectionists presume that the functioning of the mind is like the functioning of the brain. The subtlety of the "mind as brain" metaphor lies in the fact that connectionists, like symbolicists, are materialists. That is, they also hold that the mind is the brain. However, when providing psychological descriptions, it is the metaphor that matters, not the identity. In deference to the metaphor, the founders of this approach call it "brain-style" processing, and claim to be discussing "abstract networks".[2] This is not surprising since the computational and representational properties of the nodes in connectionist networks bear little resemblance to neurons in real biological neural networks.[3]

Proponents of dynamicism also rely heavily on a metaphor to understand cognitive systems. Most explicitly, van Gelder[4] employs the Watt Governor as a

---

[1] *The Language of Thought* (New York: Crowell, 1975).

[2] J. McClelland and D. Rumelhart, "Future directions" in McClelland, J. and D. Rumelhart, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Vol. 2* (Cambridge: MIT Press, 1986), pp. 547-552.

[3] As discussed in chapter 10 of W. Bechtel and A. Abrahamsen, *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks* (2nd ed., Oxford: Blackwell, 2001).

[4] "What Might Cognition Be, If Not Computation?" *this journal*, XCI, 7 (1995): p. 345-381.

metaphor for mind. It is through his analysis of the best way to characterize this dynamic system that he argues that cognitive systems, too, should be understood as non-representational, low-dimensional, dynamical systems. Like the Watt Governor, van Gelder argues, cognitive systems are essentially dynamic and can only be properly understood by characterizing their state changes through time. The "mind as Watt Governor" metaphor suggests that trying to impose any kind of discreteness, either temporal or representational, will lead to a mischaracterization of minds.

Notably, each of symbolicism, connectionism, and dynamicism, rely on metaphor not only for explanatory purposes, but also for developing their conceptual foundations in understanding the target of the metaphor; i.e., the mind. For symbolicists, the properties of Turing machines become shared with minds. For connectionists the character of representation changes dramatically. Mental representations are taken to consist of "sub-symbols" associated with each node, while "whole" representations are real-valued vectors in a high-dimensional property space.[5] Finally, for the dynamicists, because the Watt Governor is best described by dynamic systems theory, which makes no reference to computation or representation, our theories of mind need not appeal to computation or representation either.

In this paper, I want to suggest that it is time to move beyond these metaphors. We are in the position, I think, to understand the mind for what it is: the result of the dynamics of a complex, physical, information processing system, namely the brain. Clearly, in some ways this is a rather boring thesis to defend. It is just a statement of plain old "monistic materialism" or "token identity theory", call it what you will. It is, in essence, just the uncontroversial view that, so far as we know, you do not have a mind without a brain. But, I further want to argue that the best way to understand this physical system is by using a different set of conceptual tools than those employed by symbolicists, connectionists, and dynamicists individually. That is, the right toolbox will consist in an extended subset of the tools suggested by these various metaphors.

The reason we need to move beyond metaphors is because, in science, analogical thinking can sometimes constrain available hypotheses. This is not to deny that analogies are incredibly useful tools at many points during the development of a scientific theory. It is only to say that, sometimes, analogies only go so far. Take, for instance, the development of the current theory of the nature of light. In the nineteenth century, light was understood in terms of two metaphors: light as a wave, and light as a particle. Thomas Young was the best known proponent of the first view, and Isaac Newton was the best known proponent of the second. Each used their favored analogy to suggest new experiments, and develop new predictions.[6] Thus, these analogies played a role similar to that played by the analogies discussed above in contemporary cognitive science. However, as we know in the case of light, both analogies are false. Hence the famed

---

[5] See, for example, Smolensky's "On the Proper Treatment of Connectionism" *Behavioral and Brain Sciences*, XI, 1 (1988): 1-23.

[6] For a detailed description of the analogies, predictions, and experiments, see Eliasmith and Thagard, "Particles, Waves and Explanatory Coherence" *British Journal of the Philosophy of Science*, XLVIII (1997): p. 1-19.

“wave-particle duality” of light: sometimes it behaves like a particle; and sometimes it behaves like a wave. Neither analogy by itself captures all the phenomena displayed by light, but both are extremely useful in characterizing some of those phenomena. So, understanding what light *is* required moving beyond the metaphors.

I want to suggest that the same is true in the case of cognition. Each of the metaphors mentioned above has some insight to offer regarding certain phenomena displayed by cognitive systems. However, none of these metaphors is likely lead us to all of the right answers. Thus, my project in trying to move beyond these metaphors is a synthetic one. I want to provide a way of understanding cognitive systems that draws on the strengths of symbolicism, connectionism, and dynamicism. The best way of doing this is to understand minds for what they are. To phrase this as a conditional, if minds are the behavior of complex, dynamic, information processing systems, then we should use the conceptual tools that we have for understanding such systems when trying to understand minds. In this paper I outline a general theory that describes representation and dynamics in neural systems (R&D theory) that realizes the consequent of this conditional. I argue that R&D theory can help unify neural and psychological explanations of cognitive systems and that the theory suggests a need to re-evaluate standard functionalist claims.

First, however, it is instructive to see how R&D theory does not demand the invention of new conceptual tools; the relevant tools are already well-tested. So, in some ways, the theory is neither risky nor surprising. What is surprising, perhaps, is that our most powerful tools for understanding the kinds of systems that minds are, have yet to be applied to minds. I suggest that this surprising oversight is due to an over-reliance on the “mind as computer” metaphor.

2 A BRIEF HISTORY OF COGNITIVE SCIENCE

While the main purpose of this paper is clearly not historical, a brief perusal of the relevant historical landscape helps situate both the theory and the subsequent discussion.

2.1 Prehistory

While much is sometimes made of the difference between philosophical and psychological behaviorism, there was general agreement on this much: internal representations, states, and structures are irrelevant for understanding the behavior of cognitive systems. For psychologists, like Watson and Skinner, this was true because only input/output relations are scientifically accessible. For philosophers, like Ryle, this was true because mental predicates, if they were to be consistent with natural science, must be analyzable in terms of behavioral predicates. In either case, looking inside the “black box” that was the object of study, was prohibited for behaviorists.

Interestingly, engineers of the day respected a similar constraint. In order to understand dynamic physical systems, the central tool they employed was (classical) control theory. Classical control theory, notoriously, only characterizes physical systems in terms of their input/output relations in order to determine the relevant controller. Classical control theory was limited to designing non-optimal, single-variable, static controllers and depended on graphical methods, rules of thumb, and did not allow for the inclusion of noise.[7] While the limitations of classical controllers and methods are now well-known, they nevertheless allowed engineers to build systems of kinds they had not systematically built before: goal-directed systems.

While classical control theory was useful, especially in the 1940s, for building warhead guidance systems, some researchers thought it was clearly more than that. They suggested that classical control theory could provide a theoretical foundation for describing living systems as well. Most famously, the interdisciplinary movement founded in the early 1940s known as "cybernetics" was based on precisely this contention.[8] Cyberneticists claimed that living systems were also essentially goal-directed systems. Thus, closed-loop control, it was argued, should be a good way to understand the behavior of living systems. Given the nature of classical control theory, cyberneticists focused on characterizing the input/output behavior of living systems, not their internal processes. With the so-called "cognitive revolution" of the mid-1950s, interest in cybernetics waned due in part to its close association with, and similar theoretical commitments to, behaviorism.

2.2 The cognitive revolution

In the mid-1950s, with the publication of a series of seminal papers, [9] the "cognitive revolution" took place. One simplistic way to characterize this shift from behaviorism to cognitivism is that it became no longer taboo to look inside the black box. Quite the contrary: internal states, internal processes, and internal representations became standard fare when thinking about the mind. Making sense of the insides of that black box was heavily influenced by concurrent successes in building and programming computers to perform complex tasks. Thus, many early cognitive scientists saw, when they opened the lid of the box, a computer. As explored in detail by Jerry Fodor, [10] "[c]omputers show us how to connect semantical with causal properties for *symbols*", thus computers have what it takes to be intentional minds. Once cognitive scientists began to think of minds as computers, a number of new theoretical tools became available for characterizing cognition. For instance, the computer's theoretical counterpart, the Turing machine, suggested novel philosophical theses, including functionalism and multiple realizability, about the mind. More practically, the typical architecture of computers, the von Neumann architecture, was thought by many to be relevant for understanding our cognitive architecture.

---

[7] For a succinct description of the history of control theory, see Lewis's *Applied Optimal Control and Estimation* (New York: Prentice-Hall, 1992).

[8] For a statement of the motivations of cybernetics, see Rosenblueth, Wiener, and Bigelow "Behavior, Purpose, and Teleology" *Philosophy of Science*, X (1943): 18-24.

[9] These papers include, but are not limited to: A. Newell, C. Shaw, and H. Simon, "Elements of a Theory of Human Problem Solving" *Psychological Review*, LXV (1958): 151-166; G. Miller "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information" *Psychological Review*, LXIII (1956): 81-97; J. Bruner, J. Goodnow, and G. Austin, *A Study Of Thinking*. (New York: Wiley, 1956).

[10] *Psychosemantics* (Cambridge: MIT Press, 1987), p. 18.

However, adoption of the von Neumann architecture for understanding minds was seen by many as poorly motivated. As a result, the early 1980s saw a revival of the so-called "connectionist" research program. Rather than adopting the architecture of a digital computer, these researchers felt that an architecture more like that seen in the brain would provided a better model for cognition.[11] As a result of this theoretical shift, connectionists were very successful at building models sensitive to statistical structure, and could begin to explain many phenomena not easily captured by symbolicists (e.g., object recognition, generalization, learning, etc.).

For some, however, connectionists had clearly not escaped the influence of the "mind as computer" metaphor. Connectionists still spoke of representations, and thought of the mind as a kind of computer. These critics argued that minds are not essentially computational, they are essentially physical, dynamic systems.[12] They suggested that if we want to know which functions a system can actually perform in the real world, we must know how to characterize the system's dynamics. Furthermore, since cognitive systems evolved in dynamic environments, we should expect evolved control systems, like brains, to be more like the Watt Governor – dynamic, continuous, coupled directly to what they control – than like a discrete state Turing machine that computes over "disconnected" representations. As a result, these "dynamicists" suggested that dynamic systems theory, not computational theory, was the right quantitative tool for understanding minds. They claimed that notions like 'chaos,' 'hysteresis,' 'attractors,' and 'state-space' underwrite the conceptual tools best-suited for describing cognitive systems.

2.3 A puzzling oversight

In some ways, dynamicists revived the commitments of the predecessors of the cognitive revolution. Notably, the Watt Governor is a standard example of a classical control system. If minds are to be like Watt Governors, they are to be like classical control systems; just what the cyberneticists had argued. One worry with this retrospective approach is that the original problems come along with the original solutions. The limitations of classical control theory are severe, so severe that they will probably not allow us to understand a system as complex as the brain.

However, an important series of theoretical advances in control theory went completely unnoticed during the cognitive revolution. During the heyday of the computer, in the 1960s, many of the limitations of classical control theory were rectified with the introduction of what is now known as "modern" control theory.[13] Modern control theory introduced the notion of an "internal system description" to control theory.

---

[11] As discussed in both Smolensky, *op. cit.*, and the introduction to D. Rumelhart and J. McClelland, eds. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Vol. 1*(Cambridge: MIT Press, 1986).

[12] See the various contributions to R. Port and T. van Gelder, eds., *Mind as Motion: Explorations in the Dynamics of Cognition* (Cambridge: MIT Press, 1995), especially the editors' introduction.

[13] This introduction is largely credited to R. Kalman in his "A New Approach to Linear Filtering and Prediction Problems" *ASME Journal of Basic Engineering*, LXXXII (1960): 35-45.

An internal system description is one that includes *system state variables* (i.e., variables describing the state of the system itself) as part of the description (see figure 1). It is interesting that with the cognitive revolution, researchers interested in the behavior of living systems realized they needed to "look inside" the systems they were studying and, at about the same time, researchers interested in controlling engineered systems began to "look inside" as well. Both, nearly simultaneously, opened their black box. However, as already discussed, those interested in cognitive behavior adopted the computer as a metaphor for the workings of the mind. Unfortunately, the ubiquity of this metaphor has served to distance the cognitive sciences from modern control theory. Nevertheless, I argue below that modern control theory offers tools better-suited than computational theory for understanding biological systems as fundamentally physical, dynamic systems operating in changing, uncertain environments.

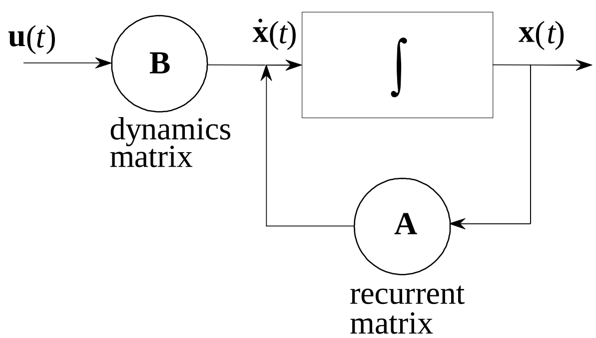

Figure 1: A modern control theoretic system description. The vector $\mathbf{u}(t)$ is the input to the system. **A** and **B** are matrices that define the behavior of the system, $\mathbf{x}(t)$ is the system state variable (generally a vector), and $\dot{\mathbf{x}}(t)$ is the derivative of the state vector. The standard transfer function in control theory, as shown in the rectangle, is integration.[14]

This is not to suggest that each of the dominant metaphors should be taken as irrelevant to our understanding of minds. Both connectionism and dynamicism highlight important limitations of the original "mind as computer" metaphor. Connectionism challenged the symbolicist conception of representation, noting how important statistical considerations are for capturing certain kinds of cognitive phenomena. Dynamicist critiques of symbolicism focused on its lack of a principled account of the temporal properties of cognitive systems.[15] Nevertheless, it was the symbolicists, armed with their metaphor, who rightly justified opening the black box. And, furthermore, both connectionism and dynamicism introduced their own misleading metaphors.

---

[14] I have simplified this diagram for a generic linear system from the typical, truly general one found in most control theory texts by excluding the feedthrough and output matrices. Nothing turns on this simplification in this context.

[15] T. van Gelder, *op. cit.*

3 REPRESENTATION AND DYNAMICS IN NEURAL SYSTEMS: A THEORY

Moving beyond metaphors, that is, taking seriously the view that minds are complex, physical, dynamic, and information processing systems, means using our best tools for describing systems with these properties. In the remainder of this section, I propose and defend a theory of representation and dynamics in neural systems (R&D theory) that takes precisely this approach. R&D theory relies on modern control theory, information theory, and recent results from neuroscience to provide an account of what minds are.[16]

Below I have broken this account into three parts. The first defines representation, the second describes computation, and the third section, on dynamics, shows how the preceding characterizations of representation and computation can be merged with control theory to provide an account of neural and cognitive function. The result, I argue, is a theory that avoids the weaknesses and capitalizes on the strengths of past approaches.

3.1 Representation

A central tenet of R&D theory is that we can adapt the information theoretic account of *codes* to understanding representation in neural systems. Codes, in engineering, are defined in terms of a complimentary encoding and decoding procedure between two alphabets in such systems. Morse code, for example, is defined by the one-to-one relation between letters of the Roman alphabet, and the alphabet composed of a standard set of dashes and dots. The encoding procedure is the mapping from the Roman alphabet to the Morse code alphabet and the decoding procedure is its inverse.

In order to characterize representation in a cognitive/neural system, we can identify each of these procedures and their relevant alphabets. The encoding procedure is quite easy to identify: it is the transduction of stimuli by the system resulting in a series of neural "action potentials", or "spikes". The precise nature of this encoding has been explored in depth via quantitative models.[17] So, encoding is what neuroscientists typically talk about. When I show a cognitive system a stimulus, some neurons or other "fire". Unfortunately, neuroscientists often stop here in their characterization of representation, but this is insufficient. We also need to identify a decoding procedure, otherwise there is no way to determine the relevance of the encoding for the system. If no information about the stimulus can be extracted from the spiking neuron, then it makes no sense to say that it represents the stimulus. Representations, at a minimum, must potentially be able to "stand-in" for their referents.

Quite surprisingly, despite typically nonlinear encoding, a good linear decoding can be found.[18] And, there are several established methods for determining linear

---

[16] For an in-depth technical description of this approach, see Eliasmith and Anderson, *Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems* (Cambridge: MIT Press, 2003).

[17] See J. Bower and D. Beeman, *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System* (Berlin: Springer Verlag, 1998) for a review of such models.

[18] As demonstrated by F. Rieke, D. Warland, R. de Ruyter van Steveninick, and W. Bialek *Spikes: Exploring the Neural Code* (Cambridge: MIT Press, 1997), pp. 76-87.

decoders given the statistics of the neural populations that respond to certain stimuli.[19] Notably, these decoders are sensitive both to the temporal statistics of the stimuli and to what other elements in the population encode. Thus, if you have multiple neurons involved in the (distributed) representation of a time-varying object, they can "cooperate" to provide a better representation.

Having specified the encoding and decoding procedures, we still need to specify the relevant alphabets. While the specific cases will diverge greatly, we can describe the alphabets generally: neural responses (encoded alphabet) code physical properties (decoded alphabet). In fact, it is possible to be a bit more specific. Neuroscientists generally agree that the basic element of the neural alphabets is the neural spike. However, there are many possibilities for how such spikes are used: average production rate of neural spikes (i.e. a rate code); specific timings of neural spikes (i.e. a timing code); population-wide groupings of neural spikes (i.e. a population code); or the synchrony of neural spikes across neurons (i.e. a synchrony code). Of these possibilities, arguably the best evidence exists for a combination of timing codes and population codes.[20] For this reason, let us take the combination of these basic coding schemes to define the alphabet of neural responses. Thus, the encoded alphabet is the set of temporally patterned neural spikes over populations of neurons.

It is much more difficult to be specific about the nature of the alphabet of physical properties. Of course, we can begin by looking to the physical sciences for categories of physical properties that might be encoded by nervous systems. Indeed, we find that many of the properties that physicists traditionally use do seem to be represented in nervous systems; e.g., displacement, velocity, acceleration, wavelength, temperature, pressure, mass, etc. But, there are many physical properties not discussed by physicists which also seem to be encoded in nervous systems; e.g. red, hot, square, dangerous, edible, object, conspecific, etc. Presumably, all of these "higher-level" properties are inferred on the basis of representations of properties more like those that physicists talk about. In other words, encodings of 'edible' depend, in some complex way, on encodings of "low-level" physical properties like wavelength, velocity, etc. While R&D theory itself does not determine precisely what is involved in such complex relations, there is reason to suppose that R&D theory provides the necessary tools for describing such relations. To see why this is so, let us consider a simple example.

It is clearly important for an animal to be able to know where various objects in its environment are. As a result, in mammals, there are a number of internal representations of signals that convey and update this kind of information. One such representation is found in parietal areas, particularly in lateral intraparietal cortex (LIP). For simplicity, let us consider the representation of only the horizontal position of an object in the environment. As a population, some neurons in this area *encode* and object's position over time. This representation can be understood as a scalar variable, whose

---

[19] As discussed in Eliasmith and Anderson, *op. cit.*

[20] For an overview of this evidence, see F. Rieke et al., *op. cit.*, Eliasmith and Anderson, *op. cit.*, and L. Abbott, "Decoding Neuronal Firing and Modelling Neural Networks" *Quarterly Review of Biophysics*, XXVII, 3 (1994): 291-331.

units are degrees from midline (decoded alphabet), that is encoded into a series of neural spikes (encoded alphabet). Using the quantitative tools mentioned earlier, we can determine the relevant decoder. Once we have such a decoder, we can then estimate what the actual position of the object is given the neural spiking in this population. Thus we can determine precisely how well (or what aspects of) the original property (in this case the actual position) is represented by the neural population. We can then use this characterization to understand the role that the representation plays in the cognitive system as a whole.

As mentioned, this is a simple example. But notice that it not only describes how to characterize representation, it also shows how we can move from talking about neurons to talking about "higher-level" variables, like object position. That is, we can move from discussing the "basic" representations (i.e., neural spikes) to "higher-level" representations (i.e., mathematical objects with units). This suggests that we can build up a kind of "representational hierarchy" that permits us to move further and further away from the neural-level description, while remaining responsible to it. For instance, we could talk about the larger population of neurons that encodes position in three dimensional space. We could dissect this higher-level description into its lower-level components (i.e., horizontal, vertical, and depth positions), or we could dissect it into the activity of individual neurons. Which description we employ will depend on the kind of explanation we need. Notably, this hierarchy can be rigorously and generally defined to include scalars, vectors, functions, vector fields, and so on.[21] The fact that all of the levels of such a hierarchy can be written in a standard form suggests that this characterization provides a unified way of understanding representation in neurobiological systems.

Note that the focus of this paper is on how to characterize representational states, computations over these states, and the dynamics of these states. As a result, I do not generally address concerns related to content determination. This is largely because such a discussion would lead me far a field. Nevertheless, it is interesting to note that R&D theory is suggestive of a particular approach to content determination. Notice, first, that both the encoding and decoding are essential for determining the identity of a representation. This means that both what causes a neural state, and how that state is used by the system are likely to play a role in content determination. This suggests that some kind of two-factor theory is consistent with R&D theory. As well, a single neural state may play a role in multiple contents concurrently, because distinct, yet related, representations (and hence contents) can be identified at different levels of organization at the same time. This may initially seem problematic, but, because the relation between levels of representation is quantitatively defined (hence we know precisely what role a single neural state is playing in each of the representations defined over it), we should expect the parallel content relations to also be well-defined. Of course, such comments only provide a hint of a theory of content, they do not constitute one.[22]

[21] This generalization is made explicit in Eliasmith and Anderson, *op. cit.*, pp. 79-80.

[22] Note that in order to understand the relation between representations at a given organizational level, we need to consider computational relations, as discussed in the next section. For an in-depth, but preliminary discussion of a theory of content that is consistent with R&D theory, see my *How Neurons*

In any case, there is no reason to consider such a theory of content if its underlying theoretical assumptions are not appropriate to cognitive systems. So, we should notice that the strength of the previous characterization of representation lies in its generality. That is, regardless of what the higher-level representations look like (i.e., what kind of mathematical objects with units they are), R&D theory will apply. So R&D theory, while having definite consequences for what constitutes a good representational story, is silent as to which particular one is correct for a given neural system. This is desirable for a theory of mind because higher-level representations are clearly theoretical postulates (at least at this point in the development of neuroscience). While we can directly measure the voltage changes of individual neurons, making claims about how they are grouped to represent the world is not easily confirmable. Presumably, the right representational story will be the most coherent and predictively successful one.

3.2 Computation

Of course, no representational characterization will be justified if it does not help us understand how the system functions. Luckily, a good characterization of neural representation paves the way for a good understanding of neural computation. This is because, like representations, computations can be characterized using decoding. But, rather than using the "representational decoder" discussed earlier, we can use a "transformational decoder". We can think of the transformational decoder as defining a kind of biased decoding. That is, in determining a transformation, we extract information *other than* what the population is taken to represent. The bias, then, is away from a "pure", or representational, decoding of the encoded information. For example, if we think that the quantity x is encoded in some neural population, when defining the representation we determine the representational decoders that estimate x. However, when defining a computation we identify transformational decoders that estimate some function, $f(x)$, of the represented quantity. In other words, we find decoders that, rather than extracting the signal represented by a population, extract some transformed version of that signal. The same techniques used to find representational decoders are applicable in this case, and result in decoders that can support both linear and nonlinear transformations.[23]

Given this understanding of neural computation, there is an important ambiguity that arises in the preceding characterization of representation. It stems from the fact that information encoded into a population may now be decoded in a variety of ways. Suppose we are again considering the population that encodes object position. Not surprisingly, we can decode that population to provide an estimate of object position. However, we can also decode that same information to provide an estimate of some function of object position (e.g. the square). Since representation is defined in terms of encoding and decoding, it seems that we need a way to pick which of these possible decodings is the relevant one for defining the representation in the original population. To

---

Mean: A Neurocomputational Theory of Representational Content (St. Louis: Washington University Ph.D. in Philosophy, 2000).

[23] As demonstrated in Eliasmith and Anderson, *op. cit.*, pp. 143-160.

resolve this issue we can specify that what a population represents is determined by the decoding that results in the quantity that all other decodings are functions of. Thus, in this example, the population would be said to represent object position (since both object position and its square are decoded). Of course, object position is also a function of the square of object position (i.e. $x = \sqrt{x^2}$)). This further difficulty can be resolved by noticing that the right physical quantities (i.e. the decoded alphabet) for representation are those that are part of a coherent, consistent, and useful theory. In other words, we characterize cognitive systems as representing positions because we characterize the world in terms of positions, and cognitive systems represent the world.

Importantly, this understanding of neural computation applies at all levels of the representational hierarchy, and accounts for complex transformations. So, for example, it can be used to define inference relations, traditionally thought necessary for characterizing the relations between high-level representations. Again consider the specific example of determining object position. Suppose that the available data from sensory receptors make it equally likely that an object is in one of two positions (represented as a bimodal probability distribution over possible positions). However, also suppose that prior information, in the form of a statistical model, favors one of those positions (perhaps one is consistent with past known locations given current velocity, and the other is not). Using the notion of computation defined above, it is straightforward to build a model that incorporates transformations between, and representations of 1) the top-down model, 2) the bottom-up data, and 3) the actual inferred position of the object (inferred based on Bayes' rule, for example). As expected, in this situation the most likely position given the *a priori* information would be the one consistent with the top-down model. However, if the bottom-up data is significantly stronger in favor of an alternate position, this will influence the preferred estimate, and so on.[24] So, although simple, performing linear decoding can support the kinds of complex transformations needed to articulate descriptions of cognitive behavior. Statistical inference is just one example.

Before moving on to a consideration of dynamics, it is important to realize that this way of characterizing representation and computation does not demand that there are “little decoders” inside the head. That is, this view does not entail that the system itself needs to decode the representations it employs. In fact, according to this account, there are no directly observable counterparts to the representational or transformational decoders. Rather, they are embedded in the synaptic weights between neighboring neurons. That is, coupling weights of neighboring neurons indirectly reflect a particular population decoder, but they are not identical to the population decoder. This is because connection weights are best characterized as determined by both the decoding of the incoming signal and the encoding of the outgoing signal. Practically speaking, this means that changing a connection weight both changes the transformation being performed and the tuning curve of the receiving neuron. As is well known from both connectionism and computational neuroscience, this is exactly what happens in such networks. In essence,

---

[24] For the technical details and results of the model described here, see *ibid*., pp. 275-283. For a brief discussion of more logic-like inference on symbolic representations, see section 4.1.

the encoding/decoding distinction is not one that neurobiological systems need to respect in order to perform their functions, but it is extremely useful in trying to understand such systems and how they do, in fact, manage to perform those functions.

3.3 Dynamics

While it may be understandable that dynamics were initially ignored by those studying cognitive systems as computational (theoretically, time is irrelevant for successful computation), it would be strange, indeed, to leave dynamics out of the study of minds as physical, neurobiological systems. Even the simplest nervous systems performing the simplest functions demand temporal characterizations (e.g. locomotion, digestion, sensing). It is not surprising, then, that single neural cells have almost always been modeled by neuroscientists as essentially dynamic systems. In contemporary neuroscience, electrophysiologists often analyze cellular responses in terms of 'onsets', 'latencies', 'stimulus intervals', 'steady states', 'decays', etc. – these are all terms describing temporal properties of a neuron's response. The fact is, the systems under study in neurobiology are dynamic systems and as such they make it very difficult to ignore time.

Notably, modern control theory was developed precisely because understanding complex dynamics is essential for building something that works in the real world. Modern control theory permits both the analysis and synthesis of elaborate dynamic systems. Because of its general formulation, modern control theory applies to chemical, mechanical, electrical, digital, or analog systems. As well, it can be used to characterize non-linear, time-varying, probabilistic, or noisy systems. As a result of this generality, modern control theory is applied to a huge variety of control problems, including autopilot design, spacecraft control, design of manufacturing facilities, robotics, chemical process control, electrical systems design, design of environmental regulators, and so on. It should not be surprising, then, that it proves useful for the analysis of the dynamics of cognitive, neurobiological systems as well.

Having identified quantitative tools for characterizing dynamics, and for characterizing representation and computation, how do we bring them together? An essential step in employing the techniques of control theory is identifying the system state variable ($\mathbf{x}(t)$ in figure 1). Given the preceding analysis of representation, it is natural to suggest that the state variable *just is* the neural representation.

However, things are not quite so simple. Because neurons have intrinsic dynamics dictated by their particular physical characteristics, we must adapt standard control theory to neurobiological systems. Fortunately, this can be done without loss of generality.[25] As well, all of the computations needed to implement such systems can be implemented using transformations as defined earlier. As a result, we can directly apply the myriad techniques for analyzing complex dynamic systems that have been developed using modern control theory to this quantitative characterization of neurobiological systems.

---

[25] For the relevant derivations see *ibid*., pp. 221-225.

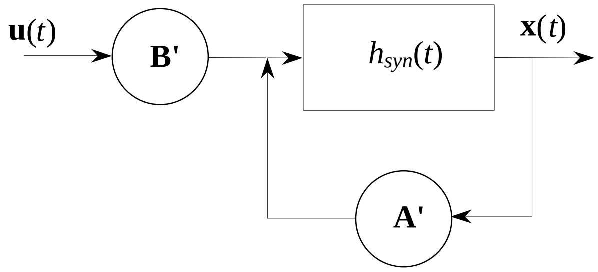

Figure 2: A control theoretic description of neurobiological systems. All variables are the same as in figure 1. However, the matrices **A**' and **B**' take into account that there is a different transfer function, $h_{syn}(t)$, than in figure 1. As well, **x**(t) is taken to be represented by a neural population.

To get a sense of how representation and dynamics can be integrated, let us revisit the simple example introduced previously; object position representation in area LIP. Note that animals often need to know not just where some object currently is, they need to remember where it was. Decades of experiments in LIP have shown that neurons in this area have sustained responses during the interval between a brief stimulus presentation and a delayed "go" signal.[26] In other words, these neurons seem to underlie the (short-term) memory of where an interesting object is in the world. Recall that I earlier characterized this area as representing $x(t)$, the position of an object. Now we know the dynamics of this area; namely, stability without subsequent input. According to R&D theory, we can let the representation be the state variable for the system whose dynamics are characterized in this manner.

Mathematically, these dynamics are easy to express with a differential equation: $x(t) = \int u(t)\,dt$. In words, this system acts as a kind of integrator. In fact, neural systems with this kind of dynamics are often called "neural integrators" and are found in a number of brain areas, including brainstem, frontal lobes, hippocampus, and parietal areas. Neural integrators act like memories because when there is no input (i.e., $u(t)$=0), the change in the output over time is 0 (i.e., $\dot{x} = \frac{dx}{dt} = 0$). Thus, such systems are stable with no subsequent inputs. Looking for the moment at figure 1, we can see that the desired values of the **A** and **B** matrices will be 0 and 1 respectively in order to implement a system with these dynamics. Since we have a means of "translating" this canonical control system into one that respects neural dynamics, we can determine the values of **A**' and **B**' in figure 2; they turn out to be 1 and τ (the time constant of the intrinsic neural dynamics), respectively. We can now set about building a model of this system at the

[26] For a detailed description and review of these experiments and their results, see C. Colby and M. Goldberg, "Space and Attention in Parietal Cortex" *Annual Review of Neuroscience*, XX (1999): 319-349, and R. Andersen, L. Snyder, D. Bradley, and J. Xing, "Multimodal Representation of Space in the Posterior Parietal Cortex and its Use in Planning Movements" *Annual Review of Neuroscience*, XX (1997): 303-330.

level of single spiking neurons which gives rise to these dynamics – originally described at a higher level. In fact, the representation in LIP is far more complex than this, but the representational characterization of R&D theory is a general one, so such complexities are easily incorporated. As well, more complex dynamics are often necessary for describing neural systems, but again, the generality of R&D theory allows these to be incorporated using similar techniques.[27] So, while the neural integrator model is extremely simple, it shows how R&D theory provides a principled means of explaining a cognitive behavior (i.e., memory) in a neurally plausible network.

3.4 Three principles

R&D theory is succinctly summarized by three principles:

1. Neural representations are defined by the combination of nonlinear encoding (exemplified by neuron tuning curves) and weighted linear decoding.
2. Transformations of neural representations are functions of variables that are represented by neural populations. Transformations are determined using an alternately weighted linear decoding.
3. Neural dynamics are characterized by considering neural representations as control theoretic state variables. Thus, the dynamics of neurobiological systems can be analyzed using control theory.

To summarize these principles, figure 3 shows a “generic neural subsystem.” This figure synthesizes the previous characterizations of representation, computation, and dynamics, across multiple levels of description.

[27] For various examples, see Eliasmith and Anderson, *op. cit*., especially chapters 6 and 8.

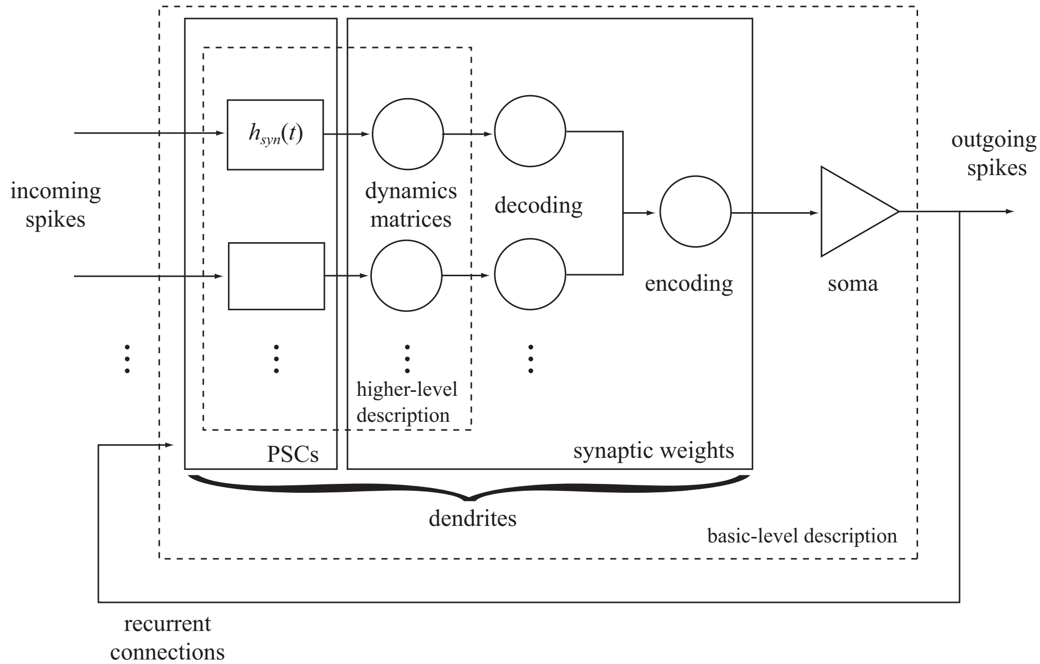

Figure 3: Generic neural subsystem. A synthesis of the preceding characterizations of representation (encoding/decoding), computation (biased decoding), and dynamics (captured by $h_{syn}(t)$ and the dynamics matrices). Dotted lines distinguish basic-level (i.e., neural) and higher-level (i.e., mathematical objects with units) descriptions.

While recent, this approach has been successfully used to characterize a number of different systems including the vestibular system, the lamprey locomotive system, the eye stabilization system, working memory,[28] and the limb control system.[29] As well, a rigorous formulation of each of these principles is available.[30]

4 COMPARISON TO CURRENT APPROACHES

Notice that R&D theory is not a description of cognitive systems as "being like" anything; cognitive systems are neurobiological systems that are best described by certain well-established quantitative tools for describing physical systems. That is, I am not proposing an analogy of "the mind as neural integrator" or even "the mind as a control system". While we may notice that certain control structures mimic some behaviors of neurobiological systems, we have to build the detailed neurobiological model and then determine if the mind really does work that way. In other words, R&D theory in no way suggests we should stop at the analogy. Rather, it gives us principled means of comparing a full-blown, neuron-level implementation with the actual neurobiological system.

[28] These examples can all be found in *ibid.*

[29] As described in Z. Nenadic, C. Anderson, and B. Ghosh, "Control of Arm Movement Using a Population of Neurons" in J. Bower, ed., *Computational Neuroscience: Trends In Research 2000* (Amsterdam: Elsevier Press, 2000).

[30] This formulation can be found in Eliasmith and Anderson, *op. cit.*, pp. 230-231.

Connectionists, symbolicists, and dynamicists typically build models that begin and end with metaphors; in particular, they do not specify how their models relate to the mind as a physical system. But, the devil is in the details.

To see what is gained by moving away from the various metaphors that have dominated theorizing about the mind, let me briefly compare R&D theory to past approaches. In doing so, I describe several of what I take to be the central strengths and weaknesses of each approach. These lists are not intended to be exhaustive, but representative. So long as I am right about at least one of each, my claim that R&D theory should be preferred follows.

4.1 Symbolicism

The central problem for symbolicists, which I have already mentioned, is that time is largely ignored, or considered only after the fact.[31] Other typical concerns about symbolic models include their brittleness (i.e., lack of ability to survive partial destruction of the system or its representations), high-level discreteness,[32] and unconvincing descriptions of low-level perceptual processes.[33] R&D theory suffers from none of these limitations due to its essential inclusion of time and its responsibility to the underlying neural architecture.

The central strengths of symbolicism are demonstrated by its many past successes (e.g. ACT, SOAR, etc.). These are largely due to its espousal of cognitivism, i.e., its willingness to peer inside the black box. Doing so has made representation an essential ingredient for providing good explanations of cognitive systems. As well, symbolicism is supported by a powerful and general theory of computation. Together, the commitment to representation and the quantitative theory of computation, make for a unified understanding of cognitive systems. R&D theory shares similar strengths. While computational theory is bolstered by including control and information theory, and the notion of representation is sharpened to relate directly to neural systems, the ability to provide unified representational explanations remains.

That being said, a typical concern of symbolicists regarding approaches that are concerned with neural implementations, is that the demonstrated symbol manipulating abilities of cognitive systems are lost in the concern for neural detail. In one sense, this issue is easily addressed in the context of R&D theory. This is because the numeric representations in R&D theory are just another kind of syntax. While it is not a typical syntax for the logic used to describe cognitive function by symbolicists, the syntax itself

---

[31] Perhaps the most concerted effort to include time in such models is in Newell's *Unified Theories Of Cognition* (Cambridge: Harvard University Press, 1990). However, this attempt is both after-the-fact and largely inconsistent, as discussed in my "The Third Contender: A Critical Examination of the Dynamicist Theory of Cognition" *Philosophical Psychology*, IX, 4 (1996): 441-463.

[32] As discussed in E. Smith, "Concepts and Induction", in M. Posner, ed., *Foundations of Cognitive Science* (Cambridge: MIT Press, 1989), pp. 501-526.

[33] This is made clear in the case of visual processes by P. Churchland, V. Ramachandran, and T. Sejnowski, "A Critique of Pure Vision", in C. Koch and J. Davis, eds., *Large-Scale Neuronal Theories of the Brain* (Cambridge: MIT Press, 1994).

does not determine the kinds of functions computable with the system.[34] Given the discussion in section 3.2, we know that this theory supports quite general computation, i.e., linear and nonlinear functions. As a result, most, if not all, of the functions computed with standard symbolicist syntax can be computed with the numerical syntax adopted by R&D theory. More to the point, perhaps, past work using numerical distributed representations has shown that structure sensitive processing of the kind demanded by Fodor and Pylyshyn[35] can be achieved in such a system.[36] Furthermore, this kind of representational system has been used to model high-level cognitive functions, like analogical mapping.[37] As a result, structured symbol manipulation is not lost by adopting R&D theory. Rather, a precise neural description of such manipulation is gained.

4.2 Connectionism

Of past approaches, connectionism is probably the most similar to R&D theory. This raises the question: Is R&D theory merely glorified connectionism? A first response is to note that glorifying connectionism (i.e., making it more neurally plausible) is no mean feat. The neural plausibility of many connectionist models leaves much to be desired. Localist models are generally not neurally plausible at all. But even distributed models seldom "look" much like real neurobiological systems. They include neurons with continuous, real-valued inputs and outputs, and often have purely linear or generic sigmoid response functions. Real neurobiological networks have highly heterogeneous, nonlinear, spiking neurons. Connectionist themselves are seldom certain precisely what the relation is between their models and what goes on the brain.[38]

Of course, such connectionist models are far more neurally plausible than symbolicist ones. As a result, they are not brittle like symbolic systems, but rather degrade gracefully with damage. As well, they are supremely statistically sensitive and are thus ideal for describing many perceptual and cognitive processes that have eluded symbolicists. And finally, connectionist models do, on occasion, consider time to be central to neural processing.[39] Again, R&D theory shares each of these strengths and, in fact, improves on a number of them (e.g., neural plausibility, and the integration of time).

---

[34] This general point has been argued in J. Girard, "Proof-Nets: The Parallel Syntax for Proof-Theory" in Ursini and Agliano, eds., *Logic and Algebra*, (Marcel Dekker: New York, 1996), and J. Girard, "Linear Logic" *Theoretical Computer Science*, L, 1 (1987): 1-102.

[35] "Connectionism and Cognitive Architecture: A Critical Analysis" *Cognition*, XXVIII (1988): 3-71.

[36] As demonstrated in T. Plate, *Distributed Representations and Nested Compositional Structure* (Toronto: University of Toronto Ph.D. in Computer Science, 1994).

[37] As in Eliasmith and Thagard, "Integrating Structure and Meaning: A Distributed Model of Analogical Mapping" *Cognitive Science*, XXV, 2 (2001): 245-286.

[38] See the various discussions by the editors in J. McClelland and D. Rumelhart, eds. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Vol. 2*.

[39] See the selection of the examples in P. Churchland and T. Sejnowski, *The Computational Brain* (Cambridge: MIT Press, 1992).

But, more importantly, R&D theory also improves on connectionism. Connectionism has been predominantly a bottom-up approach to cognitive modeling. The basic method is straightforward: connect simple nodes together and train them to compute complex functions. While this approach can provide some useful insights (e.g. determining what kinds of statistical structure can be detected in the training set), it is unlikely to lead to a useful model of a brain that consists of billions of neurons. Connecting ten billion nodes together and training them will probably not result in much. So, one of the main difficulties that connectionism suffers from is the lack of a principled method.

Progress in decomposing complex physical systems often necessitates an integration of bottom-up and top-down information.[40] So, in the case of neurobiology, it is essential to be able to test top-down hypotheses regarding brain function that are consistent with known lower-level facts. That is, we must be able to relate high-level characterizations of psychological processes (e.g., "working memory") to more specific implementational claims (e.g., that networks of certain kinds of neurons can realize such processes). Connectionists, unfortunately, have no principled method for incorporating top-down constraints on the design and analysis of their models. R&D theory, in contrast, explicitly combines both higher and lower-level constraints on models.

The third principle of R&D theory captures this synthesis. It is with this principle that the analyses of representation, computation, and dynamics come together (see figure 3). As an example, consider the recent proposal by Rao and Ballard[41] that the visual systems acts like a dynamic, optimal linear estimator (i.e., a linear control structure known as a Kalman filter).[42] Using R&D theory, we can build a large-scale, complex network to test this hypothesis. This is because the hypothesis is a precise high-level description, there is a significant amount of neural data available regarding the visual system, and principle three tells us how to combine these. Using the tools of connectionism, we simply cannot test this kind of high-level claim. It is not at all evident how we can train a network to realize an optimal estimator, or what an appropriate network architecture would be. So, R&D theory is able to test high-level hypotheses in ways not available to connectionists. This makes R&D theory much better able to bridge the gap between psychological and neural descriptions of behavior than connectionism.

Another way of making this point is to contrast the kind of characterization of dynamics principle three offers, with that typical of connectionism. While connectionists often consider the importance of time, and, in particular, have introduced and explored the relation between recurrent networks and dynamic computation, they do not have a

---

[40] As argued by W. Bechtel and R. Richardson in their *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research* (Princeton: Princeton University Press, 1993).

[41] "Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects" *Nature Neuroscience*, II, 1 (1999): 79-87.

[42] Another high-level hypothesis regarding the use of the Kalman filter has been made in the context of the construction and use of cognitive maps in hippocampus in K. Balakrishnan, O. Bousquet, and V. Honavar, "Spatial Learning and Localization in Animals: A Computational Model and its Implications for Mobile Robots" *Adaptive Behavior*, VII, 2 (1999): 173-216.

systematic means of analyzing or constructing networks with these properties. Principle three, by adopting control theory, makes such analyses possible within R&D theory. That is, control theory has a large set of well-established quantitative tools for both analyzing and constructing control structures. And, because R&D theory provides a means of inter-translating standard and "neural" control structures, such tools can be used in a neurobiological context. This is extremely important for understanding the dynamic properties, and otherwise predicting the overall behavior of a network constructed using the R&D approach. In other words, R&D relates rather imprecise connectionist notions like 'recurrence' to a specific understanding of dynamics of physical systems that is subject to well-known analytical tools. This makes it possible to rigorously design networks with highly complex dynamics (like the Kalman filter mentioned earlier), a task left mostly to chance with connectionism.

The previous discussion shows how principle three supports building networks that have the complex dynamics demanded by a higher-level hypothesis. In addition, principle three supports building networks that have the complex representations demanded by a higher-level hypothesis. For example, Eliasmith and Anderson[43] describe a model of working memory that accounts for representational phenomena not previously accounted for. Specifically, this model employed complex representations to demonstrate how working memory could be sensitive not only to spatial properties (i.e., position) but to other properties concurrently (e.g., shape). In addition, the model gives rise to both neural predictions (e.g., connectivity, and firing patterns), and psychological predictions (e.g., kinds of error, and conditions for error). This was possible only because R&D theory provides a means of determining the detailed connection weights given high-level descriptions of the system. Again, it is unclear how such a network could have been learned (that this is not an easy task is a good explanation for why it had not been previously done).

In both of these examples, R&D theory is distinguished from connectionism because it does not share the same heavy reliance on learning for model construction. Unfortunately, getting a model to learn what you want it to can be extremely challenging, even if you build in large amounts of innate information (and choosing what to build in tends to be something of an art). But, connectionists have little recourse to alternative methods of network construction, so the severe limitations and intrinsic problems with trying to learn complex networks are an inherent feature of connectionism. R&D theory, in contrast, allows for high-level characterizations of behavior to be imposed on the network that is constructed. As a result, connection weights can be analytically determined, not learned.

Nevertheless, R&D theory is also able to incorporate standard learning rules.[44] And, more than this, R&D theory can provide new insights regarding learning. This is because being able to analytically construct weights also provides some insight into

---

[43] "Beyond Bumps: Spiking Networks that Store Smooth N-Dimensional Functions" *Neurocomputting*, XXXVIII (2001): 581-586.

[44] As discussed in chapter 9 of Eliasmith and Anderson, *Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems*.

methods for deconstructing weights. So, given a set of learned weights, the techniques of R&D theory can be used to suggest what function is being instantiated by the network.[45] Often, when some input/output mapping has been learned by a connectionist network, it is very difficult to know exactly which function has been learned because the testable mappings will always be finite. Using R&D to determine which linear decoders combine to give a set of provided connection weights can be used to give an exhaustive characterization of what higher-level function is actually being computed by the network.

Such connection weight analyses are possible because R&D theory, unlike connectionism, explicitly distinguishes the encoding and decoding processes when defining representation and computation. While the relation between this distinction and the observable properties of neurobiological systems is subtle, as noted in section 3.2, the theoretical payoff is considerable.

To summarize, R&D theory should be preferred to connectionism for two main reasons. First, R&D provides for a better understanding of neural connection weights, no matter how they are generated. There is much less mystery to a network's function if we have a good means of analyzing whatever it is that determines that function. While learning is powerful, and biologically important, it cannot be a replacement for understanding what, precisely, a network is doing. Second, R&D theory provides a principled means of relating neural and psychological data. This makes representationally and dynamically complex cognitive phenomena accessible to neural level modeling. Given a high-level description of the right kind, R&D theory can help us determine how that can be realized in a neural system. Connectionists, in contrast, do not have a principled means of relating these two domains. As a result, high-level hypotheses can be difficult to test in a connectionist framework.

So, unlike connectionism, R&D theory carefully relates neural and psychological characterizations of behavior to provide new insights into both. And, while it is possible that certain hybrid models (either symbolicist/connectionist hybrids, or localist/distributed hybrids) may make up for some of the limitations of each of the components of the hybrid alone, there is an important price being paid for that kind of improvement. Namely, it becomes unclear precisely what the cognitive theory on offer is supposed to be. R&D theory, in contrast, is highly unified and succinctly summarized by three simple, yet quantifiable, principles. To put it simply, Occam's razor cuts in favor of R&D theory. But, it should be reiterated that this unification buys significantly more than just a simpler theory. It provides a unique set of conceptual tools for relating, integrating, and analyzing neural and psychological accounts of cognitive behavior.

4.3 Dynamicism

Of the three approaches, dynamicism, by design, is the most radical departure from the "mind as computer" metaphor. In some ways, this explains both its strengths and its weaknesses. Having derided talk of representation and computation, dynamicists have put in their place talk of "lumped parameters," and "trajectories through state-space." Unfortunately, it is difficult to know how lumped parameters (e.g., "motivation" and

---

[45] For a simple example, see Eliasmith and Anderson, *op. cit.*, pp. 294-298.

"preference")[46] relate to the system that they are supposed to help describe. While we can measure the arm angle of the Watt Governor, it is not at all clear how we can measure the "motivation" of a complex neurobiological system. But this kind of measurement is demanded by dynamicist models. As well, some dynamicists insist that cognitive models must be low-dimensional, in order to distinguish their models from those of connectionists.[47] But insistence on low-dimensionality greatly reduces the flexibility of the models, and does not seem to be a principled constraint.[48] Finally, because the Watt Governor, a standard example of a classical control system, has been chosen as a central exemplar of the dynamicists approach, the well-known limitations of classical control theory are likely to plague dynamicism. Clearly, these limitations are not ones shared by R&D theory.

What the replacement of the "mind as computer" metaphor by the "mind as Watt Governor" metaphor gained for dynamicists was an appreciation of the importance of time for describing the behavior of cognitive systems. No other approach so relentlessly and convincingly presented arguments to the effect that cognitive behaviors were essentially temporal.[49] If, for example, a system cannot make a decision before all of the options have (or the system has) expired, there is little sense to be made of the claim that such a system is cognitive. Furthermore, there is evidence that perception and action, two clearly temporal behaviors, provide the foundation for much of our "more cognitive" behavior.[50] While dynamicists have done a good job of making this kind of argument, the consequences of such arguments need not include the rejection of representation and computation that dynamicists espouse. R&D theory, which essentially includes precisely these kinds of dynamics, shows how representation, computation, and dynamics can be integrated in order to tell a unified story about how the mind works.

4.4 Discussion

So, in short, R&D theory adopts and improves upon the dynamics of dynamicism, the neural plausibility of connectionism, and the representational commitments of symbolicism. As such, it is a promising synthesis and extension of past approaches to

---

[46] These are two of the lumped parameters included in the model of feeding described in J. Busemeyer and J. Townsend, "Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment" *Psychological Review*, C, 3 (1993): 432-459.

[47] T. van Gelder, *op. cit.*

[48] These points are discussed in detail in my "Commentary: Dynamical Models and van Gelder's Dynamicism: Two Different Things" *Behavioral and Brain Sciences*, XXI, 5 (1998): 616-665, and my "The Third Contender: A Critical Examination of the Dynamicist Theory of Cognition".

[49] Such arguments are prominent in T. van Gelder, *op. cit.*, T. van Gelder, "The Dynamical Hypothesis In Cognitive Science" *Behavioral and Brain Sciences*, XXI, 5 (1998): 615-665, and the various contributions to R. Port and T. van Gelder, *op. cit.*

[50] This view, associated variously with terms "embodied", "embedded", or "dynamicist" has been expressed in, for example, F. Varela, E. Thompson, and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*. (Cambridge: MIT Press, 1991), and more recently in D. Ballard, M. Hayhoe, P. Pook, and R. Rao "Deictic Codes for the Embodiment of Cognition" *Behavioral and Brain Sciences*, in press.

understanding cognitive systems, because it includes the essential ingredients. Of course, it is not clear whether R&D theory combines those ingredients in the right way. Because it is a recent proposal for explaining cognitive systems, its current successes are few. While it has been used to effectively model perceptual (e.g. the vestibular system), motor (e.g. eye control), and cognitive (e.g., working memory) processes, these particular examples of perceptual, motor, and cognitive behavior are relatively simple. So, while the resources for constructing neurally plausible models of phenomena that demand complex dynamics over complex representations are available, it remains to be clearly demonstrated that such complexity can be incorporated into R&D theoretic models.

As well, R&D theory does not, in itself, satisfactorily answer questions regarding the semantics of representational states. As Dretske[51] has noted, coding theory does not solve the problem of representational semantics. Thus, R&D theory needs to be supplemented with a theory of meaning, as mentioned in section 3.1. In fact, I think R&D theory suggests a novel theory of meaning that avoids the problems of past theories.[52] Nevertheless, this remains to be clearly demonstrated.

However, even in this nascent form, R&D theory has some important theoretical implications for work in philosophy of mind and cognitive science. For example, functionalism regarding the identity of mental states may need to be reconceived. If, as R&D theory entails, the function of a mental state must be defined in terms of its time course, and not just its inputs and outputs, it is unlikely that functional isomorphism of the kind that Putnam[53] envisioned will be sufficient for settling the identity of mental states. If the dynamics of some aspects of mental life are central to their nature, then an atemporal functionalism is not warranted. Standard functionalism in philosophy of mind is clearly atemporal. And, I take it, some (if not many) aspects of mental life have their character in virtue of their dynamics (e.g., shooting pains, relaxed conversations, and recognizing friends). So, a “temporal” functionalism is necessary for properly characterizing minds. In other words, input, outputs, and their time course must all be specified to identify a mental state. While some mental functions may be not especially tied to dynamics (e.g., addition), others will be (e.g. catching a ball). Specifying ranges of dynamics that result in the successful realization of that function will allow us to determine if some mind or other can really be in a given mental state.

These considerations have further consequences for the role of the Turing machine in cognitive science.[54] While cognitive functions will still be Turing computable, they will not be realizable by every universal machine. This is because computing over time (i.e., with time as a variable in the function being computed) is

[51] *Knowledge And The Flow Of Information* (Cambridge: MIT Press, 1981).

[52] For an attempt at articulating such a theory, see my *How Neurons Mean: A Neurocomputational Theory of Representational Content*.

[53] “Philosophy and our Mental Life” in H. Putnam, ed., *Mind, Language and Reality: Philosophical Papers* (Cambridge: Cambridge University Press, 1975), pp. 291-303.

[54] These consequences are more fully explored in my “The Myth of the Turing Machine: The Failings of Functionalism and Related Theses” *Journal of Experimental and Theoretical Artificial Intelligence*, XIV (2002): 1-8.

different that computing in time (i.e., arriving at the result in a certain time frame). When this difference is acknowledged, it becomes clear that Turing machines as originally conceived (i.e., under the assumption of infinite time) are relevant theoretically, but much less so practically (i.e., for understanding and identifying real minds). I take it that more argument is needed to establish such conclusions, but that, at the very least, adopting R&D theory shows how such a position is plausible.

5 Conclusion

Perhaps, then, R&D theory or something like it can help rid us of the constraints of metaphorical thinking. Such an approach holds promise for preserving many of the strengths, and avoiding many of the weaknesses, of past approaches to understanding the mind. But, more than this, it is also suggestive of new perspectives we might adopt on some of the central issues in philosophy of mind and cognitive science.

Because cognitive science is interdisciplinary, it should not be surprising that a cognitive theory has consequences for a variety of disciplines. I have suggested some of the consequences of R&D theory for neuroscience (e.g., careful consideration of decoding), psychology (e.g., quantitative dynamic descriptions of cognitive phenomena), and philosophy (e.g., reconsidering functionalism). These are consequences that should be embraced in order to improve our understanding of cognitive systems. In other words, the time is ripe for moving beyond the metaphors.

6 Acknowledgements

Special thanks to Charles H. Anderson. Thanks as well to William Bechtel, Ned Block, David Byrd, Rob Cummins, Brian Keeley, Brian McLaughlin, William Ramsey, Paul Thagard and Charles Wallis for comments on earlier versions. Funding has been provided in part by the Mathers Foundation, the McDonnell Center for Higher Brain Function, and the McDonnell Project for Philosophy and the Neurosciences.